Some Simple Tests of Social Preferences

Gary Charness
Department of Economics, Universitat Pompeu Fabra, Barcelona

Matthew Rabin
Department of Economics, University of California—Berkeley

September 12, 2000

Abstract: Departures from pure self-interest in economic experiments have recently inspired models of "social preferences". We conduct experiments on simple games that test these theories more directly than the array of games conventionally considered. Our experiments show support for the prevalence of "quasi-maximin" preferences: People are self interested, but also sacrifice to increase the payoffs for all recipients, especially low-payoff recipients. We show that quasi-maximin preferences better capture behavior than recently proposed models that posit an aversion to differences in payoffs. People are also motivated by reciprocity: While there is little evidence of sacrifice to reciprocate good behavior beyond what they would sacrifice for neutral parties, they withdraw willingness to sacrifice to achieve a fair outcome when others are themselves unwilling to sacrifice, and sometimes punish unfair behavior. We propose some simple models based on our experimental results.

Keywords: Difference Aversion, Fairness, Inequity Aversion, Maximin Criterion, Non-Ultimatum Games, Reciprocal Fairness, Social Preferences, Ultimatum Games.

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

Acknowledgments: This paper is a substantially revised version of "Social Preferences: Some Simple Tests and a New Model." We thank Jordi Brandts, Antonio Cabrales, Colin Camerer, Martin Dufwenberg, Ed Glaeser, Brit Grosskopf, Ernan Haruvy, John Kagel, George Loewenstein, Rosemarie Nagel, Chris Shannon, an anonymous referee, and seminar participants at Harvard, Stanford GSB, Berkeley, UCSD, the June 1999 MacArthur Norms and Preferences Network meeting, the 1999 Russell Sage Foundation Summer Institute in Behavioral Economics, the March 2000 Public Choice meeting, and the April 2000 Experimental Symposium at Technion for helpful comments. We also thank Davis Beekman, Kitt Carpenter, David Huffman, Chris Meissner, and Ellen Myerson for valuable research assistance, and Brit Grosskopf and Jonah Rockoff for helping to conduct the experimental sessions in Barcelona. For financial support, Charness thanks the Spanish Ministry of Education (Grant D101-7715) and the MacArthur Foundation, and Rabin thanks the Russell Sage, Alfred P. Sloan, MacArthur, and National Science (Award 9709485) Foundations.

Contact: Gary Charness / Department of Economics and Business / Universitat Pompeu Fabra/ 25-27 Ramon Trias Fargas, 08005 Barcelona, Spain. E-mail: charness@upf.es. Web page: http://www.econ.upf.es/home/charness/. Matthew Rabin / Department of Economics / 549 Evans Hall #3880 / University of California, Berkeley / Berkeley, CA 94720-3880. E-mail: rabin@econ.berkeley.edu. Web page: http://elsa.berkeley.edu/rabin/index.html.

1. Introduction

Participants in experiments frequently choose actions that do not maximize their own monetary payoffs when those actions affect the payoffs of others. People sacrifice money in simple bargaining environments to punish those who mistreat them, share money with other parties who have no say in allocations, and make voluntary contributions to public goods. To capture such departures from narrow self-interest, several models of *social preferences* have recently been proposed. These models assume that people are self interested, but are also concerned about the payoffs of others. In this paper, we report findings from a series of simple experiments that test existing theories more directly than the conventional array of games, and formulate a new model to capture patterns of behavior that previous models don't explain.

Existing models of social preferences fall into two categories: Those that assume people care solely about the distribution of payoffs, and those that assume people are also motivated to reciprocate the intentional actions of others. Loewenstein, Thompson, and Bazerman (1989), Bolton (1991), Bolton and Ockenfels (2000), and Fehr and Schmidt (1999) develop distributional models in which a person is motivated to reduce differences among material payoffs. In these "difference-aversion" models, a player sacrifices to help others when ahead, but also engages in Pareto-damaging behavior—hurting some while helping none—when behind. An alternative hypothesis, related to the ideas discussed in Yaari and Bar-Hillel (1984) and Andreoni and Miller (1998), assumes that people like to increase social surplus and don't dislike differences in payoffs *per se*, but rather care more about helping low-payoff people than high-payoff people. Such "quasi-maximin preferences" do not induce Pareto-damaging behavior: A player wants to help others when behind, though far less so than when ahead.

There are also several models of reciprocity preferences. Rabin (1993) developed a model in which one player wishes to increase or decrease another player's payoffs based on her beliefs about whether the other player is treating her fairly, and Dufwenberg and Kirchsteiger (1998) modify and extend the model so as to be more applicable to sequential games. Falk and Fischbacher (1998) combine reciprocity with difference aversion into a sequential model.

In this paper we present data from a series of simple two- and three-player binary-choice games that will allow researchers to identify social motivations more directly than the existing set

of games studied. One motivation for our research was a concern that many alternative models were being developed based on evidence from games that have little power to identify those models. We were especially skeptical about whether the evidence used to motivate recent difference-aversion models supports those models as strongly as supposed. Our intuition was that both reciprocity and quasi-maximin preferences were a better explanation for observed behavior, and we noticed that difference aversion is confounded with these alternatives as an explanation for behavior in virtually all of the experimental evidence cited.

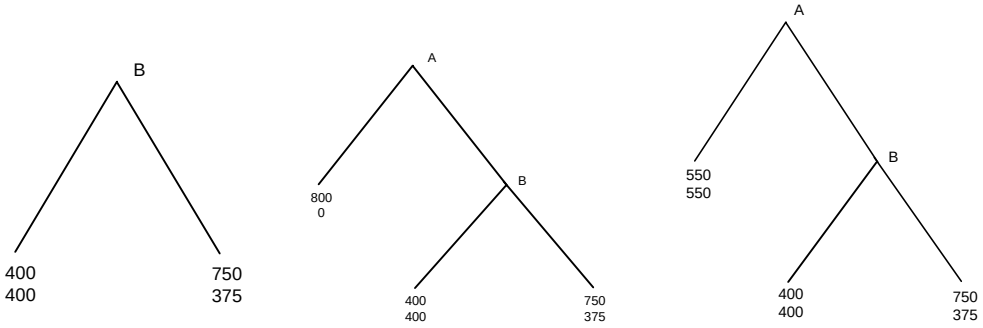
Difference aversion has been used to explain Pareto-damaging behavior, such as rejections in ultimatum games, by positing a reciprocity-free propensity for lowering the payoffs of other players who are getting a higher payoff. But the games used to support this explanation are confounded in two ways: First, these games always involve a clear motivation for negative reciprocity for the subject engaging in Pareto-damaging behavior. Second, the only form of Pareto-damaging behavior in ultimatum games is to reduce inequality, so that we cannot discern whether parties might also engage in Pareto-damaging behavior when it does not reduce inequality.

Difference aversion has also been used to explain helpful sacrifice by subjects, such as helping behavior in public-goods games and prisoner's dilemmas, as a taste for helping those with lower payoffs. We believed that the evidence invoked does indeed support the view that subjects are especially inclined to help those who are getting a lower payoff. This explanation, consonant with both difference aversion and quasi-maximin preferences share, appears to explain most observed helping behavior better than positive reciprocity. But once more there is a clear confound in all evidence invoked to use difference aversion as a theory of helping behavior: The games used to support this theory of sacrifice have only allowed helpful sacrifice that decreases inequality. Hence, we cannot learn whether inequality reduction is a necessary condition for observing helpful sacrifice.

We design simple games allowing Pareto-damaging behavior that variously doesn't occur following retaliation-worthy behavior and doesn't reduce inequality; we also design simple games allowing helpful behavior that doesn't reduce inequality. Our data consist of 29 different games, with 467 participants making 1697 decisions, giving subjects an array of choices that would allow us to isolate different social motivations that might be at play. We find that: 1)

quasi-maximin preferences alone explain our data better than does difference aversion, and that 2) reciprocity also has explanatory power, so that 3) a complete model incorporating both of these preferences does best.

To get a sense for our approach and results, consider Games 1-3:



Game 1 **Game 2** **Game 3**

In "Game" 1, Player B unilaterally determines both his own payoff and Player A's payoff by choosing between two actions yielding payoffs (750,375) and (400,400), where the first entry is A's payoff and the second is B's. Difference aversion says B prefers (400,400) over (750,375)—both on self-interest grounds and because he dislikes coming out behind. We find that about 50% of participants chose (750,375). Similarly, 69% of our subjects chose (750,400) over (400,400).[1] In Section 4, we use summary statistics from such direct tests to show that assuming that subjects have a slight preference for helping others who are getting higher payoffs explains a significantly higher percentage of behavior than assuming a preference for hurting them. We also confirm that subjects are much more willing to sacrifice to help others who are getting a lower payoff than others getting a higher payoff. At the end of Section 4 we present regression analysis that assumes homogenous preferences implemented with some error. Here the taste for helping others receiving a lower payoff once more shows up strongly. The analysis also shows that players are very close to self-interested when behind, confirming that there is relatively little taste for hurting other players but also indicating that the taste for helping others when behind is very weak.

[1] Andreoni and Miller (1998), Charness and Grosskopf (1999), and Kritkos and Bolle (1999) find similar results, with significant numbers of participants opting for inequality-increasing sacrifices to help others.

To test the role of reciprocity, we study simple response games where B's choice follows a move by A to forego an outside option, and compare B's behavior to his behavior given the same binary choice where A either forewent a different outside option or had no option at all. Game 2, for instance, involves the same choice by B as in Game 1, but follows a kind move by A to forego an (800,0) outcome. Yet only 38% chose (750,375), *less* than the proportion choosing this outcome in Game 1. These and our other findings reinforce recent experimental evidence that "positive reciprocity" is not a strong force in experimental settings.[2] One exception we find (based on limited data) to this pattern is that positive reciprocity seems to vastly reduce difference aversion when self-interest is not at stake. When A foregoes a (750,0) allocation, only 6% of Bs choose (400,400) over (750,400), down from the 31% rate in the dictator game. We return to discuss this game in the conclusion because we think it raises a serious challenge to current social-preferences models.

Other forms of reciprocity show up more strongly. Subjects exhibited a form of reciprocity we call *concern withdrawal:* They withdraw their willingness to sacrifice to allocate the fair share towards somebody who himself is unwilling to sacrifice for the sake of fairness. Consider Game 3, where A first chooses between payoffs of (550,550) or to allow B the same choice as in Games 1 and 2. A is clearly being unkind by not choosing (550,550), and here only 10% of Bs chose (750,375) over (400,400). Subjects also exhibited negative reciprocity in Pareto-damaging behavior. For instance, 0 out of 36 Bs chose (0,0) over (800,200) when neutral towards A, whereas 10% (6 out of 58) chose it following a decision by A to forego an even split. We were surprised that in many of our experiments so few participants punished others even at little or no cost of doing so. But the difference here is statistically significant, and is presumably due to negative reciprocity. In Section 5, we demonstrate that reciporcity plays a role in B's behavior by providing summary statistics on how B's propensity for helping and Pareto-damaging behavior depends on A's behavior. We also show with regression analysis that negative reciprocity is a statistically significant component of the weight B puts on A's payoff.

---

[2] We note in passing that this lack of positive reciprocity is consistent with results from trust game (e.g., Berg, Dickhaut, and McCabe 1995), and gift-exchange games, which are often interpreted as positive reciprocity. Suppose the first mover sends his entire allocation to the other player, so that the interim allocation is (0,40). Since many people in dictator games allocate more than 25% to the other player, it is not an indication of positive reciprocity that many responders send back 10 or more.

While most of our data and our formal tests concern two-player games, in Section 6 we discuss the results in the five three-player games we also studied. These games provide some evidence for a multi-person generalization of quasi-maximin preferences, and reinforce the role of reciprocity by showing that subjects' preference between two allocations is for the one where an unfair first mover gets a lower payoff. We also use one of our games to provide a direct and clear demonstration that subjects are not indifferent to the distribution of material payoffs among other people, which is a hypothesis forcefully argued by Bolton and Ockenfels (1998, 2000).

Our experiments provide raw data for building new models of social preferences, at differing levels of complexity representing different tastes for parsimony vs. accuracy. In Section 2, we provide a simple linear, two-person model of preferences that posits a person's preferences are a weighted sum of his own and the other player's preferences, where this weighting depends on who has the higher payoff. This embeds difference aversion, quasi-maximin preferences, and other preferences such as competitiveness as identically parsimonious and tractable special cases of a more general model. This formulation provides the basis for our direct comparative tests of these distributional models. We also provide a clear articulation of the reciprocity hypotheses by integrating a shift in these weights as a function of the others' behavior. In Section 6, while presenting our results for the three-person games, we provide a multi-person model of quasi-maximin preferences and how it relates to the two-person model from Section 2. In Appendix A, we develop a more complex model of *reciprocal-fairness equilibrium* that integrates quasi-maximin preferences and reciprocity. We presume that players are motivated to pursue quasi-maximin allocations, but withdraw the willingness to give others their quasi-maximin shares when these others are being unfair, and may even sacrifice to punish them.

We have stressed our view that the data indicate that difference-aversion models do not fare well; however, there are two reasons while we think that our evidence is quite far from conclusive. First, some of the differences from earlier research in both our design and in our results—especially the relative lack of Pareto-damaging behavior—demand caution in extrapolating results from our experiments. Second, it is clear that there are non-trivial numbers of subjects who exhibit difference aversion (or competitive preferences) in some circumstances, indicating that this motivation may influence behavior. Indeed, as Fehr and Schmidt (1999) and elsewhere have argued that only 40% of subjects need be difference-averse to explain the

phenomena they explain, and our data do not suggest this is an implausible figure. Hence, there may be nothing inconsistent in our results with previous explanations.

Nonetheless, we suspect our evidence does suggest that previous analyses have been misleading. The degree of willingness to sacrifice to impose equality needed to explain observed rejections in ultimatum games seems to require far more than the amount observed; to our knowledge, nobody has identified a significant willingness by laboratory subjects to sacrifice money to hurt others when negative reciprocity is not implicated. Moreover, it bears mentioning that our analysis suggests that the majority of subjects who don't exhibit difference aversion are not merely self-interested, but when behind exhibit preferences opposite to difference aversion. We believe these other preferences are both more prevalent and play a prominent role in many games that, while not the focus of recent experimental work, are likely of great economic interest. Once more, the narrow range of games which have probably misleadingly led researchers to emphasize one subset of subjects among a heterogeneous subject pool as playing a more important role than they actually do.

