Working Paper No. E04-336

Expressed Preferences and Behavior in Experimental Games

Gary Charness
Department of Economics, University of California, Santa Barbara

Matthew Rabin
Department of Economics, University of California, Berkeley

January 2003

Abstract

It is traditional in experimental games to allow participants to choose only actions or possibly communicate intended play. In sequential two-person games, we require 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. We find that such preference expression after favorable first-mover play on average increases both the social surplus and the lowest payoff received by 15-20%. Expressed preferences for favorable responder behavior by first movers who have not behaved favorably are largely ignored, however, and may even be counter-productive. Our results replicate earlier findings, in that 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 (nonmisbehaving) person's payoffs to be positive, even when one is behind. While the degree of positive reciprocity is not significant either with or without expressed preferences, there is evidence that positive reciprocity is enhanced when a preference for favorable treatment is expressed.

Keywords: beliefs, experiment, expressed preferences, positive reciprocity, social preferences

JEL Classification: A12, A13, B49, C70, C91, D63

We thank Manuel Fernandez, Brit Grosskopf, and Ellen Quarles for help with the sessions, and Ellen Quarles for her help with the data. 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.

Express yourself.
-- Madonna

1. Introduction

Many recent experimental studies have demonstrated 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 up to the available choices at hand. One thing people are concerned with is the perceived intentions of others, meaning not just what choice another player makes, but *why* that choice was made. A key means of inferring the "intentions" of another player is to compare the choice made to the set choices that player could have made but didn't.

However, even this may leave another player's intentions open to interpretation, among other reasons because the intended consequences of that player's actions depend not only on the actions but on that player's hopes and expectations about how all players 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. Most experimental studies of social preferences have, however, not allowed any communication between the parties.[2] This "no-communication protocol" therefore misses an important element of many of the real-world social and economic interactions that presumably interest researchers studying experimental games.

Cheap talk has, unsurprisingly, been found experimentally to be very effective in coordinating intended future actions in coordination games. 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.

[1] For example, see Brandts and Solà (2001), Falk, Fehr & Fischbacher (forthcoming), Charness and Rabin (2002), and Bolton and Ockenfels (2000b).
[2] Exceptions include Brandts and Charness (forthcoming), Charness and Dufwenberg (2002), and Hannan, Kagel & Moser (2002).

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).

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 somewhat responsive to the explicit hopes of people who have not been unkind to them, but don't particularly care about disappointing selfish people. This result is in line with models of utility that incorporate beliefs, but is not consistent with purely consequential models.

Our results say that differences in what is expressed lead in some circumstances to different behavior. We are also interested in a second question: What are the net differences between experiments with expressed preferences and those without? This might provide some evidence, for instance, on whether experiments banning communication may be generating misleading general conclusions about social preferences in more realistic settings. We address this question by examining the average effects of expressed preferences on social welfare from the alternative standpoints of the total payoffs received and the minimum payoff received. In response games where entry by the first mover is favorable to the responder, there is evidence that expressed preferences lead to better social outcomes – in the aggregate, the expected total monetary payoff and the expected minimum payoff both increase by 15-20%. However, expressed preferences have little overall effect in games where entry is unfavorable to the responder and the responder can punish the first mover; in addition, requiring expressed preferences in dictator games actually seems to decrease social welfare.

We also 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 significantly positive, in contrast to the specifications in the Fehr and Schmidt (1999) and Bolton and Ockenfels (2000a) 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. Omitting reciprocity, responders' preferences when behind are not significantly different than self-interested. Our regressions also confirm a major role for expressed preferences in explaining responder behavior.

Models such as Rabin (1993) and Dufwenberg and Kirchsteiger (1998) 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, few experimental studies provide any evidence whatsoever 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. We hypothesized that an explicit statement of preferences might make the 'good intentions' implicit in a first mover's favorable choice sufficiently salient to induce positive reciprocity. By comparing the difference between behavior in response to favorable plays and behavior in dictator games, we find some suggestive evidence of positive reciprocity when expressed preferences are involved. This suggests that a stronger form of communication might lead to a greater degree of positive reciprocity. But by conventional measures we still do not find statistically significant evidence of positive reciprocity, reinforcing the general finding in the literature.[3]

[3] We changed one other element of our previous experimental design, eliminating the feature of playing each game twice, with each person playing each role once, but paired with another person. While such role reversal offers insight into implied beliefs and the consistency of play across roles, many observers have expressed the concern that this design might change people's behavior. We replicated many of the games played with role reversal in Charness and Rabin (2002), and could not find any pattern of changes in behavior due to role reversal.

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 further discussion in Section 5, and conclude in Section 6.

2. Beliefs and Social Preferences

The ultimatum game (Güth, Schmittberger, and Schwarze, 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 (2000a) responders reject unfair proposals because of *per se* aversion to disparities in relative payoffs. From this perspective, the intentions of proposers are irrelevant, so *a fortiori* responders' beliefs about these intentions 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, Fehr, and Fischbacher (forthcoming) find substantial differences in both first-mover and responder behavior according to the foregone first-mover choice. 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. Perceived intentions may come into play through considerations of *reciprocity*. Studies in a variety of social science disciplines suggest that reciprocity is a basic motivational drive in social interaction.

There are a number of studies that demonstrate the importance of negative reciprocity. Kahneman, Knetsch, and Thaler (1986) had A's choose between (A,B,C) payoffs of (5,5,0) vs. (6,0,6) when they knew that B's had previously chosen an even allocation in an earlier (and independent) dictator game, while C's had chosen a selfish one. 74% of A's 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 (forthcoming) 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" employer when they received a meager allocation chosen at random.

There are fewer clear demonstrations of the impact of good intentions. McCabe, Rigdon, and Smith (2000) find evidence of significant positive reciprocity in a simple "trust" game, and Falk, Fehr, and Fischbacher (2001) observe positive reciprocity in a moonlighting game. Offerman (forthcoming) finds some evidence of positive reciprocity, though not to a statistically significant degree; Brandts and Charness (forthcoming) also find modest positive reciprocity in their cheap-talk game. But there are far more studies indicating a lack of positive reciprocity. Bolton, Brandts, and Ockenfels (1998), Bolton, Brandts, and Katok (2000) find evidence against positive reciprocity. In a rigorous test, Cox (2000) finds no evidence at all for positive reciprocity. Based on the data in our earlier study (Charness and Rabin, forthcoming), we found positive reciprocity when it was free, but strong evidence against it otherwise.