In this light, we believe the main contribution of this paper is not in firmly establishing that previous interpretations have been wrong—we don't think we have proven that point—but rather in clarifying the confounds in the previous research supporting those interpretations. Indeed, beyond our specific findings, we hope this paper helps move experimental research away from testing hypotheses solely on variants of the existing, familiar menu of experimental games. Direct and unconfounded tests of models, using a wide range of simple games, will accelerate understanding of social motivations in experimental settings. We conclude in Section 7 with a discussion of some of these issues and with some suggestions for new directions of research.

2. Hypotheses about Social Preferences

Most of this paper compares different two-player distributional models of social preferences. These different models can be represented as hypotheses about the parameter values in a simple formula for preferences. Player B's preferences, $U_B(\ )$, can be represented in terms of her own material payoffs, π_B, and Player A's material payoffs, π_A:

$$U_B(\pi_A,\pi_B) \equiv \pi_B + \rho\bullet(\pi_A-\pi_B) \equiv (1-\rho)\pi_B + \rho\pi_A \text{ when } \pi_B \geq \pi_A,$$
$$U_B(\pi_A,\pi_B) \equiv \pi_B + \sigma\bullet(\pi_A-\pi_B) \equiv (1-\sigma)\pi_B + \sigma\pi_A \text{ when } \pi_B \leq \pi_A.$$

This formulation says that a person's utility is a linear function of her own material payoff and the difference between her own payoff and the second person's, possibly weighting this difference differently when doing better or worse than the other person.

One form of distributional preferences (consistent with the psychology of status) is simple competitive preferences. These can be represented by assuming $\sigma \leq \rho < 0$, meaning Player B always prefers to do as well as possible in comparison to A, while also caring directly about her payoff. That is, people always like their payoffs to be as high relative to others' as possible.[3] Assuming $\sigma \leq \rho$ says that the preferences for gains relative to the other person is at least as high when behind as when ahead.

A more prevalent hypothesis about distributional preferences is what we call "difference aversion," and is exemplified by Loewenstein, Bazerman, and Thompson (1989), Bolton and Ockenfels (2000), and Fehr and Schmidt (1999). This approach is related to equity theory as classically formulated, as these models assume that people prefer to minimize disparities between their own monetary payoffs and those of other people. Difference aversion corresponds to $\sigma < 0 < \rho < 1$. That is, B likes money (as with all the preferences we discuss), and prefers that payoffs are equal, including wishing to lower A's payoff when A does better than B.[4] Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) show that difference aversion can match experimental data in ultimatum games, public-goods games, and some other games where many subjects sacrifice to prevent unequal payoffs.

Yet there is considerable experimental evidence that does not match these models. Andreoni and Miller (1998), for instance, test a menu of simple dictator games where many subjects give money to subjects already getting more money, which is the opposite of difference aversion. Moreover, they interpret participants who equalize payoffs to be pursuing (what we are

[3] Fehr and Schmidt (1999) discuss the possibility of competitive preferences as an alternative to difference aversion.
[4] Fehr and Schmidt (1999) add the restriction that $-\sigma > \rho$, so that B is more bothered by getting less than A than by getting more than A. We shall not impose that restriction when testing difference aversion.

calling) maximin preferences rather than difference aversion. Our notion of "quasi-maximin preferences" subsumes the different cases examined by Andreoni and Miller (1998), by letting the parameters take on the values $1 \geq \rho \geq \sigma > 0$. It is also natural to impose $\sigma \leq 1/2$, which says that B is not more concerned about A's payoff than his own when A is getting a higher payoff.[5] Here, subjects always prefer more for themselves and the other person, but are more in favor of getting payoffs for themselves when they are behind than when they are ahead. Quasi-maximin preferences are the two-player case of the more general notion, related to the ideas presented in Yaari and Bar Hillel (1984), that players want to help all players (especially themselves), but are particularly keen to help the person who is worst off.[6]

Since quasi-maximin preferences assume that people always prefer Pareto-improvements, they cannot however explain rejections in the ultimatum game or other Pareto-damaging behavior. Of course, reciprocity is a natural alternative explanation for Pareto-damaging behavior. Several models have assumed that players derive utility from reciprocal behavior, so are motivated to treat those who are fair better than those who are not.[7] Roughly put, these models say that B's values for ρ and σ vary with B's perception of player A's intentions.

Any reciprocal model must embed assumptions about distributional preferences. Rabin (1993) and Dufwenberg and Kichsteiger (1998) concentrated on modeling the general principles of reciprocity, and employed simplistic notions of fairness and distributional preferences. Falk and Fischbacher (1998) combine difference aversion and reciprocity into a model where a person is less bothered by another's refusal to come out on the short end of a split than by a refusal to share equally. Roughly put, they assume that B has preferences $\sigma < 0 < \rho < 1$ when they feel neutrally or positively towards another person, but that B's values for ρ and σ diminish if A's behavior suggests that A assigns the weight $\rho \leq 0$ to B's well-being. Importantly, Falk and Fischbacher (1998) assume that B does not resent harmful behavior by A if it seems to come only from A's unwillingness to come out behind rather than A's selfishness when ahead. That is, B

---

[5] Note that when $\rho = \sigma = 1/2$, $U_B(\pi_A,\pi_B) = (\pi_A+\pi_B)/2$, so that B puts equal weight on each player's material reward.

[6] A fourth possibility (which could be labeled "equity aversion") that also fits into our framework would be to assume a person puts more weight on a person when that person is ahead rather than behind.

[7] Studies demonstrating reciprocity that cannot be explained by distributional models include Kahneman, Knetsch, and Thaler (1986), Blount (1995), Offerman (1998), Charness (1996), Brandts and Charness (1999), Andreoni, Brown, and Vesterlund (1999), and Kagel and Wolfe (1999). Other studies, such as Bolton, Brandts, and Katok (1997) and Bolton, Brandts, and Ockenfels (1997), yield more equivocal or negative evidence regarding reciprocity.

retaliates against behavior implying that A's ρ is too small, but not against behavior indicating $\sigma < 0$.

An alternative hypothesis about reciprocal preferences follows naturally from quasi-maximin preferences. We assume that people have preferences $1 \geq \rho > \sigma > 0$ when they feel positively or neutrally towards other players, but when these others pursue self-interest at the expense of quasi-maximum preferences then they decrease the weights they put on the other players by a parameter Δ, yielding the equations:

$$U_B(\pi_A,\pi_B) \equiv (1-(\rho+\Delta q))\pi_B + (\rho+\Delta q)\pi_A \text{ when } \pi_B \geq \pi_A,$$
$$U_B(\pi_A,\pi_B) \equiv (1-(\sigma+ \Delta q))\pi_B + (\sigma+ \Delta q)\pi_A \text{ when } \pi_B \leq \pi_A.$$

Here $q = 1$ if A has "misbehaved" (violated the dictates of quasi-maximin preferences) by entering, and $q = 0$ otherwise. In our analysis in this paper, we discuss what the evidence suggests about which values of ρ and σ (and Δ, for B responding to A) can best organize the subjects' observed behavior.

We develop a full model combining quasi-maximin motivations and intentions-based reciprocity in Appendix A, where we formally introduce the notions of quasi-maximin equilibrium, which does not incorporate reciprocity, and the far more complicated *reciprocal-fairness equilibrium*. The latter notion assumes that players are motivated by QMM preferences, but abandon the desire to give the fair, QMM allocation to a player when that player is herself pursuing her self-interest rather than the quasi-maximin allocation. We also include the possibility that a player may go further in response to unjustified self-interested behavior by another, and sacrifice to punish her.

We prove that the set of reciprocal-fairness equilibria is non-empty for all parameter values and for all games, and that, for certain parameter vectors, every quasi-maximin equilibrium is a reciprocal-fairness equilibrium.

3. Experimental Procedures and Results

We report data from a series of experiments in which participants made from two to eight choices, and knew that they would be paid according to the outcome generated by one or two of their choices, to be selected at random.

A total of 14 experimental sessions were conducted at the Universitat Pompeu Fabra in Barcelona, in October and November 1998, and University of California-Berkeley, in February and March 1999.[8] There were 319 participants in the Barcelona sessions and 148 participants in the Berkeley sessions. No one could attend more than one session. Average earnings were around $9 in Barcelona and $16 in Berkeley, about $6 and $11 net of the show-up fee paid. In Barcelona, 100 units of lab money = 100 *pesetas*, equivalent to about 70 cents at the contemporaneous exchange rate; in Berkeley, 100 units of lab money = $1.00. Experimental instructions are provided in Appendix B.

We conducted no pilot studies and report all data from experiments played for financial stakes. We also collected survey responses from Barcelona students about how they would behave in hypothetical games, some of which suggested greater difference aversion than for the games we ran for stakes, and hence to contradict our results. We designed the Berkeley games after examining the Barcelona results, and modified several games after observing earlier results.[9]

Students at Pompeu Fabra were recruited by posting notices on campus; most participants were undergraduates majoring in either economics or business. Recruiting at Berkeley was done primarily through campus e-mail lists. Because an e-mail sent to randomly-selected people through the Colleges of Letters, Arts, and Sciences provided most of our participants, the Berkeley sessions included people from a broader range of academic disciplines than is common in economics experiments.[10]

---

[8] Three of the games were each run in two different sessions.

[9] Specifically, Barc4 was designed after the Barc3 results were observed and was chosen to eliminate the possibility that B could believe that A's choice to enter was motivated by an expectation of higher payoffs. In addition, after the 4th Berkeley session we deleted two planned games: 1) A chooses (375,1000) or gives B a (350,350) vs. (400,400) choice, and 2) A chooses (1000,0) or gives B a (800,200) vs. (0,0) choice. We added two games: 1) A chooses (750,750) or gives B a (800,200) vs. (0,0) choice, and 2) A chooses (450,900) or gives B a (400,400) vs. (200,400) choice. With these exceptions, we designed the entire set of games in Barcelona before conducting any experiments, and designed the entire set of Berkeley experiments after we gathered results in Barcelona and before conducting any experiments in Berkeley. We did not use the results of the survey games for design purposes.

[10] As a result of recruiting a smaller number of participants through an advertisement in *The Daily Californian*, our pool of participants also included a few colorful non-students.

Games 5-12 in Barcelona were played in one room, while comparison games were played in a simultaneous session in another room. The groups in the separate rooms were randomly drawn from the entire cohort of people who appeared. Parallel sessions were impractical in Berkeley, but some effort was made to run sessions at similar times of day and days of the week, to make the subject pools in different treatments as comparable as possible.

In all games, either one or two participants made decisions, and decisions affected the allocation to either two or three players. In two-player games, money was allocated to players A and B based either solely on a decision by B, or on decisions of both A and B. In three-player games, money was allocated to players A, B, and C, based either solely on a decision by C, or on decisions by both A and C. Participants were divided into two groups seated at opposite sides of a large room and were given instruction and decision sheets. The instructions were read aloud to the group. Prior to decisions being made in each game, the outcome for every combination of choices was publicly described (on the blackboard) to the players.

In games where more than one player had choices, these were played sequentially. Player A decision sheets were collected, then B decisions were made and the sheets were collected (or, in two cases, A decision sheets were collected, then C sheets). Following Bolton, Brandts, and Ockenfels (1998), Bolton, Brandts, and Katok (2000), and Brandts and Charness (2000), each game was played twice and each participant's role differed across the two plays. Participants were told before their first play that they would be playing in the other role as well, but to discourage reputational motivations, they were assured that pairings were changed in each period.

In games where two people make decisions, first-mover choices were made and decision sheets were collected, then second-player choices were made and these sheets were collected. Except in the case of Games 1-4, participants played more than one game in a session. Games were always presented to the participants one at a time and decision sheets were collected before the next game was revealed. In the sessions with Games 5-12, each participant played two games. In the Berkeley sessions, each participant played four games. Participants knew that the payoffs in only some of the games would be paid, as determined by a public random process after all decisions were made. One of two outcomes was selected in Games 1-4, two of four were selected in Games 5-12, and two of eight were selected in Games 13-32.

Some aspects of our experimental design may discourage comparing our results to those of other experiments. Our use of role reversal and multiple games in sessions may have generated different behavior than had each participant played just one role in one game. In addition, whereas many experiments have players make the same decision repeatedly, we had each participant make each type of decision only once.

Finally, to maximize the amount of data in response games, a responder (B or C) was not told before she made her own decision about the decisions of the first mover (A). The responder 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. This *strategy method* plausibly induces different behavior than does a *direct-response method* in which players make decisions solely in response (when necessary) to other players' decisions.[11] We imagine that there are differences in the two methods, but suspect that the use of the strategy method is not an important factor in our results.

Table 1 reports our results, organizing the games by their strategic structure and the general nature of the trade-offs involved. We label the 12 Barcelona treatments Barc1 to Barc12, where the number indicates the chronological order of the game, and label the 20 Berkeley treatments as Berk13 to Berk32. In parentheses next to the game is the number of participants in the session. The "*x*" in Barc10 and Barc12 signify that C was not told her allocation before her choice, in a design meant to discourage her from comparing A's and B's payoffs to her own.[12]

Two-Person Dictator Games		**Left**	**Right**
Berk29 (26)	B chooses (400,400) vs. (750,400)	.31	.69
Barc2 (48)	B chooses (400,400) vs. (750,375)	.52	.48
Berk17 (32)	B chooses (400,400) vs. (750,375)	.50	.50
Berk23 (36)	B chooses (800,200) vs. (0,0)	1.00	.00
Barc8 (36)	B chooses (300,600) vs. (700,500)	.67	.33
Berk15 (22)	B chooses (200,700) vs. (600,600)	.27	.73

[11] See Roth (1995, p. 323) for a hypothesis for why it might matter, and Cason and Mui (1998) and Brandts and Charness (2000) for tests where it doesn't seem to matter much. Shafir and Tversky (1992) and Croson (2000) find that people cooperate in a Prisoner's Dilemma more frequently when they are unaware of the other player's choice than when they know that the other player has cooperated (or defected). However, this reflects the effect of uncertainty, rather than differences in contingent responses and direct responses to previous play

[12] We took pains to ensure that participants did not think that their behavior influenced *x*. Participants were told that the actual value of *x*, to be revealed at the end of the experiment, was written on the back of a piece of paper that was visibly placed on a table and left untouched until the end of the experiment. *x* was actually 500.