Helpful sacrifice in laboratory games can generally be explained by distributional considerations (Bolton and Ockenfels, 2000a; Fehr and Schmidt, 1999; Charness and Rabin, 2002). Standard games cited as evidence for positive reciprocity (e.g., the gift-exchange game and the trust game) confound altruism, equity, and reciprocity, since we do not observe responder choices when a favorable choice set is merely a windfall. While a series of gift-exchange experiments beginning with Fehr, Kirchsteiger and Riedl (1993) show convincingly that people return higher effort for higher wages, the payoff structure of this game doesn't allow us to distinguish between altruistic concerns and positive reciprocity.[4] The Berg, Dickhaut & McCabe (1995) investment (or trust) game is also often considered to be evidence of positive

[4] Consider a standard design, where $\pi_{Employee}$ = wage – effort cost and π_{Firm} = (100 – wage)*effort, with wage $\in$ [0,100], effort $\in$ {.1,.2,…,1.0}, and effort cost increasing in the amount of effort chosen. Suppose the wage is 60. If the employee (responder) chooses the costless effort level (effort = .1), the (Employee, Firm) monetary payoffs would be (60,6). By choosing an effort level of, for example, .5 (effort cost = 6) the responder could change these to (54,30). We might expect that many employees

reciprocity, although the authors are careful to frame their results in terms of trust.[5] While many participants return more than was sent by the first mover, dictator game evidence indicates that many people would choose such allocations even where the other player has made no choice.[6] 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 in Fehr and Schmidt (1999), Bolton and Ockenfels (2000a), and Charness and Rabin (2002) can all explain these results without invoking reciprocity.

To the extent that one's behavior is sensitive to one's beliefs about the intentions or desires of other involved parties, 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. Clearly people may treat such communication as carrying relevant information. Cooper, DeJong, Forsythe, and Ross (1992) find a very high degree of coordination with two-way communication and a smaller, but still significant and substantial, effectiveness for one-way communication. Charness (2000a) finds that pre-play communication is very effective even when the message can clearly be self-serving.

There are several experiments in which social preferences are clearly relevant where some form of anonymous communication is permitted. Brandts and Charness (forthcoming) 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

---

would prefer (54,30) to (60,6) even if the employer had nothing to do with the choice of the wage, in line with many studies showing difference-averse, altruistic, or "social-welfare" preferences.

[5] A necessary condition for their trust definition (p. 126) is that both people must be "made better off from the transaction compared to the outcome which would have occurred if the trustor had not entrusted the trustee."

[6] Consider the case where the sender sends all 10 allocated units. Now the responder has 40 units to allocate at will. Many people playing a dictator game with 40 units to distribute would choose to allocate more than 10 units to the other person. Aside from the intention behind the sender's play, these cases are identical. In addition, a test comparing the amount sent to the percentage returned gives a Spearman rank correlation coefficient of $r_s = 0.01$, "[suggesting] no correlation between amounts sent and payback decisions" (p. 132).

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 is Hannan, Kagel & Moser (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.[7] 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 $0 \leq y \leq 20$.[8] A's were then asked to guess the average y chosen by B's and, simultaneously, B's were asked to guess the average guess of A's. Guesses were rewarded monetarily according to *ex post* accuracy. 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. In our context, this suggests that behavior might be different if first-mover preferences are clarified and expectations brought into sharper focus.

[7] Fehr, Klein & Schmidt (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.

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

3. Experimental Design and Results

We conducted 11 sessions (with 48 distinct games) in Berkeley. Participants played eight games in a session, and knew that they would be paid according to the outcome generated in only two of these games, randomly selected. Recruiting at Berkeley was done primarily through the use of campus e-mail lists. Each person could participate in at most one session; altogether there were 289 participants in the sessions. Average earnings were around $16 in Berkeley, about $11 net of the show-up fee paid, for a session lasting about an hour. Experimental instructions are provided in Appendix A. Each of 40 simple binary-choice extensive-form games was played in two of Sessions 1-10. 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 chosen to fill in missing conditions whose absence diminished our ability to draw inferences from the original array of games.[9] We conducted no "pilot studies" prior to the experiments in these games except the closely-related games reported in Charness and Rabin (2002).

We also conducted four sessions in Barcelona. Here participants played four games in a session, and knew that they would be paid according to the outcome in only one game, randomly selected. Recruiting in Barcelona was by posting announcements around campus. Each person could participate in at most one session; altogether there were 117 participants in the sessions. Average earnings were around $7 in Barcelona, about $4 net of the show-up fee paid for a 40-minute session. Each of 10 simple binary-choice extensive-form games was played in two of the sessions.[10]

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

[9] Some games in Session 11 were designed after observing outcomes in the first 10 sessions. The Barcelona sessions we designed took place after the Berkeley sessions, and were conducted to complete comparisons with games conducted there and reported in Charness and Rabin (2002).

[10] 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.

randomly paired, and people were told (truthfully) that they would never be matched twice with the same person. 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 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, an eight-sided die was rolled (at least) twice to determine which two of the eight games would be chosen for actual payments and these were calculated. People were paid individually and privately.

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.[11] We conducted games both with expressed preferences and without. 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.

Note that A was not allowed to remain silent. We chose this design, rather than allowing silence by A, to enhance the power and simplicity of our tests by limiting the number of choices. Clearly this is less realistic than allowing silence, and makes our results harder to interpret than would be ideal. We speculate later on the possible effects of not permitting silence by A. We asked responders to make choices following hypothetical A2/prefer B1 and A2/prefer B2 selections. In the latter case, A simply indicated A1 or A2, and B indicated B1 or B2.

[11] 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. See Charness and Rabin (2002) for a brief discussion on this point, and Cason and Mui (1998) and Brandts and Charness (2000), where this difference in elicitation methods does not appear to affect behavior. Brandts and Charness (forthcoming) find that punishment levels are lower with the strategy method, although all qualitative results reported there hold for both the strategy and direct-response elicitation methods.

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 and 2 show all the results from this paper. In these and all tables, 100 units of lab money equal $1.00 in the Berkeley sessions, and equals 100 pesetas (worth 57 cents at the time of the experiments) in the Barcelona sessions.

In this Table (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").[12] 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.

Table 1: Games with preference expression

	A preference			**B's helping A**	
	(for B to play Left)			A hopes Left	A hopes Right
Dictator games					
(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%)
	A play & preference			**B's helping A**	
Response Games	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%)

Barcelona games are in italics.