Berk26 (32)	B chooses (0,800) vs. (400,400)	.78	.22

Two-Person Response Games—B's Payoffs Identical		**Out**	**Enter**	**Left**	**Right**
Barc7 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	.47	.53	.06	.94
Barc5 (36)	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	.39	.61	.33	.67
Berk28 (32)	A chooses (100,1000) or lets B choose (75,125) vs. (125,125)	.50	.50	.34	.66
Berk32 (26)	A chooses (450,900) or lets B choose (200,400) vs. (400,400)	.85	.15	.35	.65

Two-Person Response Games—B's Sacrifice Helps A		**Out**	**Enter**	**Left**	**Right**
Barc3 (42)	A chooses (725,0) or lets B choose (400,400) vs. (750,375)	.74	.26	.62	.38
Barc4 (42)	A chooses (800,0) or lets B choose (400,400) vs. (750,375)	.83	.17	.62	.38
Berk21 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,375)	.47	.53	.61	.39
Barc6 (36)	A chooses (750,100) or lets B choose (300,600) vs. (700,500)	.92	.08	.75	.25
Barc9 (36)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	.69	.31	.94	.06
Berk25 (32)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	.62	.38	.81	.19
Berk19 (32)	A chooses (700,200) or lets B choose (200,700) vs. (600,600)	.56	.44	.22	.78
Berk14 (22)	A chooses (800,0) or lets B choose (0,800) vs. (400,400)	.68	.32	.45	.55
Barc1 (44)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.96	.04	.93	.07
Berk13 (22)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.86	.14	.82	.18
Berk18 (32)	A chooses (0,800) or lets B choose (0,800) vs. (400,400)	.00	1.00	.44	.56

Two-Person Response Games—B's Sacrifice Hurts A		**Out**	**Enter**	**Left**	**Right**
Barc11 (35)	A chooses (375,1000) or lets B choose (400,400) vs. (350,350)	.54	.46	.89	.11
Berk22 (36)	A chooses (375,1000) or lets B choose (400,400) vs. (250,350)	.39	.61	.97	.03
Berk27 (32)	A chooses (500,500) or lets B choose (800,200) vs. (0,0)	.41	.59	.91	.09
Berk31 (26)	A chooses (750,750) or lets B choose (800,200) vs. (0,0)	.73	.27	.88	.12
Berk30 (26)	A chooses (400,1200) or lets B choose (400,200) vs. (0,0)	.77	.23	.88	.12

Three-Person Dictator Games		**Left**	**Right**
Barc10 (24)	C chooses (400,400,x) vs. (750,375,x)	.46	.54
Barc12 (22)	C chooses (400,400,x) vs. (1200,0,x)	.82	.18
Berk24 (24)	C chooses (575,575,575) vs. (900,300,600)	.54	.46

Three-Person Response Games		**Out**	**In**	**Left**	**Right**
Berk16 (15)	A chooses (800,800,800) or lets C choose (100,1200,400) or (1200,200,400)	.93	.07	.80	.20
Berk20 (21)	A chooses (800,800,800) or lets C choose (200,1200,400) or (1200,100,400)	.95	.05	.86	.14

Table 1 – Game-by-Game Results

This array of games was chosen to provide a broad range of simple tests that have some power to differentiate among various social preferences. The seven dictator games isolate distributional preferences from reciprocity concerns, and variously allow a responder to sacrifice to decrease inequality through Pareto-damaging behavior, to sacrifice to increase inequality and total surplus, and to affect inequality at no cost to himself. These provide a useful range upon which to test the value of ρ and σ.

The twenty response games have an even wider range of options by B and a wide range of options by A. There are games where entry by A hurts B and where entry helps B, and where this help or harm is or isn't compatible with difference aversion or quasi-maximin preferences. We use these games as further tests of the distributional models by examining both B and A behavior, and can examine reciprocity by seeing how B's response depends on the choice A has made. To aid inferences about reciprocity, we have many sets of games where B's choices are identical, but A's prior choice (or lack thereof) is varied.

In the next two sections, we analyze our results to highlight our central findings as they pertain to the hypotheses discussed in the previous section. In the process of providing general analysis, we will gloss over many plausibly important issues and alternative hypotheses about what explains the behavior we observe in particular games. Many games invite many alternative hypotheses, and our analysis does not do justice to the full range of plausible explanations of what is motivating players in these games. While it is of course somewhat arbitrary to compare models on this set of games, this set clearly offers a greater variety of games than much of the previous literature. For each pair of hypotheses about social preferences, we have games where these preferences make different predictions. We cannot define a "fair" test of the different distributional preferences because we do not know the most appropriate array of games to study. Our primary goal in our experimental design was to create a very diverse list of games giving scope for the widest array of social motivations to play out, and providing scope for the models to fail.[13]

4. Comparison of Distributional Models

In this section, we compare the explanatory power of self-interest and the distributional models (competitive, difference-averse, and quasi-maximin) with respect to our data. We mostly consider how many observations in our games are consistent with the values of ρ and σ permitted

[13] In Charness and Rabin (1999), we provide endless play-by-play commentary interpreting the results, emphasizing especially how the selection of games we chose might affect our overall results, and discuss how hypotheses we are arguing against could be reconciled with the observed behavior.

by the restrictions for each type of social preferences, without placing any further restrictions on the specific value.[14] This approach accommodates any parameter values within the relevant range restrictions, permitting individual heterogeneity for these values without estimating specific values for these parameters. At the end of the section we analyze the data by positing fixed underlying preferences which subjects implement with error, estimating the best-fit values of ρ and σ.

Table 2 shows the explanatory power of various models, under the appropriate restrictions for ρ and σ. These particular statistics are obviously highly dependent on the set of games chosen.[15]

[14] In the 19 two-person games where both players make a decision, each participant makes a choice (in separate cases) as both a first-mover and a responder. Tracking each person's combination of play might tell us something about both participants' beliefs about other players' choices, and the motivations behind their own choices. This is a potentially important source of evidence, and we present the data in Appendix C. We discuss this data in Charness and Rabin (1999). Beyond showing that behavior in the A role is correlated with behavior in the B role, we found relatively little of interest. Observed correlations appeared typically to be compatible with many different models.

[15] Our determination of which choices are consistent with which models, upon which we base the following statistics, is shown in Appendix D. Because we include narrow self-interest as a special case of each of the other distributional preferences, the number of choices consistent with any of these classes of preferences will be at least as large as the number consistent with narrow self interest in games without exact ties. In the many games in which B's payoffs for his two options are the same, however, each of these models is a restriction on self-interest, and hence the numbers we report are variously larger and smaller than the numbers for narrow self interest.

	Total # Observations	Narrow Self interest	Competitive	Difference Aversion	Quasi-Maximin
B's behavior in the dictator games	232	158 (68%)	140 (60%)	175 (75%)	224 (97%)
B's behavior in the response games	671	532 (79%)	439 (65%)	510 (76%)	612 (91%)
B's behavior in all games	903	690 (76%)	579 (64%)	685 (76%)	836 (93%)
A's behavior, any predictions	671	636 (94%)	579 (86%)	671 (100%)	661 (99%)
A's behavior, correct predictions	671	466 (69%)	488 (73%)	603 (90%)	649 (97%)
All behavior, any predictions by A	1574	1326 (84%)	1158 (74%)	1356 (86%)	1497 (95%)
All behavior, correct predictions	1574	1156 (73%)	1067 (68%)	1288 (82%)	1485 (94%)

Table 2 – Consistency of Behavior with Distributional Models

As we are not yet considering reciprocity motivations, which may influence preferences in response games, it is most appropriate to make comparisons using only the seven dictator games. The first line indicates that QMM preferences are far more effective than the others in explaining behavior when reciprocity issues are absent.

Discussing some individual dictator games provides some intuition for our findings. Berk29, in which B chooses between (750,400) and (400,400), shows that a substantial number of subjects refuse to receive less than another person when such refusal is costless, and provides the strongest evidence in our data for difference aversion. But in Berk29 and elsewhere we never observe more than 1/3 of people exhibiting *any* degree of difference aversion. Note that an exact tie in B's payoff provides the best possible chance of revealing any degree of difference aversion, since it eliminates self-interest and everything else as a countervailing motive. Berk23, where 0

of 36 B's chose (0,0) over (800,200), was an attempt to test the willingness of participants to reject offers of the sort rejected in many ultimatum-game experiments, but in a reciprocity-free context. There is obviously no support for difference aversion in this experiment. However, inducing negative reciprocity motives for B making the same choice did not lead to very high rejection rates, so Berk23 provides only limited evidence that punishment in the ultimatum games doesn't come from difference aversion.

The remaining two-player dictator games examine B's willingness to sacrifice to help A. Barc2 and Berk17, where B chooses between (400,400) and (750,375) provide a challenge to difference aversion. About one half of B's sacrifice money to *increase* their deficit with respect to A. Berk8 and Berk15 both show significant willingness by B to help A, where this help is consistent with both difference aversion and QMM. The contrast in behavior between Barc8 and Berk15 shows that Player B is far less willing ($p \approx .00$) to sacrifice 100 to help A by 400 when by doing so she receives a lower payoff than A.[16] A higher proportion of B's take a 100% share in Berk26 than in traditional dictator experiments. But the 22% rate observed for even splits is not unusual in a dictator game, and no intermediate split was available.

In this and the other comparisons in Table 2, the proportion of observations explained by quasi-maximin preferences is significantly higher than the proportions explained by the other three types of preferences. Except for the case of unrestricted A behavior, all the comparisons between quasi-maximin preferences and the other three categories would be statistically significant at $p < .0001$ if each observation were treated as independent.[17]

These proportions compare how the distributional models do in explaining all behavior. But when both choices are compatible with a model, its ability to match the data may merely reflect its lack or predictive power. In this light, perhaps a more relevant test is how well a model matches behavior when it makes a unique prediction. Note that, because each model

---

[16] 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. As we generally have a directional hypothesis, the p-value given reflects a one-tailed test. Where there is no directional hypothesis, we use a two-tailed test and state that we do so.

[17] If we assume that each individual's choices are only one independent observation, we can calculate a minimum level of statistical significance by dividing the test statistic by $\sqrt{8}$, since we can have as many as eight observations for each individual. Doing so, we find statistical significance at $p < .05$ in each case except for unrestricted A behavior.

embeds self-interest, it makes a unique prediction only when there is an exact tie in payoffs or when the distributional preference matches self-interest. Table 3 shows how each model performs in our data in each class of choices among those choices where the model predicts only one of the two choices is compatible with the model. Again, we see that QMM substantially outperforms the other models.

Class of Games	Narrow Self Interest	Competitive	Difference Aversion	Quasi-Maximin
B's behavior in the dictator games	132/206 (64%)	104/196 (53%)	49/106 (46%)	54/62 (87%)
B's behavior in the response games	346/479 (72%)	319/551 (58%)	350/517 (68%)	304/363 (84%)
B's behavior in all games	478/685 (70%)	423/747 (57%)	399/623 (64%)	358/425 (84%)
A's behavior, any predictions	172/226 (76%)	212/304 (70%)	32/32 (100%)	74/84 (88%)
A's behavior, correct predictions	466/671 (69%)	364/553 (66%)	181/249 (73%)	134/150 (89%)
All behavior, any predictions by A	650/911 (71%)	635/1051 (60%)	431/655 (66%)	432/509 (85%)
All behavior, correct predictions	944/1356 (70%)	787/1300 (61%)	580/872 (67%)	492/575 (86%)

Table 3 – Consistency of Behavior with Distributional Models When the Prediction is Unique
(Entries are chances taken over total chances)

Line 1 shows that QMM clearly outperforms both difference aversion and competitive preferences in dictator games. Of course, one may desire a model that does better than to explain accurately the behavior in dictator games. Distributional models may be appropriate in response games where reciprocity is likely to be aroused, either because reciprocity is relatively weak or

because the models are meant to be proxies for reciprocity.[18] While we discuss the specific findings on the various types of response games in the next section, line 2 of both Table 2 and Table 3 provides aggregate statistics on behavior in response games. Here we see that quasi-maximin preferences and even narrow self-interest outperform difference aversion. Line 3 of both tables shows the aggregate of all B behavior.

While we have emphasized B's behavior in reaching our strongest conclusions, obviously A's behavior may also be motivated by social preferences. Interpreting A behavior is more problematic, since A's perceived distributional consequences of his choice depends on his beliefs about what B will do. One approach is to make no assumptions about what A believes B will do—and say that A's choice is consistent with a distributional preference if his choice is consistent given any belief about what B might do. The strongest and most common—and most tenuous—way to interpret A's choices is to assume that A's correctly anticipated the empirically observed responses by B's and hence that A's made a binary choice between that expected payoff and the payoff from the outside option. Appendix D presents our classification of A's choices in all the two-player response games using each of these two methods, and Tables 2 and 3 assess A's behavior using both methods.

Referring to Table 2, under the liberal interpretation of consistency, few choices by A are entirely inconsistent with any of the models, but clearly difference aversion and quasi-maximin do very well, narrow self interest does a little worse, and competitiveness does relatively poorly. The more restrictive consistency interpretation seems to indicate the superiority of quasi-maximin preferences. However, we urge caution in making this interpretation, as there are more observations where intuitively implausible parameter values are needed to reconcile choices with quasi-maximin equilibrium than with difference aversion.

The behavior by A's in our experiments help shed light on the much-emphasized observation that in ultimatum games proposer behavior is not discernibly inconsistent with narrow self-interest. This is because proposers have an incentive to make generous offers out of fear of having their offers rejected by responders. It is not clear what the generalization of this fact would be beyond the ultimatum game, but the hypothesis that first-mover behavior is

[18] One possibility, for instance, is that difference aversion may not be literally correct, but may be a parsimonious proxy for complicated intentions-based reciprocity models. However, as demonstrated by Tables 2 and 3, and

approximately self-interested is, as with many hypotheses, not sustainable when analyzing games besides the ultimatum game. In our data, 27% of A's take the action that, given actual B behavior, involved an expected sacrifice. By this measure, A behavior is less self-interested than B behavior. While this could, of course, be an artifact of misprediction by A's, note that of A's whose sacrifice helps B, 35% sacrificed, whereas only 15% sacrificed to hurt B's. This difference (179/517 vs. 22/144) is significant at $p \approx .00$. Even more directly, note that in the eight games in which A's decision to enter could only lose her money but could help B, 33% (92/276) sacrificed. In the two cases where entry by A could not help either player, 19% (10/52) entered.[19] Tables 2 and 3 show that departure from self-interest, depending how one measures it, seems just as common for A's as for B's.

The last two rows of Tables 2 and 3 tally up the consistency of all choices in two-player games by adding A's choices to B's choices in the second row, and measuring consistency using each of the two methods discussed above.[20]

Table 4 shows a useful way to parse our results to help see why difference aversion performs poorly, breaking down both Pareto-damaging and helpful behavior by B into its effects on inequality.

---

especially Table 4 below, our experiments call into question even this weaker case for difference aversion.

[19] The eight games where entry could help B are Barc4, Barc6, Barc7, Barc9, Berk14, Berk19, Berk21, and Berk25; the two games where it hurts both are Berk30 and Berk32.

[20] As the number of participants in each game varied, our percentages could be correspondingly distorted by weighting different games differently. Thus, we also checked these percentages by assigning an equal weight to each game (and eliminating duplicate games). We find that the percentages changed very little—with this approach, the penultimate row of Table 2 becomes 84%, 73%, 87%, 94%, and the last row becomes 73%, 67%, 82%, 94%.

Class of Games	Sacrifices/Chances	Probability of Sacrifice
Games allowing Pareto-damage	**59/357**	**17%**
Decreases inequality	34/228	15%
No effect on inequality	4/35	11%
Increases inequality	21/94	22%
Games where sacrifice helps A	**199/546**	**36%**
Decreases inequality	99/212	47%
No effect on inequality	8/68	12%
Increases inequality	92/266	35%
All Games	**268/903**	**30%**
Decreases inequality	133/440	30%
No effect on inequality	12/103	12%
Increases inequality	123/360	34%

Games allowing Pareto damage are: 5, 7, 11, 22, 23, 27, 28, 29, 30, 31, and 32. Games in which a sacrifice helps A are: 1, 2, 3, 4, 6, 8, 9, 13, 14, 15, 17, 18, 19, 21, 25, and 26.

Table 4: B's Sacrifice Rate by Effect on Inequality

Table 4 shows that B's chose Pareto damage in 17% of their opportunities. Calling into question difference aversion as an explanatory variable in Pareto-damaging behavior, in our sample B's are *less* likely to cause Pareto damage when this decreases inequality than when Pareto damage increases inequality. We don't believe this would be the pattern more generally, but we also suspect the role for inequality reduction in punishment behavior has been exaggerated, and our results highlight the overwhelming confound between inequality reduction and Pareto-damaging behavior even in previous research that disentangles Pareto damage from negative reciprocity.

B sacrifices to help A 36% of the time when he has the opportunity to do so. There is a significant relationship ($p \approx .00$) between helping behavior and whether such helping increases or decreases inequality, consistent with the predictions of both difference aversion and quasi-maximin preferences. The fact that, overall, 34% of inequality-increasing opportunities to sacrifice are taken, however, indicates much stronger support for quasi-maximin preferences than for difference aversion, as reflected in the statistics reported in Tables 2 and 3.

A final test of the consistency of our data with different distributional models is to parse results according to how well the different models predict sacrifice behavior, removing all the cases where B is indifferent. This can provide a partial test of the strength of the different social motivations. Table 5 provides such data, and also directly compares quasi-maximin to difference aversion when the two models make differing predictions about sacrifice.

Class of Games	Sacrifices/Chances	Probability of Sacrifice
All games where B can sacrifice	213/737	29%
When Sacrifice is…		
Consistent with Competitive	10/156	6%
Inconsistent with Competitive	203/581	35%
Consistent with DA	108/332	33%
Inconsistent with DA	105/405	12%
Consistent with QMM	191/478	40%
Inconsistent with QMM	22/259	8%
Consistent with DA but not QMM	9/120	8%
Consistent with QMM but not DA	92/266	35%

Games where B can sacrifice are: 1, 2, 3, 4, 6, 8, 9, 11, 13, 14, 15, 17, 18, 19, 21, 22, 23, 25, 26, 27, 30, and 31.

Table 5: Distributional Models as Explanations for B's Sacrifice

It is clear that competitive preferences do a poor job of explaining sacrifices by B. Difference aversion explains sacrifice much better. B sacrifices 40% of the time when doing so is consistent with QMM preferences, but only 8% of the time when a sacrifice is inconsistent with QMM. The last two rows of the Table are revealing, and strongly suggest that QMM preferences play a more prominent role in B's decision to sacrifice money, although in our set of games the average sacrifice needed to promote difference aversion is greater than that need to promote quasi-maximin preferences.

Similar evidence from elsewhere also supports our findings about the relative frequency of behavior consistent with quasi-maximin preferences, difference aversion, and competitiveness.

Charness and Grosskopf (1999) found that while about 33% of subjects chose (Other,Self) allocations of pesetas of (600,600) over (900,600), only about 11% of subjects chose (Other,Self) allocations of (400,600) over (600,600). This suggests that about 1/3 of subjects who chose to equalize payoffs when behind are competitive rather than difference averse. In a variant where each of 108 choosers receives 600 but can choose any payoff for the other person between 300 and 1200, 80 (74%) chose 1200, 8 (7%) chose a number between 600 and 1200, 11 (10%) chose 600, and 9 (8%) chose a number less than 600.

If interpreted as error-free reflections of stable behavior, these experiments that test distributional preferences when no self-interest is at stake indicate that something like 70% of people are quasi-maximin, 20% difference averse, and 10% competitive. Other results from Charness and Grosskopf (1999) in which a small amount of money was at stake are perhaps even more telling. While 67% of 108 subjects chose (Other,Self) payoffs of (1200,600) over (625,625), only 12% chose payoffs of (600,600) over (1200,625). That is, of the two thirds of subjects who had quasi-maximin rather than difference-averse or competitive preferences, virtually all were willing to sacrifice 25 pesetas to implement those preferences. Of the one third of subjects who had either difference-averse or competitive preferences, two thirds were unwilling to sacrifice 25 pesetas to implement those preferences.

Our comparisons of models above assume that all behavior reflects stable underlying preferences of the individual, and then analyzes the frequency of different preferences that can explain the data. 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 estimate the population means for ρ and σ by performing maximum-likelihood estimation on our binary-response data. We use the logit regression

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

to determine the values that best match predicted probabilities of play with the observed behavior, where γ is a precision parameter reflecting sensitivity to differences in utility (see

McFadden 1981). The higher the value of γ, the sharper the predictions—when γ is 0, the probability of either action must be 50%; when γ is arbitrarily large, the probability of the action yielding the highest utility approaches 1. The utility is estimated from the Section 2 equations that excluded reciprocity:

$$U_B(\pi_A,\pi_B) \equiv (1-\rho)\pi_B + \rho\pi_A \text{ when } \pi_B \geq \pi_A,$$
$$U_B(\pi_A,\pi_B) \equiv (1-\sigma)\pi_B + \sigma\pi_A \text{ when } \pi_B \leq \pi_A.$$

We estimate the values in these equations by imposing restrictions on parameters implied by the variety of models that can be encompassed within this framework, using the data for B behavior in all games.[21] Since we have the same number of observations in each case, in addition to observing the estimated value of γ in each of our models, we can compare the log-likelihood values to gain some insight into the explanatory power of the parameters and the models.

This approach allows us to compare models that make different predictions about the parameter values, and to investigate the power of different models and the costs of the restrictions they impose. While the allowance for "noise" in maximizing utility provides a crude proxy for heterogeneity, it does not accommodate the heterogeneity that certainly exists among participant's parameter values.[22] As such, we believe that our regression results provide a strong indication of general patterns in our data and help select among models, but are not adequate for grasping an accurate sense of the relative frequency of preferences that describe subsets of the subjects. In addition, we reiterate that, while we believe we have chosen a broader array of games than any previous papers with which we are familiar, as with all previous empirical tests of social-preferences models, the fitted values for these parameters will be influenced by our selection of games.

Table 6 reports the regression results for different restrictions that we have investigated in this paper and that have appeared previously in the literature:

---

[21] We follow an approach similar to that used in Charness and Haruvy (1999).

Model	Restrictions	ρ	σ	γ	LL
Self-interest	$\rho = \sigma = 0$	-	-	.004 (9.07)	-593.4
Single parameter— "altruism"	$\rho = \sigma$	.212 (7.20)	.212 (7.20)	.005 (8.65)	-574.5
Single parameter— "behindness aversion"	$\rho = 0$	-	.118 (1.76)	.004 (8.53)	-591.5
Single parameter— "charity"	$\sigma = 0$	.422 (25.5)	-	.014 (11.6)	-527.9
Two-parameter ρ,σ model	none	.423 (25.5)	-.014 (-0.73)	.014 (11.6)	-527.7

t-statistics are in parentheses. γ is the precision parameter, and LL is the log-likelihood function.
Games where A's entry is QMM-misbehavior are: 1, 5, 11, 13, 22, 27, 28, 30, 31, and 32.

Table 6: Regression estimates for B behavior without reciprocity (N=903)

In the first line, we report how well the pure self-interest model fits the data. Its rather low level of precision, as measured by either γ or the log-likelihood ratio, serves as a benchmark for the other models. The next three lines report on three different ways of allowing one additional parameter to account for a person's concern for the payoffs of others. On line 2, we investigate "simple altruism"—a model that has been employed sporadically over the years by economists—that says B cares about a weighted sum of his own payoffs and A's payoffs. This model has clear explanatory power beyond the pure self-interest model, lowering the log-likelihood ratio and marginally raising γ. The estimation also confirms that the best-fit single parameter has B putting significant positive weight on A's payoff.

[22] Estimation of separate ρ and σ values for each individual would be rather problematic. The number of observations for each individual (typically 5 or 6 in Berkeley, and 3 in Barcelona) is not much greater than the number of parameters to be estimated, so that the effectiveness of this approach is rather limited.

Lines 3 and 4 examine how well a model would fit if we restricted a person's concern for the other to the case where, respectively, she is behind the other or she is ahead. Line 3 imposes the restriction that $\rho = 0$, accommodating the model developed by Bolton (1991) to match the data in the ultimatum and other bargaining games. The results show that this model 1) does significantly worse than the simple altruism model, and has no significant explanatory power beyond pure self-interest, and 2) that the best fit value of σ is positive, rather than negative as posited by Bolton (1991), but is only marginally significant.[23] As argued in different ways above, this too helps indicate that those models built on the assumption that $\sigma < 0$ do not usefully organize the data on broad sets of games where this hypothesis for Pareto-damaging behavior is not confounded with other explanations. Line 4 tests the "charity" model, which posits that people only care about the payoff of those others who receive less than they do. As can be seen, this model does significantly better altruism or pure self-interest, indicating that there is indeed much less concern for those who are getting a better payoff.

Indeed, Line 5, in which we estimate the linear distributional model without restrictions, explains the data no better than the charity model does. That is, neither the positive concern for others when behind as incorporated into quasi-maximin preferences nor the behindness-aversion preferences as incorporated into the difference-aversion models seem an important explanatory variable when reciprocity is ignored. The estimated value of ρ is virtually identical whether or not σ is included, and the value of σ is very small and insignificantly different than 0.

Overall, the major gains come from allowing ρ to vary independently of σ. In lines 4-6, the LL is much better and the precision is much greater. Line 4 indicates that people tend to be charitable toward those who are less fortunate, but feel differently when such charity would not increase the minimum of the players' material payoffs. Lines 4 and 5 together show that (overall, and absent A's misbehavior) σ is not much of a driving force in our games. Removing the restriction that $\sigma = 0$ gains us very little: Although the likelihood-ratio goes down slightly, a significance test gives $\chi^2 = .54$, far from the 5% significance level of 3.84. Thus, any explanation for nonpecuniary behavior that relies upon σ being typically negative seems inadequate.

---

[23] It seems clear from our other results, however, that the result that σ is significantly greater than 0 is caused by the restriction that $\rho = 0$. In some games either parameter could explain behavior, and hence it appears that σ is reflecting the positive value of ρ in those games.

5. The Role of Reciprocity

In this section we analyze our results in terms of evidence for reciprocity. We designed our experiments so as to have many examples of games with identical choices for B following different choices by A, and in comparison to dictator choices by B. This lets us compare the choice B makes as a function of what choice A made. We first discuss specific games to give an intuitive sense of the behavior observed. We then present some aggregate statistics examining B's response as a function of A's behavior. We conclude the section with a logit regression analysis of the values of ρ and σ. The difference in these parameter values across types of games sheds light on the role of reciprocity in shaping choices.

Examining the games in which B chooses between (750,400) and (400,400) offers an indication of how reciprocity is implicated in responder behavior:

Games With the Choice Between (400,400) and (750,400)		**(400,400)**	**(750,400)**
Berk29 (26)	B chooses (400,400) vs. (750,400)	.31	.69
Barc7 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	.06	.94
Barc5 (36)	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	.33	.67

Barc7 tests the relative strength of positive reciprocity versus difference aversion when self-interest is not implicated. In contrast to the 31% of B's who choose (400,400) in the dictator game Berk29, only 6% do so following a generous move by A.[24] The difference between Barc7 and Berk29 is significant at $p \approx .00$. Again, there is no reason to consider B's choice between (750,400) and (400,400) anything but a strong invitation to B to pursue difference aversion. We

[24] Note that the dictator version was in Berkeley, not Barcelona. While we did not run a (400,400) vs. (750,400) dictator game in Barcelona, the Charness and Grosskopf result of 34% vs. 66% in the (600,600) vs. (900,600) dictator game in Barcelona was nearly identical to the 31% vs. 69% result in Berk29.

show below that positive reciprocity is nowhere else a strong motivation in our data, so that its dominance here over difference aversion seems to indicate that the 30% of behavior consistent with difference aversion (or competitiveness) is not a strong factor when in conflict with other social motivations. The results from Barc5 surprised us, as Bs were no more likely than in Berk29 to choose (400,400). Punishment for the unfair entry by A would be free here, yet is not employed. We do not know if this is a statistical aberration.

Turning to games where B can sacrifice to help A, consider first those games letting B choose between (400,400) and (750,375).

Games With the Choice Between (400,400) and (750,375)		**(400,400)**	**(750,375)**
Barc2 (48)	B chooses (400,400) vs. (750,375)	.52	.48
Berk17 (32)	B chooses (400,400) vs. (750,375)	.50	.50
Barc3 (42)	A chooses (725,0) or lets B choose (400,400) vs. (750,375)	.62	.38
Barc4 (42)	A chooses (800,0) or lets B choose (400,400) vs. (750,375)	.62	.38
Berk21 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,375)	.61	.39
Barc1 (44)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.93	.07
Berk13 (22)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.82	.18

The games in which B chooses between (400,400) and (750,375) provides the starkest illustration of our two main findings about reciprocity. A large percentage of B's here are willing to sacrifice to pursue the quasi-maximin allocation when they feel neutrally towards A's. There is clearly no evidence of positive reciprocity in comparing Barc2 and Berk17 to Barc3, Barc4, and Berk21. B is in fact *less* likely to sacrifice in pursuit of the quasi-maximin outcome following kind behavior by A than in the dictator context (the difference is collectively significant in at two-tiled test at $p \approx .14$). However, we see evidence of *concern withdrawal*: B is likely to withdraw his willingness to sacrifice to give the quasi-maximin allocation to A if A has behaved selfishly. Comparing within subject pools, the percentage of B's that sacrifice to help A following a selfish action drops from 48% to 7% (from Barc2 to Barc1) and from 50% to 18% (from Berk17 to Berk13). These differences are both significant at $p < .01$.