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

Table 2: Games without preference expression

	A play	B's helping A
Dictator games		
(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%)
	A's entering	**B's helping A**
Response Games		
(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); (750,375)-(400,400)	3/25 (12%)	3/25 (12%)
(550,550); (750,400)-(400,400)	5/26 (19%)	12/26 (46%)
(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%)

Barcelona games are in italics.

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 in Charness and Rabin (2002) and without role reversal here. Table B1 in Appendix B reports results 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). In the B case, 6 of the 11 comparisons give a two-tailed p-value above .50, quite consistent with the *ex ante* random prediction; in the A case, 6 of the 7 comparisons give a p-value above .50, suggesting remarkably little difference. 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 3 presents this evidence:

Table 3: Dictator games and preferences

Game	B's helping A (by A preference)			
	No express	Help me	Don't help me	Agg. Pref
(750,375)-(400,400)	13/30 (43%)	8/26 (31%)	3/25 (12%)	25%
(400,400)-(0,800)	17/27 (63%)	17/27 (63%)	11/27 (41%)	58%
(750,400)-(400,400)	17/25 (68%)	12/25 (48%)	6/25 (24%)	39%
(600,600)-(200,700)	16/22 (73%)*	14/20 (70%)	5/20 (25%)	68%
(450,350)-(350,450)	1/20 (5%)	5/20 (25%)	6/20 (30%)	26%
Aggregated total	**64/124 (52%)**	**56/118 (47%)**	**31/117 (26%)**	**43%**

The aggregated preference outcomes were calculated as follows: Multiply the proportion of those A's expressing preferences for Left by the proportion of B's then playing Left, and do the same for those A's expressing preferences for Right.

*In all tables, asterisked entries refer to data from Charness and Rabin (2002).

In four of the five dictator games in Table 3, the proportion of B players who maximize A's payoff is considerably higher following "Help Me" than following "Don't Help Me". The fifth game suggests that it is not so acceptable to ask for favorable treatment when the payoffs

are symmetric and there is no social benefit involved – here B is actually slightly less likely to help A when this preference is expressed. Nevertheless, the differences in B play are significant in the other four individual games at $p = .05$ or better; if we aggregate the data by column, the difference is significant at $p = .00$. [13] Thus, B choices are generally sensitive to A's expressed preferences, even though these are unaccompanied by any action. However, note that A is worse off when forced to express preferences than when forced to be mute, with the likelihood of a favorable B choice reduced to 43% from 52%; this 9% difference between aggregated silence behavior and aggregated preference-expression behavior is marginally significant ($Z = 1.34$, $p = .09$, one-tailed test).

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 B's payoff, while lowering A's own payoff.[14, 15]

[13] Throughout this and subsequent sections, the p-value is approximated to two decimal places 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.

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. Thus, for statistical purposes, each observation in our aggregated comparisons within Table 3 and within other Tables organized by categories is largely independent.

[14] 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.

[15] In this Table and others, we sometimes 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 re-run in our current study.

Table 4: Favorable A play and preferences

Game	B's helping A (by A preference)			
	No express	Help Me	Don't Help Me	Agg. Pref^
A(800,0); B(750,375)-(400,400)	9/27 (33%)	7/24 (29%)	2/25 (8%)	(21%)
A(800,0); B(400,400)-(0,800)	9/30 (30%)	16/26 (62%)	4/25 (16%)	(62%)
A(450,0); B(450,350)-(350,450)	7/27 (26%)	12/24 (50%)	4/25 (16%)	(43%)
A(700,200); B(600,600)-(200,700)	25/32 (78%)*	16/20 (80%)	8/19 (42%)	(80%)
A(450,0); B(450,350)-(350,450)	2/36 (6%)*	14/26 (54%)	8/26 (31%)	(54%)
A(750,100); B(700,500)-(300,600)	9/36 (25%)*	16/27 (59%)	7/27 (26%)	(59%)
Aggregated total	**61/188 (32%)**	**81/147 (55%)**	**33/147 (22%)**	**(53%)**

Barcelona games are italicized.

^Here, and in later Tables, the aggregated preference outcomes were calculated as follows: Multiply the proportion of those A's expressing preferences for Left by the proportion of B's then playing Left, and do the same for those A's expressing preferences for Right. *indicates results from Charness and Rabin (2002)

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. This is true in every case, and each of the six comparisons is significant at $p = .05$ (one-tailed tests); the aggregated 33% difference is highly significant ($Z = 5.75$, $p = .00$).

Perhaps more noteworthy is that when A has made a favorable choice, an expressed preference for help substantially improves the likelihood of a favorable response in comparison to the no-expression case. In five of six games where A's choice to enter is favorable to B, the difference is significant in the predicted direction at $p = .05$; if we aggregate the data, the average difference of 23% is highly significant ($Z = 4.16$, $p = .00$). In addition, when we take into account the actual preferences expressed and compute the aggregate result, preference expression leads to a considerably higher likelihood of a favorable response than the silence case ($Z = 3.79$, $p = .00$).

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

Table 5: Unfavorable A play and preferences

Games where B can help A at little or no cost	B's helping A (by A preference)			
	No express	Help Me	Don't Help Me	Agg. Pref
(550,550); (750,375)-(400,400)	3/25 (12%)	4/27 (15%)	1/27 (4%)	(15%)
(100,1000); (125,125)-(75,125)	22/26 (85%)	24/30 (80%)	14/30 (47%)	(78%)
(550,550); (750,400)-(400,400)	12/26 (46%)	15/28 (54%)	6/26 (23%)	(54%)
Aggregated total	**37/77 (48%)**	**43/85 (51%)**	**21/83 (25%)**	**(49%)**

Games where B can make a costly sacrifice to punish A	B's hurting A (by A preference)			
	No express	Punish Me	Don't Punish Me	Agg. Pref
(700,1300); (800,200)-(0,0)	4/25 (16%)	0/11 (0%)	0/11 (0%)	(0%)
(400,1200); (400,200)-(0,0)	4/25 (16%)	3/22 (14%)	4/25 (16%)	(15%)
(375,1000); (400,400)-(250,350)	3/26 (12%)*	0/11 (0%)	2/11 (18%)	(14%)
Aggregated total	**11/77 (14%)**	**3/44 (7%)**	**6/47 (13%)**	**(10%)**

*indicates results from Charness and Rabin (2002)

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. We are a bit surprised by this and we had no directional hypothesis (and so use a two-tailed test). In any case, each of the three comparisons is significant at $p = .10$; the aggregated column responses are significantly different at $p = .00$.