The lack of positive reciprocity is a pattern that also holds for comparing Barc6 to Barc8, the games where B chooses (300,600) vs. (700,500) and (200,700) vs. (600,600).

Games Where B Chooses Between (300,600) and (700,500)		**(300,600)**	**(700,500)**
Barc8 (36)	B chooses (300,600) vs. (700,500)	.67	.33
Barc6 (36)	A chooses (750,100) or lets B choose (300,600) vs. (700,500)	.75	.25

Games Where B Chooses Between (200,700) and (600,600)		**(200,700)**	**(600,600)**
Berk15 (22)	B chooses (200,700) vs. (600,600)	.27	.73
Berk19 (32)	A chooses (700,200) or lets B choose (200,700) vs. (600,600)	.22	.78

The set of games where B chooses between (400,400) and (0,800) provides a confusing picture about the role of positive reciprocity:

Games Where B Chooses Between (0,800) and (400,400)		**(0,800)**	**(400,400)**
Berk26 (32)	B chooses (0,800) vs. (400,400)	.78	.22
Berk14 (22)	A chooses (800,0) or lets B choose (0,800) vs. (400,400)	.45	.55
Berk18 (32)	A chooses (0,800) or lets B choose (0,800) vs. (400,400)	.44	.56

The results from Berk14, where 55% choose (400,400) over (0,800) in contrast to the 22% who choose (400,400) in the dictator game Berk26, significant at $p \approx .01$, would seem to indicate positive reciprocity. But the results from Berk18 call this interpretation into question. We thought B's willingness to sacrifice would be roughly equal to that in the dictator version of the game, but it is much greater, significant at $p \approx .01$.[25] In addition, in Barc9 and Berk25, only 8 of 68 Bs choose (450,350) over (350,450).

Our final grouping of games where B's payoffs are identical were meant to test difference aversion as an explanation of Pareto damage in a simplified form of the ultimatum game:

Games Where B Chooses Between (800,200) and (0,0)		**(800,200)**	**(0,0)**
Berk23 (36)	B chooses (800,200) vs. (0,0)	1.00	.00
Berk27 (32)	A chooses (500,500) or lets B choose (800,200) vs. (0,0)	.91	.09
Berk31 (26)	A chooses (750,750) or lets B choose (800,200) vs. (0,0)	.88	.12

0% (0 of 36) chose the (0,0) outcome outside the context of retaliation, while 6/58 chose (0,0) in the two treatments where retaliation is a motive. The difference between Berk23 and each of the other two games is significant separately at $p < .03$. But together with Barc11 and Barc22, the other games were B can sacrifice to hurt A, we find relatively little negative

[25] The only sense we can make of this is that A has unambiguously stated a preference against the (0,800) payoff, reducing B's ability to rationalize taking everything. However, this is a weak explanation, and we are puzzled by this result.

reciprocity. In all of these games, B has the option to cause Pareto damage following what we felt would be perceived by B as an unfair entry decision by A.

Games where B can punish A for free also show only weak negative reciprocity. As in Barc5, we were surprised by our findings in Berk28 and Berk32. In each case, an apparent "mean" action by A could be punished for free by B, but only about 35% of Bs do so. Doing so contradicts quasi-maximin preferences in Barc5 and both quasi-maximin preferences and difference aversion in Berk28 and Berk32. These are indicative of many of our results: For whatever reason, we observed relatively few instances of retaliatory decreases in others' payoffs unless they benefited the retaliators materially.[26]

As a first pass at summarizing the evidence on reciprocity, Table 7 specifies a distributional parsing of Pareto damage and positive sacrifice in terms of how A has treated B:

Class of Games	Sacrifices/Chances	Probability of Sacrifice
All games allowing Pareto-damage	**59/357**	**17%**
A has helped B	2/36	6%
A has had no play	8/62	13%
A has hurt B	49/259	19%
All games where sacrifice by B helps A	**199/546**	**36%**
A helped B	100/278	36%
A had no play*	88/202	44%
A hurt B in violation of QMM	7/66	11%

*We include Berk18 in this classification, since A play was obvious and universal.

Table 7: B's Response as a function of A's help or harm

Table 7 shows that when A hurts B, B is more likely to hurt A than otherwise and more likely to withdraw willingness to sacrifice to help A. The difference in Pareto-damaging B

[26] Perhaps the way our games are framed has the effect of obscuring the take-it-or-leave-it aspect of the ultimatum here. However, other studies with a foregone payoff design (e.g., Brandts and Solà 1998 and Falk, Fehr, and Fishbacher 1999) should also share this problem, but have higher rejection rates for 80/20 proposals.

behavior when A helps B and when A hurts B is significant at $p \approx .02$; comparing B behavior when A hurts B and when A either has no play or helps B is also significant at $p \approx .02$.

B sacrifices to help A 36% of the time when he has the opportunity to do so. The data support the view that positive reciprocity plays little role in helping behavior, and that negative reciprocity does play a role. The table crystallizes the fact that our data show that a nice prior choice by A is *less* likely to yield nice treatment by B than is no choice by A at all—reducing helping behavior from 44% to 36%. By contrast, when A has hurt B, helping behavior reduces to 11%. Hence, we see that violation of quasi-maximin norms plays a stronger role in determining when a person sacrifices to help another player than it plays in determining when a player sacrifices to harm another. While involving only two games and 66 observations, this last comparison forms part of the basis for our incorporation of "concern withdrawal" as the primary form of reciprocity in our model of the next section.

To give a more precise analysis of the role of reciprocity, we return to our regression-analysis estimation of the best-fit values of ρ and σ, and investigate how these depend on the reciprocity motivation. Using the same technique as in the previous section, we now estimate the utility functions from the Section 2 equations that include a reciprocity component:

$$U_B(\pi_A,\pi_B) \equiv (1-(\rho+\Delta q))\pi_B + (\rho+\Delta q)\pi_A \text{ when } \pi_B \geq \pi_A,$$
$$U_B(\pi_A,\pi_B) \equiv (1-(\sigma+\Delta q))\pi_B + (\sigma+\Delta q)\pi_A \text{ when } \pi_B \leq \pi_A,$$

where Δ is the parameter measuring how "QMM-misbehavior" by A affects B's weight on A's payoff. Table 8 reports the regression results for different restrictions that we have investigated in this paper and that have appeared previously in the literature:

Model	Restrictions	ρ	σ	Δ	γ	LL
Self-interest	$\rho = \sigma = \Delta = 0$	-	-	-	.004 (9.07)	-593.4
ρ,σ model without reciprocity	$\Delta = 0$	.423 (25.5)	-.014 (-0.73)	-	.014 (11.6)	-527.7
"Reciprocal charity"	$\sigma = 0$	.425	-	-.089	.015	-523.7

		(27.9)		(-2.98)	(11.3)	
ρ,σ model with reciprocity	none	.424	.023	-.111	.015	-523.1
		(28.3)	(1.10)	(-3.11)	(11.6)	

t-statistics are in parentheses. γ is the precision parameter, and LL is the log-likelihood function.
Games where A's entry is QMM-misbehavior are: 1, 5, 11, 13, 22, 27, 28, 30, 31, and 32.

Table 8: Regression estimates for reciprocity by B (N=903)

In the first line, we again report the fit of the benchmark pure self-interest model. The next three lines report on three different ways of allowing one additional parameter to account for a person's concern for the payoffs of others. The level of precision (γ) is much higher for each of these reciprocity regressions than for the self-interest model regression.

More importantly, lines 2, 3, and 4 together show that reciprocity (in the form of Δ) plays a much greater role than σ *per se*. When A has misbehaved, the coefficient for σ diminishes significantly and becomes substantially negative.[27] The coefficient for the reciprocity term (Δ) is strongly significant, and the likelihood-ratio test (line 2 vs. line 4) gives $\chi^2 = 9.18$ ($p \approx .00$).[28] By comparison, allowing a nonzero σ (line 4 vs. line 3) does not produce a substantial difference: The other parameter values do not change much, and the likelihood-ratio test gives $\chi^2 = 1.26$ (not significant, $p \approx .27$). Thus, while the pure distributional models have appeal in limited subsets of games, an analysis over a broad range of games indicates that reciprocity considerations are an important component of behavior.

6. Multi-Player Games

[27] Although our definition of "misbehaved" builds on quasi-maximin preferences, we note that results are quite similar when q reflects misbehavior by difference-aversion standards, as built into the reciprocity model developed by Falk and Fischbacher (1998). There are very few instances when the two models make different predictions, and when they do there is nothing in our data that supports our model over their model.

Though we emphasize two-player distributional preferences throughout the paper, we also ran several games with three players, whose results shed light on the issues discussed in previous sections, and on hypotheses specific to three-player games. While the model discussed in Section 2 and tested above relates to two-person games, it is motivated by the more general multi-person model that is outlined in Appendix A. Of special interest in multi-person models are questions about how players feel about changes in the distribution among others' payoffs given their own payoffs. We presume that B cares more about A's payoff when A earns less than when A earns more. This is the two-player projection of the more general notion that (absent negative reciprocity and in addition to self-interest) people like to improve the payoffs of everybody, but are more concerned about raising the payoffs of those with lower payoffs. In simplified and extreme form, they like to maximize the minimum payoff among players.

Barc10 and Barc12 offer a test of people's "disinterested" views of fairness. The results indicate that people care about both the total surplus and the minimum payoff among others. In both cases, many subjects chose to increase total surplus at the expense of minimum payoff, while others chose to maximize the minimum payoff. The results in Barc10 are of special interest in light of our two-player results. Our results above show that about 50% of Bs choose (400,400) over (750,375) is consistent with those subjects different averse, self-interested, or competitive. None of these motivations would explain the choice by Cs to choose (400,400), suggesting that a good proportion of Bs are choosing (400,400) for "disinterested", quasi-maximin reasons rather than just to get more money.

Bolton and Ockenfels (2000) assume that social preferences extend only to the average payoff of all other players, so that people are unconcerned with the distribution of those payoffs. Bolton and Ockenfels (1998, 2000) provide examples from Güth and van Damme (1998) and elsewhere, in which players seem relatively unconcerned with the distribution of payoffs among other parties. Because we did not believe that rejections in the ultimatum game are a manifestation of distributional preferences rather than reciprocity, and more generally found it

---

[28] Although the multiple-observation caveat to statistical significance may still apply, a comparison between the likelihood-ratio tests nevertheless indicates that allowing Δ to be nonzero has a much greater impact than allowing σ to be nonzero.

surprising to posit that subjects were indifferent to the allocations among others, we designed Berk24 as a simple and direct test of their hypothesis.[29]

Berk24 demonstrates that subjects care about the allocation among other parties: 54% of the participants sacrificed 25 to equalize payoffs with each of the other players, without changing the difference (zero) between a player's own payoff and the average of other players. Under the assumption that virtually no participants would (without reciprocal motivations) choose (575,575,575) over (600,600,600), these results support quasi-maximin preferences and Fehr and Schmidt (1999) difference aversion, but reject Bolton and Ockenfels (2000) difference aversion. Since the sacrifice involved is small, it may be hard to say how strong the motive is. In the context of our other results, however, we are not inclined to call it small: 54% is a higher proportion than we found are inclined to sacrifice *nothing* to eliminate disadvantageous inequality against themselves. Hence, our results suggest that people are more concerned about this aspect of the distribution among other players' payoffs than about equalizing the self-other payoffs in the sense captured by difference-aversion models.

Finally, our two three-person response games also offer strong evidence of reciprocity in responder behavior. Berk16 and Berk20 test the explanatory power of distributional preferences versus reciprocity, disentangled from self-interest. In both games, C receives a payoff of 400 regardless of her choice, and has identical choices among the distribution of the other two players' payoffs—1200 and 100, or 1200 and 200. While the difference-aversion models make different predictions in these two games, the evidence shows that all of the models are wrong. Notice that the proportion of C's choosing the 1200/400/100 combination over the 1200/400/200 combination jumped from 14% to 80% when the choice meant A would get the low payoff instead of B. C's were unhappy with A's greed, and chose to give A the lower payoff irrespective of the distributional consequences, punishing A's 83% of the time overall. This difference in distributional preferences is significant at $p \approx .00$. Because the differences in

[29] Kagel and Wolfe (1999) designed a clever variant of a three-person ultimatum game and find a form of insensitivity to third-party allocations when the Bolton and Ockenfels (2000) and the Fehr and Schmidt (1999) models of difference aversion predict high sensitivity to these allocations. The clear interpretation of Kagel and Wolfe's (1999) data is that the observed insensitivity to payoff distributions is due not to the functional form of difference aversion (as claimed by Bolton and Ockenfels 2000), but rather because difference aversion in any form does not explain the behavior they discussed. Nevertheless, the willingness of participants to assign (as a

distributional consequences of behavior were minor, we do not consider this a very discerning test of the general relative strength of distributional vs. reciprocity motivations. Rather, it shows that reciprocity can overwhelm distributional concerns in some circumstances.

Our data on three-person games suggests a more general model that applies to more general games. We now turn to a model that generalizes the simple distributional two-person quasi-maximin preferences presented above. We begin by positing distributional preferences that combine a formulation of a person's "disinterested social ideal" with a specification of the weight the person puts on this social ideal relative to her self-interest.

We denote by $W(\pi_1,\pi_2,\ldots,\pi_N)$ a disinterested social-welfare function, and can write down the utility function, and assume that a person's distributional preferences assign a weight that players put on self-interest versus social interest. Consider Player *i*'s "reciprocity-free" preferences:

$$V_i(\pi_1,\pi_2,\ldots,\pi_N) \equiv (1-\gamma)\cdot\pi_i + \gamma\cdot W(\pi_1,\pi_2,\ldots,\pi_N),$$

where $\gamma \in [0,1]$ measures how much Person *i* cares about pursuing the social ideal vs. pursuing his self interest.[30]

The quasi-maximin criterion is:

$$W(\pi_1,\pi_2,\ldots,\pi_N) = \delta\cdot\text{Min}[\pi_1,\pi_2,\ldots,\pi_N] + (1-\delta)\cdot(\pi_1+\pi_2+\ \ldots\ +\pi_N),$$

where $\delta \in (0,1)$ is a parameter measuring the degree of concern for helping the worst-off person versus maximizing the total social surplus.[31] Setting $\delta = 1$ corresponds to the pure maximin criterion; setting $\delta = 0$ corresponds to total-surplus maximization. Combined with the quasi-maximin social preferences, the function V_i translates into:

$$V_i(\pi_1,\pi_2,\ldots,\pi_N) \equiv (1-\gamma)\cdot\pi_i + \gamma\cdot[\delta\cdot\text{Min}[\pi_1,\pi_2,\ldots,\pi_N] + (1-\delta)\cdot(\pi_1+\pi_2+\ \ldots\ +\pi_N)].$$

---

consequence of a rejection) negative payoffs (in their experiment 2) to innocent third parties also goes against quasi-maximin preferences, and can only be rationalized as a strong willingness to punish the proposer.