However, the pattern is quite different in the three games where B can choose a costly punishment. Here, although the difference is not statistically significant, responders are nearly twice as likely to punish when A states a preference against punishment. Perhaps the responder is more charitable when A appears confused, or perhaps this pattern of play stems from negative reciprocity. In any case, responders do not 'respect' A's preference when 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 Table B3 (Appendix B). Here there is no consistent pattern when A acts favorably and asks for an unfavorable response. The difference is significant in one direction in two of the five games, but is significant in the other direction in a third game. Pooling across columns gives a difference with $Z = 1.12$, perhaps suggesting a tendency to comply with A's expressed wishes. The effect on responder choices after A acts unfavorably and asks for an unfavorable response depends on

whether it is costly to comply with the expressed preference. 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.[16]

One issue that interested us is whether preference expression produces better social outcomes. We consider two proxies for social welfare: 1) the total material payoffs received, and 2) the minimum payoff received. These two parameters are the fundamental components of the purely distributional element of the Charness and Rabin (2002) model, and have been identified as key social factors in studies such as Engelmann and Strobel (2002).

We consider three classes of games: dictator games in which sacrifice is beneficial to the other player, response games in which responder sacrifice is beneficial to the first mover, and response games in which the responder can choose to punish the first mover for an unfavorable entry choice. In Table 6, efficiency is defined as the expected percentage achieved of the potential joint payoff sum (for all possible combinations of A and B choices), where 0% represents the minimum feasible payoff sum and 100% is the maximum. The expected minimum is defined in terms of the actual minimum payoff as a percentage of the highest minimum payoff achievable in the game.

[16] We examine the effect of preference expression on A play in Table B2 of Appendix B. 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 6. Social welfare in *non-punishment* games: Expected efficiency and expected minimum payoff, expressed preferences vs. silence

	Efficiency			Minimum		
	Pref	*NP*	*NP'*	*Pref*	*NP*	*NP'*
Dictator games						
(750,375)-(400,400)	.250	.433	.500	.484	.567	.500
(400,400)-(0,800)	-	-	-	.576	.440	.219
(750,400)-(400,400)	.394	.680	.692	-	-	-
(600,600)-(200,700)	.678	-	.727	.678	-	.727
(450,350)-(350,450)	-	-	-	-	-	-
Average dictator game	**.440**	**.576**		**.579**	**.530**	
	Pref	*NP*	*NP'*	*Pref*	*NP*	*NP'*
Response Games						
(800,0); (750,375)-(400,400)	.094	.111	-	.434	.326	-
(800,0); (400,400)-(0,800)	-	-	-	.142	.110	.174
(450,0); (450,350)-(350,450)	.560	.407	.375	.560	.407	.375
(700,200); (600,600)-(200,700)	.240	-	.342	.240	-	.342
(750,100); (700,500)-(300,600)	.152	-	.030	.186	-	.052
(450,0); (450,350)-(350,450)	.310	-	.306	.310	-	.306
Average response game	**.271**	**.236**		**.312**	**.260**	

Barcelona games are in italics.
NP refers to no-preference games in our current results, and *NP'* refers to no-preference results for the same games in the same location in our previous study.

The NP and NP' percentages are pooled for each game, and these six games are then averaged for the numbers in the bottom row. We do this to facilitate comparisons between the preference-expression and no-preference-expression cases. Given that the subject pool and location was identical and the protocols were very similar (see footnote 14), we feel justified in following this procedure.[17] The outcomes were calculated by multiplying the applicable percentage choices of A's and B's.

Preference expression actually seems to be detrimental in the dictator games, where A is helpless without preference expression, but has a voice with it. Previous papers have noted a possible tendency for B to be more generous if A has had no choice than if A has made a choice,

[17] There are seven direct comparisons available between NP and NP' sessions for the same game. In three of these, the measure was higher for NP.

even if A's choice is favorable.[18] Here simply having a voice without an action is sufficient to reduce generosity. Interestingly, this negative impact is seen primarily with respect to efficiency. Asking someone to sacrifice for you to increase the total payoff is ineffective at best, particularly when this sacrifice would lower the minimum payoff. However, we see that aggregate outcomes are improved by preference expression when A has acted favorably by entering. In the aggregate, the expected efficiency and the expected minimum payoff increase by 15-20% with preference expression.

What happens by this measure in punishment games? Here preference expression is largely irrelevant, as is shown in Table 7:

Table 7. Social welfare in *punishment* games: Expected efficiency and expected minimum payoff , expressed preferences vs. silence

	Efficiency			**Minimum**		
	Pref	*NP*	*NP'*	*Pref*	*NP*	*NP'*
Response Games						
(100,1000); (125,125)-(75,125)	.553	.487	.518	.629	.686	.578
(700,1300); (800,200)-(0,0)	.792	.791	-	.702	.726	-
(550,550); (750,400)-(400,400)	.771	.781	-	.733	.808	.389
(400,1200); (400,200)-(0,0)	.891	.726	.846	.908	.768	.871
(375,1000); (400,400)-(250,350)	.487	--	.678	.854	-	.918
Aggregate Average Percentage	**.699**	**.708**		**.765**	**.739**	

Barcelona games are in italics.
NP refers to no-preference games in our current results, and *NP'* refers to no-preference results for the same games in the same location in our previous study.

So it may well be that people are reluctant to disappoint others, but this seems to be mainly the case when these other people have acted cooperatively. Giving someone a voice without an action actually seems to make things worse from a social standpoint.

We next examine whether preference expression helps to induce positive reciprocity. Charness and Rabin (2002) replicated earlier research in finding little evidence for positive reciprocity without expressed preferences. When a favorable move comes bundled with a stated preference for favorable treatment, will the good intentions present behind A's entry be brought more into play with an expressed preference?

[18] Charness and Rabin (2002) observe some weak evidence of such *complicity effects*. Charness (2000b) presents evidence that people may be more generous when the responsibility for an allocation rests

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. We first replicate the results in Charness and Rabin (2002), without expressed preferences:

Table 8 – Positive Reciprocity without Expressed Preferences

Games without Expressed Preferences	B's helping A when:		Z (one-tailed p-value)
	A has a play	Dictator version	
A(800,0); B(400,400)-(0,800)	9/30 (30%)	11/25 (44%)	-1.07 (.86)
A(800,0); B(750,375)-(400,400)	9/27 (33%)	13/30 (43%)	-0.77 (.78)
A(450,0); B(450,350)-(350,450)	7/27 (26%)	1/20 (5%)	1.89 (.03)
A(450,0); B(450,350)-(350,450)	2/36 (6%)	3/29 (10%)	-0.72 (.76)
Aggregated total	**27/120 (22%)**	**28/104 (27%)**	**-0.77 (.78)**

Barcelona games are italicized.