[30] Please note that this γ is not at all related to the precision parameter γ in our logit regressions.

Setting $\gamma = 1$ corresponds to purely "disinterested" preferences, in which players care no more (or less) about her own payoffs than others' payoffs, and setting $\gamma = 0$ corresponds to pure self interest. This weight placed on social interests versus self-interest will play a very large role in our analysis; other players' evaluation of Player *i*'s behavior will be measured in terms of how high his γ seems to be.

To see the connection between this model and the two-player specification of Section 2, note that the full model (without reciprocity preferences) reduces to:

$$V_B(\pi_A,\pi_B) \equiv (1-\gamma\delta)\pi_B + \gamma\delta\pi_A \text{ when } \pi_B \geq \pi_A,$$
$$V_B(\pi_A,\pi_B) \equiv \pi_B + \gamma(1-\delta)\pi_A \quad \text{when } \pi_B \leq \pi_A.$$

If we normalize these two equations by dividing by $1 + \gamma(1-\delta)$, so that as in the Section 2 model $V_B = \pi_B$ when $\pi_B = \pi_A$, we see that $\rho = \frac{\gamma}{1+\gamma(1-\delta)}$ and $\sigma = \frac{\gamma(1-\delta)}{1+\gamma(1-\delta)}$.

These have a natural interpretation. When γ increases (meaning B puts more weight on the social good and less on his own material payoffs), both ρ and σ increase. When δ increases (so that B puts relatively more weight on the maximin component and less on total surplus), then ρ increases and σ decreases, both parameters showing more concern for the person who has less, whether this is A or B. Indeed, looking at $\frac{\rho}{\sigma} = \frac{1}{1-\delta}$ makes this even clearer. Increasing ρ and σ together indicates an increase in γ; increasing the ratio indicates an increase in δ.

7. Summary and Conclusion

---

[31] It would be both more realistic and more complicated to assume that people care about not just the lowest payoff, but the full distribution of payoffs, giving more and more weight to the well-being of those with lower and lower payoffs.

This paper continues recent research delineating the nature of social preferences in laboratory behavior. As we have made clear, one of our motivations was to demonstrate that the apparent adequacy of non-reciprocity distributional models in general—and difference-aversion models in particular—has likely been an artifact of the systematic confounds in the narrow range of games used to construct these models. Behavior recently attributed to difference aversion is really attributable to either quasi-maximin preferences or reciprocal preferences. The approach we have taken is to expand the set of games tested, choosing simple games that disentangle and identify players' motives. Although the wealth of data provides some contradictory evidence and puzzles, there are patterns that emerge.[32]

Our data are rich and complicated and we have not analyzed them thoroughly. We have not run individual differences and correlation across games. Heterogeneity of preferences raises many issues. For instance, if it is common knowledge that players share a different norm of fairness, there would be two different directions in which to extend the model. We could assume that one is not angered by another person's behavior that violates one's own preferred norm of fairness, so long as one is convinced that the other person was adhering to a genuine norm of fairness she would hold even if it did not benefit her. Or we could assume that a person is angered whenever others violate his own norm.

Further, our data seem to indicate a dependence on the behavior of others that does not lend itself to any sort of natural reciprocity interpretation. We see some evidence of a *complicity effect*: The mere fact of another player being involved in a decision seems to change a player's behavior, generally in the direction of making him more selfish. Does a person act more favorably when she knows that the other person has had no opportunity for a decision, so that the full responsibility for a final allocation rests with the decider? There is some evidence

[32] Our view that difference aversion is unlikely to prove to be a strong factor in laboratory behavior does not mean that we believe comparable phenomena are unimportant in the real world. Indeed, we suspect the inherent limitations of laboratory experiments prevent full realization of phenomena such as jealousy, envy, and self-serving assessments of deservingness, that are likely to create *de facto* difference aversion in the real world. On the other hand, there is also reason to believe that experimental settings may exaggerate difference aversion since the very nature of the careful, controlled designs and use of monetary rewards makes relative payoffs salient.

suggesting that impulses towards pro-social behavior are diminished when an agent does not feel the full responsibility for an outcome.[33]

We feel that the range of games typically studied has not only been too narrow, but has also typically been too complicated to lend themselves to easy interpretation. One benefit of the sort of simple games we run is that it is easier to discern what subjects believe are the consequences of their actions. But even in our simple games—and inherently in any games with enough strategic structure to make reciprocity motives operative—we could not reach sharp conclusions about the motivations of first movers because we could not be sure how they thought the responders would play. Hence, we feel one avenue for research would be to pay more attention in experimental design to ways to more directly discern participants' beliefs about the intentions or likely behavior of others in their group or session. For example, Dufwenberg and Gneezy (2000) measure both A's expectation about B behavior and B's expectation about the expectation of A; they find B's expectation of A's expectation is positively correlated with B's response.

Although re-reading evidence in the experimental literature and our own experiments indicate that much of the evidence for positive reciprocity has been misidentified fairness preferences or forms of concern withdrawal, we believe positive reciprocity may in fact play an important role in many situations. For example, one hypothesis is that people are willing to sacrifice to achieve the fair outcome, and willing to sacrifice for the sake of positive reciprocity—but *not* willing to sacrifice more to achieve both. This would in turn suggest that we should see reciprocity in contexts where sacrificing is neither required by fairness, nor manifestly in contradiction to fair treatment of oneself.

We hope to encourage researchers to employ alternative experimental games to test hypotheses. An important reason to study one particular class of games is that they are more economically realistic or relevant. The ultimatum game and the prisoner's dilemma are parsimonious representations of important phenomena of bargaining and public-goods situations,

---

[33] See Charness (2000) for a discussion of *responsibility alleviation*, and a review of papers with evidence related to the phenomena.

and hence it may be argued that it is most important to develop models that do well in explaining behavior in those contexts. But they are not the only representations of these phenomena.[34]

Indeed, to discuss both positive reciprocity and our worries about the range of games we study, we wish to return to Game Barc7. In this game A chose between (750,0) allocation or giving B a choice between (750,400) and (400,400) allocation. This strikes as a simplified form of a very common social and economic situation: A "wealthy" party can do something for a less well off party and hope that second party won't take advantage of a chance for petty, low-cost punishment just to hurt the first party. This is a common situation.[35] We suspect it is far more common than bilateral ultimatum games, and suspect it is partly an accident of recent experimental history that the ultimatum game happens to have become the game of choice to test.

Many researchers (ourselves included) have concluded that positive reciprocity has virtually no explanatory power in many of the conventional games studied. While the data from Barc7 are only one session with 36 subjects on only this one game, in the context of the current literature on social preferences the findings are striking. Yet if the results of this game hold up, it suggests a strong challenge indeed to the explanatory power of reciprocity-free distributional models in general and of difference aversion in particular. Will subjects who have just been treated kindly engage in petty acts of Pareto-damage just to equalize payoffs? Our suspicion is that the answer is broadly "no", and that even in models allowing heterogeneity of preferences for which it is appropriate to assume a substantial minority of subjects are difference-averse when neutral, these subjects will abandon that taste in a way that can only be viewed as reciprocity.

All said, it is clear that a broad array of additional games and methods would be useful for studying social preferences. Clearly, more research funding is needed.

---

[34] We surmise that one reason that a poor set of games has been used to differentiate among social preferences is that the games studied were originally studied in the context of either assuming narrow self-interest, or to test for the *existence*—not the *nature*—of departures from narrow self-interest. But when social preferences enter into the picture, the ultimatum game no longer serves as an adequate model of such a situation. The ultimatum game is, for instance, a poor proxy for employer-employee bargaining, where any accepted take-it-or-leave-it wage offer by an employer will be followed by opportunities for disgruntled employees to undermine the employer's profits.

[35] More generally, we believe that situations where one party has the chance to greatly affect another's payoffs at little effect of one's own are *very* common in the economic world. Perhaps representations of these, along with more

incremental bargaining games, ought to replace the ultimatum game as the game template for experimental researchers.

APPENDIX A: A MORE GENERAL MODEL

Before defining the full model that incorporates reciprocity, we first define an equilibrium notion based just on the quasi-maximin preferences. This is straightforward. We begin with the preferences for each player corresponding to those presented in Section 6. To put preferences in the context of games, let A_i be Player i's pure strategies, S_i be Player i's mixed strategies, and $S_{-i} \equiv \times_{j \neq i} S_j$ be the set of strategies for all players besides Player i. The material payoffs are determined by actions taken, where $\pi_i(a_1,\ldots,a_N)$ represents Player i's payoffs given actions $(a_1,\ldots,a_N)$.

Definition: For given parameters $(\gamma,\delta) \in [0,1]$, a *quasi-maximin equilibrium* (QME) of the material game $(A_1,\ldots A_N;\pi_1\ldots\pi_N)$ is a strategy profile $(s_1,\ldots s_N)$ that corresponds to Nash equilibrium of the game $(A_1,\ldots A_N;V_1(\pi)\ldots V_N(\pi))$, where $V_i(\pi)$ is Player i's (γ,δ)-quasi-maximin utility function.

Because $\pi_1,\ldots\pi_n$ are continuous in the players' actions, the functions $V_i(\pi)$ are well-defined and continuous in the players' actions. Hence, a QME always exists.

As with other distributional models, one could readily define a range of solution concepts with respect to quasi-maximin utility functions. Both refinements of Nash equilibrium (such as subgame-perfect Nash equilibrium) and less restrictive concepts (such as rationalizability) can be applied directly to the transformed games. Our model does not incorporate any sophisticated notion of sequential rationality, as have some recent reciprocity models, such as Dufwenberg and Kirchsteiger (1998) and Falk and Fischbacher (1998). We do not do so, partly to keep our model simple, and partly because some of the better predictions made by these models are obtained in our model as well without sequential refinements, by assuming that players are motivated to help others even in the absence of sacrifice by others. Moreover, we suspect that much of the intuition in these models—and the evidence invoked in favor of these intuitions—derive from heterogenous and non-equilibrium play in experiments, rather than from a notion of how players should behave at points in a game that really are "off the equilibrium path". If it is unrealistic to assume that the second mover in a sequential prisoner's dilemma will play a strategy of unconditional cooperation no matter what a first mover does, it is probably not because unconditional cooperation is not a best response to certainty that the first mover will cooperate.

It seems more likely that the real positive probability (due either to heterogenous preferences or disequilibrium) that a first mover will defect induces the second mover to defect in response to an interpretable on-the-equilibrium-path play by the first mover, rather than as part of an off-the-equilibrium-path strategy.

QME is a useful alternative to difference-aversion models in both reciprocity-free environments—where players are unlikely to be motivated by reciprocity—and in "simple-model environments"—where researchers want the most tractable model possible—QME can provide more explanatory power than other distributional models. But QME also serves as a foundation for our reciprocity model. Indeed, with an important restriction placed on the parameters of our model, every QME will be an equilibrium in our full reciprocity model.

To begin to incorporate reciprocity, consider a strategy profile $s \equiv (s_1,s_2,\ldots,s_n)$, as well as a *demerit profile*, $\rho \equiv (\rho_1,\ldots\rho_n)$, where $\rho_k \in [0,1]$ for all k. Below ρ will be determined endogenously. For now, ρ_k can be interpreted roughly as a measure of how much Player k deserves, where the higher the value of ρ_k, the less others think Player k deserves. With this interpretation, we define players' preferences as a function of both there underlying quasi-maximin preferences and how they feel about other players:

$$U_i(s,\rho) \equiv (1-\gamma)\cdot\pi_i + \gamma\cdot[\delta\cdot\text{Min}[\pi_i, \text{Min}_{m\neq i}\{\pi_m+d\rho_m\}] + (1-\delta)\cdot(\pi_i + \Sigma_{m\neq i}\,\max[1-k\rho_m,0]\pi_m) - f\,\Sigma_{m\neq i}\,\rho_m\cdot\pi_m],$$

where d, k, and f are non-negative parameters of the model. The key new aspect to these preferences is that the greater is ρ_j for $j\neq i$, the less weight Player i places on Player j's payoff. Hence, these preferences say that the more Player i feels that a Player j is being a jerk, the less Player i wants to help him. When the parameter f is positive, Player i may in fact wish to hurt Player j when Player j is being a jerk. The nature of these preferences, and how they match our data and intuitive discussions, can be seen most starkly by setting $f = 0$, and assuming that d and k are both very large. Then the preferences $U_i(s,\rho)$ imply that Player i maximizes the disinterested quasi-maximin allocation among all those other players for which $\rho_j = 0$—that is, among all the deserving others—and ignores the payoffs among those who are underserving.

We begin endogenizing the demerits ρ by defining, for every profile of strategies s_{-i} and demerits ρ_{-i} for other players, and every $g \in [0,1]$, the set of Player i's strategies that would maximize her utility *if* she put weight g on the social good and weight 1-g on her own payoff:

$$S_i^*(s_{-i},\rho_{-i};g) \equiv \{s_i \in S_i \mid s_i \in \text{argmax}\ \{(1\text{-}g)\,\pi_i + g[\delta\,\text{Min}[\pi_i,\text{Min}_{m\neq i}\,\{\pi_m + d\rho_m\}] + (1\text{-}\delta)\,[\Sigma_{j=1\ldots n}\,\pi_j - k\Sigma_{m\neq i}\,\rho_m\cdot\pi_m] - f\,\Sigma_{m\neq i}\,\rho_m\cdot\pi_m]\}\},$$

where π is the profile of material payoffs. "Typically", $S_i^*(s_{-i},\rho_{-i},g)$ will be a singleton set. The material payoffs are a function of players' actions, and hence strategies; we suppress this fact in our notation.

We let $g_i(s,\rho)$ be some upper hemi-continuous and convex-valued correspondence from (s,ρ) into the set [0,1] such that, for values (s,ρ) where $\{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$ is non-empty, $g_i(s,\rho) \approx \{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$. That is,

This convoluted formulation embeds a "smoothing" procedure that is a common trick to assure continuity in reciprocity models (see, e.g., Rabin (1993) and Falk and Fischbacher (1998)), assuring here that there exists such a correspondence meeting the criteria of upper hemi-continuity and convexity.