We see that B is slightly *less* likely to help A after a favorable play in three of the four cases, which is evidence against positive reciprocity. If we aggregate the rows in each column and compare across columns, we get a difference of Z = -0.77 (p = .78, one-tailed test). There is no evidence of positive reciprocity without expressed preferences.

Turning to behavior with expressed preferences, we see a slightly different picture:

Table 9 –Positive Reciprocity with Expressed Preferences

Entering is a favorable play and A prefers a favorable response	B's helping A		Z (one-tailed p-value)
	Preference case	Dictator version	
(800,0); (400,400)-(0,800)	16/26 (62%)	16/27 (59%)	0.17 (.43)
(800,0); (750,375)-(400,400)	7/24 (29%)	8/26 (31%)	-0.12 (.55)
(450,0); (450,350)-(350,450)	12/24 (50%)	5/20 (25%)	1.70 (.04)
(700,200); (600,600)-(200,700)	16/20 (80%)	14/20 (70%)	0.73 (.23)
Aggregated total	**51/94 (54%)**	**43/93 (46%)**	**1.10 (.14)**

entirely on one's shoulders than when one can rationalize that someone else shares the responsibility.

Overall, there appears to be some tendency for positive reciprocity *per se* to be triggered by a stated preference for favorable treatment. This is not statistically significant; nevertheless, it appears that expressing a preference for a favorable play is beneficial, when we consider the 13% difference in the difference across columns in the aggregated data in Tables 8 and 9. The test of proportions finds this difference to be statistically significant (Z = 2.02, p = .02). Expressed preferences do appear to make good intentions salient to some degree, although the bulk of the improvement in B behavior would appear to stem from an unwillingness to go against the hopes (or expectations) of someone who has not misbehaved.

5. Regression Analysis and Discussion

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 use the simple conceptual model of social preferences in two-person games presented in Charness and Rabin (2002). Letting π_A and π_B be Player A's and B's money payoffs, consider the following formulation of Player B's preferences:

$$U_B(\pi_A,\pi_B) \equiv (\rho \cdot r + \sigma \cdot s + \theta \cdot q) \cdot \pi_A + (1 - \rho \cdot r - \sigma \cdot s - \theta \cdot q) \cdot \pi_B,$$

where $r = 1$ if $\pi_B > \pi_A$, and $r = 0$ otherwise;
$s = 1$ if $\pi_B < \pi_A$, and $s = 0$ otherwise;
$q = -1$ if A has misbehaved, and $q = 0$ otherwise.

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, and θ capture various aspects of social preferences. The parameter θ provides a mechanism for modeling reciprocity. The parameters ρ and σ allow for a range of different "distributional preferences", that rely solely on the outcomes and not on any notion of reciprocity.

We estimate the population means for ρ and σ, the respective weight one assigns to the material payoff of the other player when this payoff is less than or greater than one's own, and θ, the weight one assigns to reciprocating, by performing maximum-likelihood estimation on our binary-response data. In this approach, the logit regression

$$P(action\,1) = \frac{e^{\gamma \cdot u(action1)}}{e^{\gamma \cdot u(action1)} + e^{\gamma \cdot u(action2)}}$$

determines the values that best match predicted probabilities of play with the observed behavior.[20] Table 10 reports regression results for these and other variables under a spectrum of different restriction assumptions.[21]

[19.] Another way of writing this utility function that some readers might find more intuitive is to break it down into two cases: When $\pi_B \geq \pi_A$, $U_B(\pi_A,\pi_B) \equiv (1-\rho-\theta q)\pi_B + (\rho+\theta q)\pi_A$; when $\pi_B \leq \pi_A$, $U_B(\pi_A,\pi_B) \equiv (1-\sigma-\theta q)\pi_B + (\sigma+\theta q)\pi_A$.

[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 also tried specifications including a dummy variable for whether player B has a unilateral choice of allocations. However, the estimated coefficient on this dummy is tiny, and we omit these specifications. This provides no overall support for the conjectured complicity effects mentioned above.

Table 10. Regression estimates for B behavior (N=1491)

Model	Restrictions	ρ	σ	θ	HN	HM	γ	LL
Self-interest	$\rho = \sigma = \theta =$ HN = HM = 0	-	-	-	-	-	.003 (10.2)	-988.2
ρ only	$\sigma = \theta =$ HN = HM = 0	.389 (22.1)	-	-	-	-	.009 (12.8)	-932.4
ρ, σ only	θ = HN = HM = 0	.390 (22.3)	-.010 (-0.55)	-	-	-	.009 (12.3)	-932.3
ρ, θ only	σ = HN = HM = 0	.395 (25.8)	-	-.093 (-3.27)	-	-	.011 (12.7)	-927.6
ρ, σ, θ	HN = HM = 0	.394 (25.9)	.033 (1.47)	-.127 (-3.53)	-	-	.011 (12.7)	-926.3
ρ, σ, θ, Hope Nice, Hope Mean	None	.447 (17.4)	.064 (2.21)	-.134 (-3.61)	.048 (1.75)	-.218 (-7.02)	.012 (13.0)	-882.9

t-statistics are in parentheses. HN ≡ Hope Nice, HM ≡ Hope Mean, γ is the precision parameter, and LL is the log-likelihood function.

In all the regressions, ρ is estimated to be around .4 and is always highly significant. σ is always small, ranging from -.01 to .06. In the regression at the bottom of the chart, it is actually significantly positive. We see a large and highly significant negative coefficient on the dummy for Hope Mean, and a smaller and marginally significant positive one for the Hope Nice dummy.

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. Doing so, we obtain Table 11, which essentially confirms these assessments:

Table 11. Likelihood-ratio tests on parameter restrictions, B behavior (N=1491)

Restriction	LL	χ^2	*p*-value
None	-882.853	-	-
Hope Nice = 0	-884.460	3.21	.08
$\sigma = 0$	-885.433	5.16	.03
$\theta = 0$	-888.857	12.0	.00
Hope Mean = 0	-911.427	57.1	.00
$\rho = 0$	-942.272	118.8	.00

LL is the log-likelihood function.

The strongest explanatory power for behavior comes from ρ, so that the most important non-selfish motive is a player's desire when ahead to increase the other player's payoff. The next most powerful factor is that of an expressed preference for unfavorable treatment. We also see that A's misbehavior is a very significant factor. The significantly *positive* baseline level for σ suggested in Table 10 is confirmed by this test. An expressed preference for favorable treatment is marginally significant.

6. Conclusion

Behavior by responders in our simple games is quite sensitive to the preference expressed by a first mover who has not misbehaved. The greatest effects observed result from the clarification of intentions behind a favorable play. Responders are much more likely to help first movers after a favorable play when there has been a preference expressed for help than when no such preference can be indicated. Expressed preferences by A affect B's behavior even when A has made no play, so to some extent the responder is simply complying with the first mover's stated hope.

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 (1999) and Charness and Rabin (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 the 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 in this case. On the other hand, social welfare in punishment games is unaffected by preference expression, and social welfare in dictator games seems to be adversely affected.

On this last point, perhaps people consider it to be inappropriate to express a preference for financial sacrifice on one's behalf when one has not taken an action. In fact, there is a significantly greater likelihood that A's express a preference for unfavorable treatment in dictator games than in games where entry by A is unfavorable to B (21% compared to 10%, $Z = 2.01$, $p = .04$). In the dictator-game comparisons in Table 3, we see that, in the aggregate, B's are more likely to help A's when no preference can be expressed than when a preference for favorable treatment is expressed. This leads us to speculate on the outcomes that would occur if participants in the preference games were also given the option of not expressing a preference. Our suspicion is that would not have changed behavior much in games with favorable entry, but that a substantial proportion of A's would have chosen to stay mute in dictator games and punishment games if they had been given this option.

We do not observe substantial positive reciprocity *per se* for a favorable first-mover play either without preference expression or when it is accompanied by a preference for a favorable response. However, there is significant movement in this direction, as seen by comparing Table 8 and 9. We feel it is quite possible that a more personal form of communication might well induce a greater degree of positive reciprocity. 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 often an important element in real-world social 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.

Appendix A: Sample Instructions (Barcelona)

INSTRUCTIONS

Thank you for participating in this experiment. You will receive 400 pesetas for your participation, in addition to other money to be paid as a result of decisions made in the experiment.

You will be involved in 4 situations where decisions are made. You will make binary decisions in most or all of these situations. Each decision (and outcome) is independent from each of your other decisions, so that your decisions and outcomes in one situation will not affect your outcomes in any other situation.

In every case, you will be anonymously paired with one (or more) other people, so that your decision may affect the outcomes of others, just as the decisions of the other people in your group may affect your outcomes. For every decision task, you will be paired with a <u>different</u> person or persons than in previous decisions.

There are A and B "roles" in each situation. Every person will receive the decision forms for each role in each situation. These sheets are color-coded by role. Please read each one (for each situation in turn; we will proceed one at a time) and ask question(s) if you wish a further explanation. When everyone has become familiarized with these, we will flip a coin to determine the role for each participant.

If a situation has multiple decisions (some situations only have decisions for one role), these decisions will be made sequentially, in alphabetical order: "A" persons will complete their decision sheets first and their decision sheets will then be collected. Next, "B" persons complete their decision sheets and these will be collected.

When you have made a decision, please turn your decision sheet over, so that we will know when people have finished.

You will not be informed of the results of any previous decisions prior to making subsequent decisions.

Although there will be a total of 4 outcomes, only 1 of these outcomes will be selected for determining monetary reward. A die will be rolled at the end of the experiment for this purpose.

At the end of the session, you will be given a receipt form to be filled out and you will be paid individually and privately.

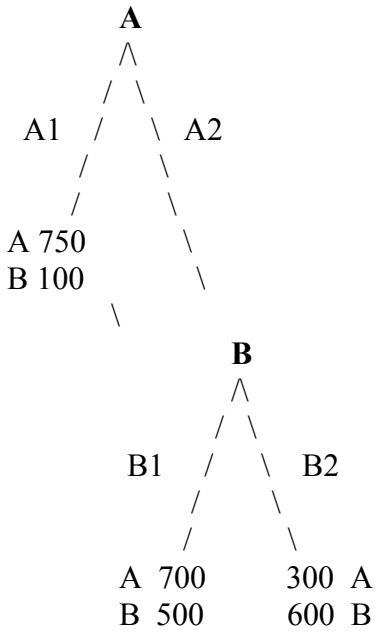
Please feel free to ask questions at any point if you feel you need clarification. Please do so by raising your hand. Please DO NOT attempt to communicate with any other participants in the session until the session is concluded.

We will proceed to the decisions once the instructions are clear. Are there any questions?

DECISION 1

You are **person A**.

You may choose A1 or A2. If you choose A1, you would receive 750 and person B would receive 100. If you choose A2, then B's choice of B1 or B2 would determine the outcome. If you choose A2 and B chooses B1, you would receive 700 and B would receive 500. If you choose A2 and person B chooses B2, you would receive 300 and he or she would receive 600. **Person B knows that his or her choice only affects the outcome if you choose A2, so that he or she will choose B1 or B2 on the assumption that you have chosen A2 over A1.** Please mark a choice below. Person B will make a decision contingent on your decision/preference.



DECISION/ PREFERENCE

I choose **A1** _____

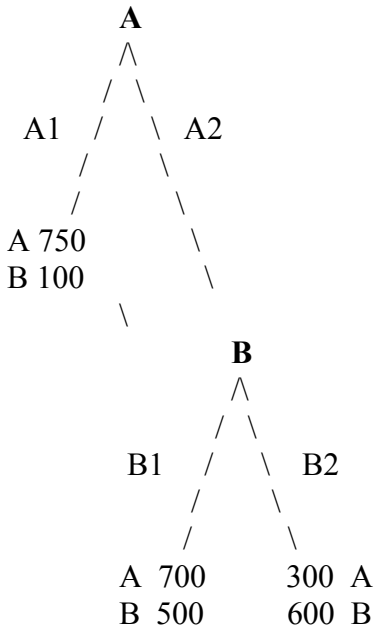
I choose **A2 with the hope that B selects B1**,
because I prefer the outcome (700,500) to (750,100)_____

I choose **A2 with the hope that B selects B2**,
because I prefer the outcome (300,600) to (750,100)_____

DECISION 1

You are **person B**.