The full definition of $g_i(s,\rho)$ is as follows. Let $\varepsilon(s,\rho)$ be the neighborhood around (s,ρ) with all components within $\varepsilon > 0$ of (s,ρ). We then let $g_i(s,\rho)$ be any upper hemi-continuous and convex-valued correspondence such that $\{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\} \subseteq g_i(s,\rho) \subseteq G(\varepsilon,s,\rho)$, where $G(\varepsilon,s,\rho)$ is the convex hull of $\{g \mid t_i \in S_i^*(t_{-i},\chi_{-i},g)$ for some $(t,\chi) \in \varepsilon(s,\rho)\}$ if $\{g \mid t_i \in S_i^*(t_{-i},\chi_{-i},g)$ for some $(t,\chi) \in \varepsilon(s,\rho)\}$ is non-empty, and $G(\varepsilon,s,\rho) = [0,1]$ if $\{g \mid t_i \in S_i^*(t_{-i},\chi_{-i},g)$ for some $(t,\chi) \in \varepsilon(s,\rho)\}$ is empty. This is entirely unrestrictive when $\{g \mid t_i \in S_i^*(t_{-i},\chi_{-i},g)$ for some $(t,\chi) \in \varepsilon(s,\rho)\}$ is empty. But, assuming as we do that ε is small, $g_i(s,\rho) \approx \{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$ when $\{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$ is non-empty.

The function $g_i(s,\rho)$ will serve as a measure of how appropriately other players feel that Player i is behaving when they determine how to reciprocate. It can be interpreted as the degree to which Player i is pursuing the social good (that is, pursuing the disinterested quasi-maximin criterion) by choosing s_i in response to s_{-i}, given that she has disposition ρ_{-i} towards the other

players. Except for a technical fix to assure that $g_i(s,\rho)$ is upper hemi-continuous and convex-valued, this interpretation holds when there *exists* some degree of concern for the social good that, combined with self interest, can explain Player *i*'s choice. But some strategies may not be consistent with any such weighting—as, for instance, when a person chooses a Pareto-inefficient allocation even when the others have no demerits. In such cases, our model does not pin down a particular functional form, and hence in some cases can be unrestrictive.

The unrestrictiveness of our model in such cases is partly for technical convenience and because it doesn't matter much. This unrestrictiveness would be more problematic if we were to use it to predict non-equilibrium outcomes, or outcomes for heterogeneous preferences. But we don't restrict $g_i(s,\rho)$ when $\{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$ is empty also because we don't feel we know the right psychology for how people interpret seemingly unmotivated Pareto-damaging behavior or behavior that seems motivated by different norms of fairness than expected.

To derive demerit profiles from these functions, we assume that other players compare each $g_i(s,\rho)$ to some selflessness standard, γ^*, the weight they feel a decent person puts on social good. Specifically, we assume that other players' level of animosity towards Player *i* corresponds to $r_i(s,\rho,\gamma^*) \in \{\text{Max}[\gamma^* - g, 0] \mid g \in g_i(s,\rho)\}$. That is, whenever $\text{Max}\{g \mid g \in g_i(s,\rho)\} < \gamma^*$, Player *i* will generate some degree of animosity in others, since he is judged to be hurting others relative to what they would get if he were pursuing quasi-maximin preferences with $\gamma = \gamma^*$. When $\text{Min}\{g \mid g \in g_i(s,\rho)\} \geq \gamma^*$, others will feel no animosity towards Player *i*. Requiring elements of $r_i(s,\rho,\gamma^*)$ to be non-negative greatly simplifies the model. It is, however, also a substantive assumption that essentially rules out positive reciprocity. But given the lack of positive reciprocity in our data and those of others, it may not be a costly restriction in many situations. We can now define our solution concept:

Definition: The strategy profile *s* is a *reciprocal-fairness equilibrium* (RFE) with respect to parameter profiles γ, γ^*, δ, *d*, *k*, *f* and correspondence $g_i(s,\rho)$ if there exists ρ where, for all *i*, there exists $g_i \in g_i(s,\rho)$ such that

1) $s_i \in \text{Argmax } U_i(s,\rho)$, and
2) $\rho_i = \text{Max}[\gamma^* - g_i, 0]$.

A strategy profile is a RFE if every player is maximizing her expected utility given other players' strategies and given some demerit profile that is itself consistent with the profile of strategies. While not stated in that framework, this definition implicitly corresponds to a psychological Nash equilibrium of a psychological game as formulated by Geanakoplos, Pearce, and Stacchetti (1989). Were we to define a non-equilibrium notion of players' preferences, the entire formal apparatus would be needed. Because we just define the equilibrium concept, suppressing the psychological-game apparatus is both feasible and tractable.

The implications of RFE depend, of course, on the specific parameter values assumed, and hence it is unrestrictive insofar as there are many degrees of freedom in interpreting behavior as consistent with RFE. It is too restrictive to be directly applied to experimental evidence, on the other hand, because it does not allow for other social preferences, heterogeneity in players' preferences, or non-equilibrium play. But two results enhance the applicability of reciprocal-fairness equilibrium:

Theorem 1: For all parameter values and for all games, the set of RFE is non-empty.

Proof: Let h be the mapping from (s,ρ) into itself defined by the best-response correspondences $s_i \in$ Argmax $U_i(s,\rho)$ and the demerit functions $\rho_i(s,\rho) \in \{r \mid \exists\, g \in g_i(s,\rho)$ such that $r = \text{Max}[\gamma^* - g_i, 0]\}$. If this mapping is upper hemi-continuous and convex-valued, then it will have a fixed point, and this fixed point will be a RFE. By the continuity of $U_i(s,\rho)$ and the expected-utility structure, Argmax $U_i(s,\rho)$ is upper hemi-continuous and convex-valued. The component $\rho_i(s,\rho)$ is upper hemi-continuous and convex-valued because $g_i(s,\rho)$ is, by assumption, upper hemi-continuous and convex-valued. Hence, h is upper hemi-continuous and convex-valued, proving the theorem.

Existence clearly enhances the applicability of the solution concept. A second feature also enhances the applicability of the model despite potential complications due to incorporating reciprocity. Above we noted that quasi-maximin equilibria would play a prominent role in our model. Because of the reciprocity component in preferences (operative when $\rho_k > 0$ for some k), reciprocal-fairness equilibria might not correspond to quasi-maximin equilibria. Outcomes such as non-cooperation in the prisoners' dilemma can be "concern-withdrawal equilibria". Indeed, if players hold each other to very high standards of selflessness—if γ^* is very high—it may be that such negative outcomes are the only RFE. But if all players' intrinsic desire, γ, to pursue the

social good rather than self interest is at least as great as the standard, γ^*, to which people hold each other, then all quasi-maximin equilibria will be reciprocal-fairness equilibria:

Theorem 2: For all vectors of parameters such that $\gamma^* \leq \gamma$, every quasi-maximin equilibrium is a reciprocal-fairness equilibrium.

Proof: Consider a QME s^*. Each Player i is playing a best response given $\rho_{-i} = 0$, so that $\gamma \in g_i(s,\rho)$. If $\gamma \geq \gamma^*$, this means that $0 = \text{Max}[\gamma^* - \gamma, 0]$. Hence, s^* is a RFE with respect to the demerit profile $\rho = 0$.

Theorem 2 indicates that QME may serve as a good heuristic to predict the types of "cooperative" equilibria that can occur. Of course, there may additionally be negative equilibria, and (more importantly for interpreting experimental data) there may be either disequilibrium play or heterogenous preferences, where $\gamma < \gamma^*$ for some of the participants, so that some bad behavior, and corresponding retaliation, may be observed.

APPENDIX B - SAMPLE INSTRUCTIONS

INSTRUCTIONS

Thank you for participating in this experiment. You will receive $5 for your participation, in addition to other money to be paid as a result of decisions made in the experiment.

You will make decisions in several different situations ("games"). Each decision (and outcome) is independent from each of your other decisions, so that your decisions and outcomes in one game will not affect your outcomes in any other game.

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

There are "roles" in each game - generally A or B, although some games also have a C role. If a game has multiple decisions (some games only have decisions for one role), these decisions will be made sequentially, in alphabetical order: "A" players will complete their decision sheets first and their decision sheets will then be collected. Next, "B" players complete their decision sheets and these will be collected. Etc.

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

There will be two "periods" in each game and so you will play each game twice, with a different role (and a different anonymous pairing) in each case. You will not be informed of the results of any previous period or game prior to making your decision.

Although you will thus have 8 "outcomes" from the games played, only two of these outcomes will be selected for payoffs. An 8-sided die will be rolled twice at the end of the experiment and the (different) numbers rolled will determine which outcomes (1-8) are used for payoffs.

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.

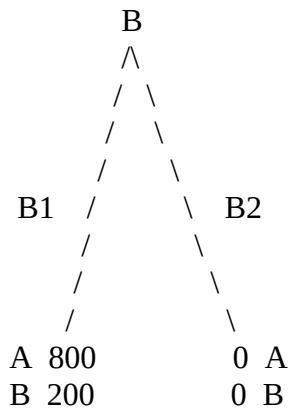
We will proceed to the decisions once the instructions are clear. Are there any questions?

PERIOD 1

GAME 3

In this period, you are person A.

You have no choice in this game. Player B's choice determines the outcome. If player B chooses B1, you would receive 800 and player B would receive 200. If player B chooses B2, you would each receive 0.



DECISION

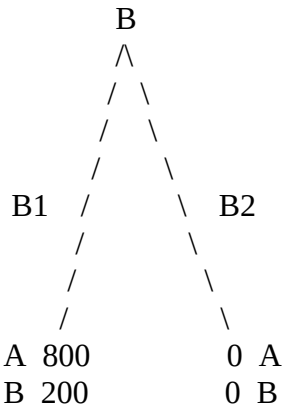
I understand I have no choice in this game _____

PERIOD 1

GAME 3

In this period, you are person B.

You may choose B1 or B2. Player A has no choice in this game. If you choose B1, you would receive 200 and player A would receive 800. If you choose B2, you would each receive 0.



DECISION

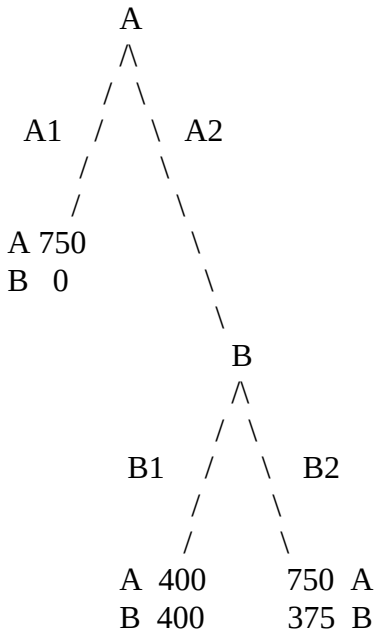
I choose: B1 B2

PERIOD 1

GAME 1

In this period, you are person A.

You may choose A1 or A2. If you choose A1, you would receive 750 and player B would receive 0. If you choose A2, then player B's choice of B1 or B2 would determine the outcome. If you choose A2 and player B chooses B1, you would each receive 400. If you choose A2 and player B chooses B2, you would receive 750 and he or she would receive 375. Player B will make a choice without being informed of your decision. Player 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.



DECISION

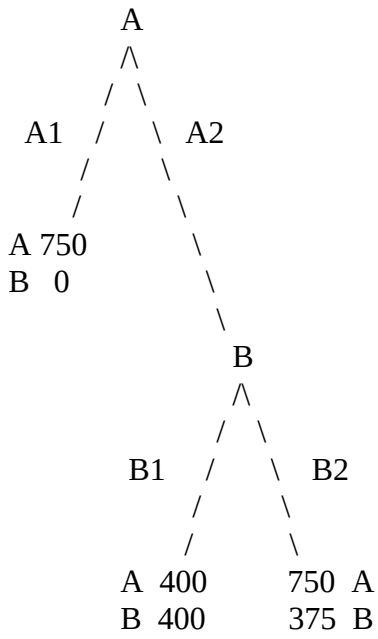
I choose: A1 A2

PERIOD 1

GAME 1

In this period, you are person B.

You may choose B1 or B2. Player A has already made a choice. If he or she has chosen A1, he or she would receive 750 and you would receive 0. Your decision only affects the outcome if player A has chosen A2. Thus, you should choose B1 or B2 on the assumption that player A has chosen A2 over A1. If player A has chosen A2 and you choose B1, you would each receive 400. If player A has chosen A2 and you choose B2, then player A would receive 750 and you would receive 375.



DECISION

I choose: A1 A2

APPENDIX C – Role Reversal

Tables 4.1-4.3 show the role-reversal data for each of the 19 games. The two-tailed p-value included reflects the likelihood that the difference in rates would occur randomly:

For each type of behavior as A, did the person help A as B?		if Out	if Enter	p-value
Helping A Doesn't Affect B's Payoff				
Barc5 (36)	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	5/14	19/22	.00
Barc7 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	15/17	19/19	.12
Berk28 (32)	A chooses (100,1000) or lets B choose (75,125) vs. (125,125)	10/16	11/16	.73
Berk32 (26)	A chooses (450,900) or lets B choose (200,400) vs. (400,400)	16/22	1/4	.06
Helping A is Costly to B				
Barc3 (42)	A chooses (725,0) or lets B choose (400,400) vs. (750,375)	10/31	6/11	.19
Barc4 (42)	A chooses (800,0) or lets B choose (400,400) vs. (750,375)	11/35	5/7	.05
Berk21 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,375)	3/17	11/19	.01
Barc6 (36)	A chooses (750,100) or lets B choose (300,600) vs. (700,500)	8/33	1/3	.73
Barc9 (36)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	2/25	0/11	.24
Berk25 (32)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	3/20	3/12	.48
Berk19 (32)	A chooses (700,200) or lets B choose (200,700) vs. (600,600)	13/18	12/14	.36
Berk14 (22)	A chooses (800,0) or lets B choose (0,800) vs. (400,400)	6/15	6/7	.05
Barc1 (44)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	1/42	2/2	.00
Berk13 (22)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	1/19	3/3	.00
Berk18 (32)	A chooses (0,800) or lets B choose (0,800) vs. (400,400)	0/0	14/32	
Helping A is Beneficial to B				
Barc11 (35)	A chooses (375,1000) or lets B choose (350,350) vs. (400,400)	15/19	16/16	.05
Berk22 (36)	A chooses (375,1000) or lets B choose (250,350) vs. (400,400)	13/14	22/22	.20
Berk27 (32)	A chooses (500,500) or lets B choose (0,0) vs. (800,200)	11/13	18/19	.34
Berk31 (26)	A chooses (750,750) or lets B choose (0,0) vs. (800,200)	16/19	7/7	.26
Berk30 (26)	A chooses (400,1200) or lets B choose (0,0) vs. (400,200)	19/20	4/6	.06

APPENDIX D: Game-by-Game Consistency with Distributional Models

In this Table, we allow A to have any beliefs about B's response to Enter.

Game	A Exit		A Enter		B plays Left		B plays Right	
	N	*Prefs.*	*N*	*Prefs.*	*N*	*Prefs.*	*N*	*Prefs.*
1 A(550,550); B(400,400)-(750,375)	42	C,D,Q,$	2	C,D,Q,$	41	C,D,Q,$	3	Q
2 B(400,400)-(750,375)	-		-		25	C,D,Q,$	23	Q
3 A(725,0); B(400,400)-(750,375)	31	C,D,Q,$	11	C,D,Q,$	26	C,D,Q,$	16	Q
4 A(800,0); B(400,400)-(750,375)	35	C,D,Q,$	7	D,Q	26	C,D,Q,$	16	Q
5 A(550,550); B(400,400)-(750,400)	18	C,D,Q,$	28	C,D,Q,$	15	C,D,$	31	Q,$
6 A(750,100); B(300,600)-(700,500)	33	C,D,Q,$	3	D,Q	27	C,D,Q,$	9	D,Q
7 A(750,0); B(400,400)-(750,400)	17	C,D,Q,$	19	D,Q,$	2	C,D,$	34	Q,$
8 B(300,600)-(700,500)	-		-		24	C,D,Q,$	12	D,Q
9 A(450,0); B(350,450)-(450,350)	25	C,D,Q,$	11	D,Q,$	34	C,D,Q,$	2	
11 A(375,1000); B(400,400)-(350,350)	19	C,D,Q,$	16	C,D,Q,$	31	C,D,Q,$	4	
13 A(550,550); B(400,400)-(750,375)	19	C,D,Q,$	3	C,D,Q,$	18	C,D,Q,$	4	Q
14 A(800,0); B(0,800)-(400,400)	15	C,D,Q,$	7	D,Q	10	C,D,Q,$	12	D,Q
15 B(200,700)-(600,600)	-		-		6	C,D,Q,$	16	D,Q
17 B(400,400)-(750,375)	-		-		16	C,D,Q,$	16	Q
18 A(0,800); B(0,800)-(400,400)	0		32	C,D,Q,$	14	C,D,Q,$	18	D,Q
19 A(700,200); B(200,700)-(600,600)	18	C,D,Q,$	14	D,Q	7	C,D,Q,$	25	D,Q
21 A(750,0); B(400,400)-(750,375)	17	C,D,Q,$	19	D,Q,$	22	C,D,Q,$	14	Q
22 A(375,1000); B(400,400)-(250,350)	14	C,D,Q,$	22	C,D,Q,$	35	C,D,Q,$	1	C
23 B(800,200)-(0,0)	-		-		36	C,D,Q,$	0	C,D
25 A(450,0); B(350,450)-(450,350)	20	C,D,Q,$	12	D,Q,$	26	C,D,Q,$	6	
26 B(0,800)-(400,400)	-		-		25	C,D,Q,$	7	D,Q
27 A(500,500); B(800,200)-(0,0)	13	C,D,Q,$	19	C,D,Q,$	29	C,D,Q,$	3	C,D
28 A(100,1000); B(75,125)-(125,125)	16	C,D,Q,$	16	C,D,Q,$	11	C,D,$	21	Q,$
29 B(400,400)-(750,400)	-		-		8	C,D,$	18	Q,$
30 A(400,1200); B(400,200)-(0,0)	20	C,D,Q,$	6	C,D,$	23	C,D,Q,$	3	C,D
31 A(750,750); B(800,200)-(0,0)	19	C,D,Q,$	7	C,D,Q,$	23	C,D,Q,$	3	C,D
32 A(450,900); B(200,400)-(400,400)	22	C,D,Q,$	4	C,D	9	C,$	17	D,Q,$

Total A choices = 671	C = 579	D = 671	Q = 661	$ = 636
Total B choices = 903	C = 579	D = 685	Q = 836	$ = 690

In this Table, we assume A correctly assesses actual B play when choosing.

Game	A Exit		A Enter		B plays Left		B plays Right	
	N	*Prefs.*	*N*	*Prefs.*	*N*	*Prefs.*	*N*	*Prefs.*
1 A(550,550); B(400,400)-(750,375)	42	C,D,Q,$	2	C	41	C,D,Q,$	3	Q
2 B(400,400)-(750,375)	-		-		25	C,D,Q,$	23	Q
3 A(725,0); B(400,400)-(750,375)	31	C,D,Q,$	11	D,Q	26	C,D,Q,$	16	Q
4 A(800,0); B(400,400)-(750,375)	35	C,D,Q,$	7	D,Q	26	C,D,Q,$	16	Q
5 A(550,550); B(400,400)-(750,400)	18	D,Q	28	C,D,Q,$	15	C,D,$	31	Q,$
6 A(750,100); B(300,600)-(700,500)	33	C,D,Q,$	3	D,Q	27	C,D,Q,$	9	D,Q
7 A(750,0); B(400,400)-(750,400)	17	C,D,Q,$	19	D,Q	2	C,D,$	34	Q,$
8 B(300,600)-(700,500)	-		-		24	C,D,Q,$	12	D,Q
9 A(450,0); B(350,450)-(450,350)	25	C,D,Q,$	11	D,Q	34	C,D,Q,$	2	
11 A(375,1000); B(400,400)-(350,350)	19	Q	16	C,D,Q,$	31	C,D,Q,$	4	
13 A(550,550); B(400,400)-(750,375)	19	C,D,Q,$	3	C	18	C,D,Q,$	4	Q
14 A(800,0); B(0,800)-(400,400)	15	C,D,Q,$	7	Q	10	C,D,Q,$	12	D,Q
15 B(200,700)-(600,600)	-		-		6	C,D,Q,$	16	D,Q
17 B(400,400)-(750,375)	-		-		16	C,D,Q,$	16	Q
18 A(0,800); B(0,800)-(400,400)	0		32	C,D,Q,$	14	C,D,Q,$	18	D,Q
19 A(700,200); B(200,700)-(600,600)	18	C,D,Q,$	14	D,Q	7	C,D,Q,$	25	D,Q
21 A(750,0); B(400,400)-(750,375)	17	C,D,Q,$	19	D,Q	22	C,D,Q,$	14	Q
22 A(375,1000); B(400,400)-(250,350)	14	Q	22	C,D,Q,$	35	C,D,Q,$	1	C
23 B(800,200)-(0,0)	-		-		36	C,D,Q,$	0	C,D
25 A(450,0); B(350,450)-(450,350)	20	C,D,Q,$	12	D,Q	26	C,D,Q,$	6	
26 B(0,800)-(400,400)	-		-		25	C,D,Q,$	7	D,Q
27 A(500,500); B(800,200)-(0,0)	13	D,Q	19	C,D,Q,$	29	C,D,Q,$	3	C,D
28 A(100,1000); B(75,125)-(125,125)	16	Q	16	C,D,Q,$	11	C,D,$	21	Q,$
29 B(400,400)-(750,400)	-		-		8	C,D,$	18	Q,$
30 A(400,1200); B(400,200)-(0,0)	20	C,D,Q,$	6	C,D	23	C,D,Q,$	3	C,D
31 A(750,750); B(800,200)-(0,0)	19	C,D,Q,$	7	C	23	C,D,Q,$	3	C,D
32 A(450,900); B(200,400)-(400,400)	22	C,D,Q,$	4	C,D	9	C,$	17	D,Q,$

Total A choices = 671	C = 579	D = 671	Q = 661	$ = 636
Total B choices = 903	C = 579	D = 685	Q = 836	$ = 690

APPENDIX E: First-Mover Behavior

Table 3.1: A's Sacrifice Helps B					**Maximize**	**Sacrifice**
Barc5 (36)	A chooses	(634,400)	or	(550,550)	.61	.39
Barc7 (36)	A chooses	(750,0)	or	(729,400)	.47	.53
Berk28 (32)	A chooses	(108,125)	or	(100,1000)	.50	.50
Barc3 (42)	A chooses	(725,0)	or	(533,390)	.74	.26
Barc4 (42)	A chooses	(800,0)	or	(533,390)	.83	.17
Berk21 (36)	A chooses	(750,0)	or	(536,390)	.47	.53
Barc6 (36)	A chooses	(750,100)	or	(400,575)	.92	.08
Barc9 (36)	A chooses	(450,0)	or	(356,444)	.69	.31
Berk25 (32)	A chooses	(450,0)	or	(369,431)	.62	.38
Berk19 (32)	A chooses	(700,200)	or	(512,622)	.56	.44
Berk14 (22)	A chooses	(800,0)	or	(216,584)	.68	.32
Berk18 (32)	A chooses	(224,576)	or	(0,800)	1.00	.00
Barc11 (35)	A chooses	(394,394)	or	(375,1000)	.46	.54
Berk22 (36)	A chooses	(396,398)	or	(375,1000)	.61	.39
Berk27 (32)	A chooses	(728,182)	or	(500,500)	.59	.41

Table 3.2: A's Sacrifice Hurts B					**Maximize**	**Sacrifice**
Berk32 (26)	A chooses	(450,900)	or	(330,400)	.85	.15
Barc1 (44)	A chooses	(550,550)	or	(424,398)	.96	.04
Berk13 (22)	A chooses	(550,550)	or	(463,396)	.86	.14
Berk31 (26)	A chooses	(750,750)	or	(704,176)	.73	.27
Berk30 (26)	A chooses	(400,1200)	or	(352,176)	.77	.23

REFERENCES

Andreoni, James and John Miller, "Giving According to GARP: An Experimental Study of Rationality and Altruism," 1998, mimeo, Social Systems Research Institute, University of Wisconsin, Madison.

Andreoni, James, Paul Brown, and Lise Vesterlund, "What Makes an Allocation Fair? Some Experimental Evidence," 1999, mimeo.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," Games and Economic Behavior, 1995, **10**, 122-42.

Blount, Sally (1995), "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," Organizational Behavior and Human Decision Processes **63**, 131-144.

Bolton, Gary and Axel Ockenfels, "Strategy and Equity: An ERC-analysis of the Güth-van Damme game," Journal of Mathematical Psychology, 1998, **42**, 215-226.

Bolton, Gary and Axel Ockenfels, "ERC: A Theory of Equity, Reciprocity, and Competition," American Economic Review, 2000, **90**, 166-193.

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

Bolton, Gary, Jordi Brandts, and Axel Ockenfels, "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game," Experimental Economics, 1998, **1**, 207-219.

Brandts, Jordi and Gary Charness, "Hot vs. Cold: Sequential Responses in Simple Experimental Games," Experimental Economics, 2000, **2**, 227-238.

Brandts, Jordi and Gary Charness, "Retribution in a Cheap-talk Game," 1999, mimeo.

Brandts, Jordi and Carles Solà, "Reference Points and Negative Reciprocity in Simple Sequential Games," 1998, forthcoming in *Games and Economic Behavior*.

Cason, Timothy and Vai-Lam Mui, "Social Influence in the Sequential Dictator Game," Journal of Mathematical Psychology, 1998, **42**, 248-265.

Charness, Gary, "Attribution and Reciprocity in an Experimental Labor Market: An Experimental Investigation," 1996, mimeo, University of California at Berkeley.

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

Charness, Gary and Brit Grosskopf, "Relative Payoffs and Happiness: An Experimental Study," 1999, forthcoming in Journal of Economic Behavior and Organization.

Charness, Gary and Ernan Haruvy, "Altruism, Fairness, and Reciprocity in a Gfit-exchange Experiment: An Encompassing Approach" 1999, mimeo.

Charness, Gary and Matthew Rabin, "Social Preferences: Some Simple Tests and a New Model," 1999, mimeo, Universitat Pompeu Fabra and University of California at Berkeley.

Croson, Rachel, "The Disjunction Effect and Reason-Based Choice in Games," 2000, Organizational Behavior and Human Decision Processes, **80**, 1999, 118-133.

Dufwenberg, Martin and Uri Gneezy, "Measuring Beliefs in an Experimental Lost Wallet Game," Games and Economic Behavior, 2000, **30**, 163-182.

Dufwenberg, Martin and Georg Kirchsteiger, "A Theory of Sequential Reciprocity," 1998, mimeo.

Falk, Armin and Urs Fischbacher, "A Theory of Reciprocity," 1998, mimeo.

Falk, Armin, Enrst Fehr, and Urs Fischbacher, "On the Nature of Fair Behavior," 1999, mimeo.

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

Geanakoplos, John, David Pearce, and Ennio Stacchetti (1989), "Psychological Games," Games and Economic Behavior, 1989, **1**, 60-79.

Glasnapp, Douglas and John Poggio, Essentials of Statistical Analysis for the Behavioral Sciences, 1985, Columbus, Merrill.

Güth, Werner and Eric van Damme, "Information, Strategic Behavior, and Fairness in Ultimatum Bargaining: An Experimental Study," Journal of Mathematical Psychology, 1998, **42** Jun-Sep: 227-247.

Kagel, John and Katherine Wolfe, "Testing Between Alternative Models of Fairness: A New Three-Person Ultimatum Game," 1999, mimeo.

Kritikos, Alexander and Friedel Bolle, "Approaching Fair Behavior: Self-Centered Inequality Aversion Versus Reciprocity and Altruism," 1999, mimeo.

Loewenstein, George, Max Bazerman and Leigh Thompson, "Social Utility and Decision Making in Interpersonal Contexts," Journal of Personality and Social Psychology, 1989, **57**, 426-441.

McFadden, Daniel, "Econometric Models of Probabilistic Choice," in Structural Analysis of Discrete Data with Econometric Applications, Charles Manski and Daniel McFadden, eds., 1981, (page numbers), Cambridge, MIT Press.

Offerman, Theo, "Hurting Hurts More than Helping Helps: The Role of the Self-serving Bias," mimeo, 1998.

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

Roth, Alvin, "Bargaining Experiments," in Handbook of Experimental Economics, J. Kagel and A. Roth, eds., 1995, 253-348.

Siegel, Sidney and N. John Castellan, Nonparametric Statistics for the Behavioral Sciences, 1988, Boston, McGraw-Hill.

Shafir, Eldar and Amos Tversky, "Thinking Through Uncertainty: Nonconsequentialist Reasoning and Choice," Cognitive Psychology, 1992, **23**, 449-474.

Yaari, Menahem and Maya Bar-Hillel, "On Dividing Justly," Social Choice and Welfare, 1984, **1**, 1-24