You may choose B1 or B2. Person A has already made a choice. If he or she has chosen A1, he or she would receive 750 and you would receive 100. **Your decision only affects the outcome if person A has chosen A2. Thus, you should choose B1 or B2 on the assumption that A has chosen A2 over A1.** If A has chosen A2 and you choose B1, you would receive 500 and A would receive 700. If person A has chosen A2 and you choose B2, then you would receive 600 and person A would receive 300.



If person A has chosen A2, he or she has made one of two possible comments. Please indicate your choice of B1 or B2 in each of these cases:

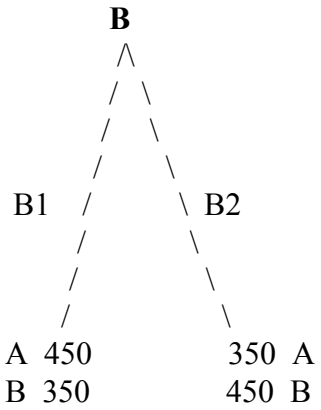
DECISION
(mark one in <u>each</u> row)

1) If A has stated "I choose A2 with the hope that B selects B1, because I prefer the outcome (700,500) to (750,100)," I choose:	**B1**	**B2**
2) If A has stated "I choose A2 with the hope that B selects B2, because I prefer the outcome (300,600) to (750,100)," I choose:	**B1**	**B2**

DECISION 2

You are **person A**.

You have no choice in this situation. Person B's choice determines the outcome. If person B chooses 1, you would receive 450 and person B would receive 350. If person B chooses 2, you would receive 350 and B would receive 450.



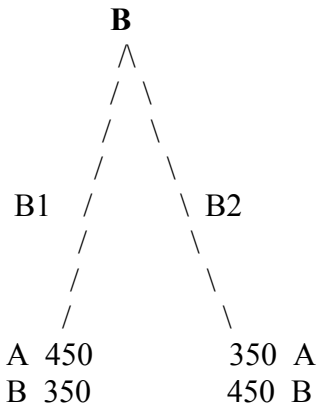
DECISION

I understand I have no choice in this situation _____

DECISION 2

You are **person B**.

You may choose B1 or B2. Person A has no choice in this situation. If you choose B1, you would receive 350 and person A would receive 450. If you choose B2, you would receive 450 and person A would receive 350.



DECISION

I choose: **B1** **B2**

Appendix B – Further Detail

Table B1 – The Effect of Role Reversal

B play	B's helping A		Z (two-tailed p-value)
	Role Reversal	No Role reversal	
B (750,375)-(400,400)	16/32 (50%)	13/30 (43%)	0.53 (.60)
B (750,400)-(400,400)	18/26 (69%)	17/25 (68%)	0.09 (.93)
B (800,200)-(0,0)	36/36 (100%)	11/11 (100%)	0.00 (1.00)
B (400,400)-(0,800)	7/32 (22%)	11/25 (44%)	-1.78 (.07)
(100,1000); (125,125)-(75,125)	21/32 (66%)	22/26 (85%)	-1.64 (.10)
(450,0); (450,350)-(350,450)	6/32 (19%)	7/27 (26%)	-0.66 (.51)
(800,0); (400,400)-(0,800)	12/22 (55%)	9/30 (30%)	1.78 (.07)
(550,550); (750,375)-(400,400)	4/22 (18%)	3/25 (12%)	0.48 (.63)
(0,800); (400,400)-(0,800)	18/32 (56%)	3/12 (25%)	1.85 (.06)
(500,500); (800,200)-(0,0)	29/32 (91%)	28/30 (93%)	-0.39 (.70)
(400,1200); (400,200)-(0,0)	23/26 (88%)	25/25 (100%)	-1.75 (.08)
Aggregated total	**190/324 (59%)**	**149/266 (56%)**	**0.64 (.52)**

A play	A's entering		Z (two-tailed p-value)
	Role Reversal	No Role reversal	
(100,1000); (125,125)-(75,125)	16/32 (50%)	14/26 (54%)	-0.29 (.77)
(450,0); (450,350)-(350,450)	12/32 (38%)	11/27 (41%)	-0.25 (.80)
(800,0); (400,400)-(0,800)	7/22 (32%)	11/30 (37%)	-0.36 (.72)
(550,550); (750,375)-(400,400)	3/22 (14%)	2/25 (8%)	0.62 (.54)
(0,800); (400,400)-(0,800)	32/32 (100%)	11/11 (100%)	0.00 (1.00)
(500,500); (800,200)-(0,0)	19/32 (59%)	16/30 (53%)	0.48 (.63)
(400,1200); (400,200)-(0,0)	6/26 (23%)	10/25 (40%)	-1.30 (.19)
Aggregated total	**95/198 (48%)**	**75/174 (43%)**	**0.94 (.35)**

Table B2 – The Effect of Expressing a Preference on A Play

Games where entering is a favorable play	A's entering		Z (one-tailed p-value)
	Preference	Silence	
(450,0); (450,350)-(350,450)	14/25 (56%)	11/27 (41%)	1.10 (.14)
(800,0); (400,400)-(0,800)	6/26 (23%)	11/30 (37%)	-1.10 (.86)
(800,0); (750,375)-(400,400)	11/25 (44%)	9/27 (33%)	0.79 (.21)
A(700,200); B(600,600)-(200,700)	6/20 (30%)	14/32 (44%)*	-0.99 (.84)
A(450,0); B(450,350)-(350,450)	9/29 (31%)	11/36 (31%)	0.04 (.02)
A(750,100); B(700,500)-(300,600)	7/30 (23%)	3/36 (8%)	1.69 (.05)
Total for favorable play	**53/155 (34%)**	**59/188 (31%)**	**0.55 (.29)**
Games where entering is an unfavorable play	A's entering		Z (two-tailed p-value)
	Preference	Silence	
(550,550); (750,375)-(400,400)	7/27 (26%)	3/25 (12%)	1.27 (.20)
(100,1000); (125,125)-(75,125)	14/30 (47%)	14/26 (54%)	-0.54 (.59)
(550,550); (750,400)-(400,400)	8/30 (27%)	5/26 (19%)	0.66 (.51)
(700,1300); (800,200)-(0,0)	5/12 (42%)	9/25 (36%)	0.33 (.74)
(400,1200); (400,200)-(0,0)	4/25 (16%)	10/25 (40%)	-1.89 (.06)
Total for unfavorable play	**38/124 (31%)**	**41/127 (32%)**	**-0.28 (.78)**

Barcelona games are italicized.

*Indicates results from Charness and Rabin (2002)

Table B3 – The Effect of Unfavorable Expressed Preferences on Responder Play

A has made a favorable play	B's helping A (by A preference)		Z (two-tailed p-value)
	Don't Help Me	No Preference	
(450,0); (450,350)-(350,450)	4/25 (16%)	7/27 (26%)	0.88 (.38)
(800,0); (400,400)-(0,800)	2/25 (8%)	9/30 (30%)	2.03 (.04)
(800,0); (750,375)-(400,400)	3/25 (12%)	9/27 (33%)	1.82 (.06)
A(450,0); B(450,350)-(350,450)	7/27 (26%)	2/36 (6%)	-2.29 (.02)
A(750,100); B(700,500)-(300,600)	7/28 (25%)	9/36 (25%)	0.00 (1.00)
Aggregated total	**23/130 (18%)**	**51/233 (23%)**	**1.12 (.26)**

A has made an unfavorable play, B can punish at no cost	B's hurting A (by A preference)		Z (one-tailed p-value)
	Punish Me	No Preference	
(550,550); (750,375)-(400,400)	26/27 (96%)	22/25 (88%)	1.12 (.13)
(100,1000); (125,125)-(75,125)	16/30 (53%)	4/26 (15%)	2.96 (.00)
(550,550); (750,400)-(400,400)	20/26 (77%)	14/26 (54%)	1.75 (.04)
Aggregated total	**62/83 (75%)**	**40/77 (52%)**	**2.99 (.00)**

A has made an unfavorable play, B must pay to punish	B's helping A (by A preference)		Z (one-tailed p-value)*
	Punish Me	No Preference	
(700,1300); (800,200)-(0,0)	0/11 (0%)	4/25 (16%)	1.41 (.08)
(400,1200); (400,200)-(0,0)	3/22 (14%)	4/25 (16%)	0.23 (.41)
Aggregated total	**3/33 (9%)**	**8/50 (16%)**	**0.91 (.18)**

* Here the prediction is less punishment when it is requested.

References

Berg, J., J. Dickhaut & K. McCabe (1995), "Trust, Reciprocity, and Social History," Games and Economic Behavior, **10**, 122-42.

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.

Bolton, G., J. Brandts, and E. Katok (2000), "How Strategy Sensitive are Contributions? A Test of Six Hypotheses in a Two-Person Dilemma Game," Economic Theory, **15**, 367-387.

Bolton, G., J. Brandts, and A. Ockenfels (1998), "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game," Experimental Economics, **1**, 207-219.

Bolton, G. and A. Ockenfels (2000a), "ERC: A Theory of Equity, Reciprocity, and Competition," American Economic Review, **90**, 166-193.

Bolton, G. and A. Ockenfels (2000b), "A Stress Test of Fairness Measures in Models of Social Utility," mimeo.

Brandts, J. and G. Charness (2000), "Hot vs. Cold: Sequential Responses in Simple Experimental Games," Experimental Economics, **2**, 227-238.

Brandts, J. and G. Charness (forthcoming), "Truth or Consequences; An Experiment," Management Science.

Brandts, J. and C. Solà (2001), "Reference Points and Negative Reciprocity in Simple Sequential Games," Games and Economic Behavior, **36**, 138-157.

Cason, T. and V. Mui (1998), "Social Influence in the Sequential Dictator Game," Journal of Mathematical Psychology, **42**, 248-265.

Charness, G. (forthcoming), "Attribution and Reciprocity in an Experimental Labor Market," Journal of Labor Economics.

Charness, G. (2000a), "Self-serving Cheap Talk: A Test of Aumann's Conjecture," Games and Economic Behavior, **33**, 177-194.

Charness, G. (2000b), "Responsibility and Effort in an Experimental Labor Market," Journal of Economic Behavior and Organization, **42**, 375-384.

Charness, G. and M. Rabin (2002), "Understanding Social Preferences with Simple Tests," Quarterly Journal of Economics, **117**, 817-869.

Cooper, R., D. DeJong, R. Forsythe & T. Ross (1992), "Communication in Coordination Games," Quarterly Journal of Economics, **53**, 739-771.

Cox, J. (2000), "Trust and Reciprocity: Implications of Game Triads and Social Contexts," mimeo.

Dufwenberg, M. and U. Gneezy (2000), "Measuring Beliefs in an Experimental Lost Wallet Game," Games and Economic Behavior, **30**, 163-182.

Dufwenberg, M. and G. Kirchsteiger (1998), "A Theory of Sequential Reciprocity," mimeo.

Engelmann, D. and M. Strobel (2002), "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," mimeo.

Falk, A. and U. Fischbacher (1998), "A Theory of Reciprocity," mimeo.

Falk, A., E. Fehr & U. Fischbacher (forthcoming), "On the Nature of Fair Behavior," Economic Inquiry.

Falk, A., E. Fehr & U. Fischbacher (2001), "Informal Sanctions," mimeo.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993), "Does Fairness Prevent Market Clearing? An Experimental Investigation," Quarterly Journal of Economics, **108**, 437-459.

Fehr, E., A. Klein & K. Schmidt (2001), "Fairness, Incentives and Contractual Incompleteness" mimeo.

Fehr, E. and K. Schmidt (1999), "A Theory of Fairness, Competition, and Cooperation," Quarterly Journal of Economics, **114**, 769-816 according to cover; 817-868 in truth.

Glasnapp, D. and J. Poggio (1985), Essentials of Statistical Analysis for the Behavioral Sciences, Columbus: Merrill.

Güth, W., R. Schmittberger & B. Schwarze (1982), "An Experimental Analysis of Ultimatum Game Bargaining," Journal of Economic Behavior and Organization, **3**, 367-388.

Hannan, L., J. Kagel & D. Moser (2002), "Partial Gift Exchange in an Experimental Labor Market: Impact of Subject Population Differences, Productivity Differences, and Effort Requests on Behavior," Journal of Labor Economics, **20**, 923-951.

Kahneman, D., J. Knetsch & R. Thaler (1986), "Fairness and the Assumptions of Economics," Journal of Business, **59**, S285-S300.

McCabe, K., M. Rigdon, and V. Smith (2000), "Positive Reciprocity and Intentions in Trust Games," mimeo.

Offerman, T. (2002), "Hurting Hurts More than Helping Helps," European Economic Review, **46**, 1423-1437.

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