

# UNDERSTANDING SOCIAL PREFERENCES WITH SIMPLE TESTS\*

Gary Charness and Matthew Rabin

AUGUST 2001

**Abstract:** Departures from self-interest in economic experiments have recently inspired models of “social preferences”. We design a range of simple experimental games that test these theories more directly than existing experiments. Our experiments show that subjects are more concerned with increasing social welfare—sacrificing to increase the payoffs for all recipients, especially low-payoff recipients—than with reducing differences in payoffs (as supposed in recent models). Subjects are also motivated by reciprocity: They withdraw willingness to sacrifice to achieve a fair outcome when others are themselves unwilling to sacrifice, and sometimes punish unfair behavior.

**Keywords:** Difference Aversion, Fairness, Inequity Aversion, Social Welfare, Non-Ultimatum Games, Reciprocal Fairness, Social Preferences, Ultimatum Games.

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

\* This paper is a revised version of the related working papers Charness and Rabin [1999, 2000]. We thank Jordi Brandts, Antonio Cabrales, Colin Camerer, Martin Dufwenberg, Ernst Fehr, Urs Fischbacher, Simon Gächter, Edward Glaeser, Brit Grosskopf, Ernan Haruvy, John Kagel, George Loewenstein, Rosemarie Nagel, Christina Shannon, Lise Vesterlund, an anonymous referee, and seminar participants at Harvard University, Stanford University Graduate School of Business, University of California at Berkeley, University of California at San Diego, 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, the April 2000 Experimental Symposium at Technion, and the January 2001 ASSA meeting for helpful comments. We also thank Davis Beekman, Christopher Carpenter, David Huffman, Christopher 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/ University of California, Santa Barbara/ 2127 North Hall, Santa Barbara, CA 93106-9210. E-mail: [charness@econ.ucsb.edu](mailto:charness@econ.ucsb.edu). 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](mailto:rabin@econ.berkeley.edu). Web page: <http://elsa.berkeley.edu/rabin/index.html>.

# I. Introduction

Participants in experiments frequently choose actions that do not maximize their own monetary payoffs when those actions affect others' payoffs. They sacrifice money in simple bargaining environments to punish those who mistreat them and share money with other parties who have no say in allocations.

One hopes that the insights into the nature of non-self-interested behavior gleaned from experiments can eventually be applied to a variety of economic settings, such as consumer response to price changes, attitudes towards different tax schemes, and employee response to changes in wages and employment practices. To facilitate such applications, researchers have begun to develop formal models of *social preferences* that assume people are self-interested, but are also concerned about the payoffs of others. Different types of models have been formulated. “Difference-aversion models” assume that players are motivated to reduce differences between theirs and others' payoffs; “social-welfare models” assume that people like to increase social surplus, caring especially about helping those (themselves or others) with low payoffs; reciprocity models assume that the desire to raise or lower others' payoffs depends on how fairly those others are behaving.

In this paper, we report findings from some simple experiments that test existing theories more directly than the array of games commonly studied. We then fit our evidence to a simple, stylized model that encapsulates variants of existing models as special cases, and formulate a more complicated new model to capture patterns of behavior that previous models don't explain.

A major motivation for our research was a concern about pervasive and fundamental confounds in the experimental games that have inspired recent social-preferences models. Most notably, papers presenting difference-aversion models have argued that Pareto-damaging behavior—such as rejecting unfair offers in ultimatum games, where subjects lower both their own and others' payoffs—can be explained by an intrinsic preference to minimize differences in payoffs. But this explanation is almost universally confounded in two ways: First, opportunities for inequality-reducing Pareto-damaging behavior arise in these games solely when a clear motivation for retaliation is aroused. Second, the only plausible Pareto-damaging behavior permitted is to reduce inequality. Difference aversion has also been used to explain helpful

sacrifice—such as cooperation in prisoner’s dilemmas—as a taste for helping those with lower payoffs. But here again two confounds are nearly universal: The games studied only allow efficient helpful sacrifice that decreases inequality, and only when a motive for retaliation is *not* aroused.

All of these confounds mean that the tight fit of these models may merely reflect the fact that in many of the games studied their predictions happen both to be the only way that subjects can depart from self-interest, and to be the same as the predictions of reciprocity.<sup>1</sup>

To provide a more discerning examination of social preferences, our games offer an array of choices that directly test of the role of different social motivations, by testing a fuller range of possible departures from self-interest, by eliminating confounds within games, and by inviting crisp, revealing comparisons across games. Our data consist of 29 different games, with 467 participants making 1697 decisions.

In Section II, we provide a simple linear, two-person model of preferences that assumes that players’ propensity to sacrifice for another player is characterized by three parameters: The weight on the other’s payoff when she is ahead, the weight when she is behind, and the change in weight when the other player has misbehaved. This embeds difference aversion, social-welfare preferences, and other preferences as identically parsimonious and tractable special cases of a more general model. By way of the shift parameter, it also embeds a simple form of reciprocity. In Section III we explain our experimental procedures and raw results. We interpret our results without invoking intentions-based reciprocity in Section IV, and with reciprocity in Section V. We analyze our results both by comparing the percentage of data that different models explain, and with regression analysis of the best-fit parameter values of the model of Section II.

Our findings suggest that the role of inequality-reduction in motivating subjects has been exaggerated. Few subjects sacrifice money to reduce inequality by lowering another subjects’ payoff, and only a minority do so even when this is free. Indeed, we observed Pareto-damaging behavior more often when it *increased* inequality than when it *decreased* inequality. While this comparison is itself confounded by other explanations, our data strongly suggest that inequality

---

<sup>1</sup> The analysis articulated in developing these models, on the other hand, usefully demonstrates that the interpretations of authors (such as Rabin [1993]) that helpful sacrifice is based on positive reciprocity are misleading—since such helpful sacrifice is for the most part as strong when no positive feelings are aroused.

reduction is not a good explanation of Pareto-damaging behavior.<sup>2</sup> By contrast, difference-aversion models do provide an elegant insight into players' willingness to sacrifice when ahead of other players. Yet social-welfare preferences provide an even better theory of helpful sacrifice. By positing far greater concern for those who are behind than those who are ahead, they also predict helpful sacrifice by those with higher payoffs. By positing a concern for efficiency, however, social-welfare preferences predict that even if players are behind they may sacrifice small amounts to help those ahead. Unlike difference aversion, therefore, social-welfare preferences can explain the finding in our data that about half of subjects make inequality-*increasing* sacrifices when these sacrifices are efficient and inexpensive.<sup>3</sup>

To test the role of reciprocity, we study simple response games where Player B's choice follows a move by Player 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. Our data replicate recent experimental evidence that positive reciprocity is not a strong force in experimental settings.<sup>4</sup> But 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. Subjects also significantly increased their Pareto-damaging behavior following selfish actions by A.

Overall, straightforward interpretation of specific games and summary descriptive statistics show that social-welfare preferences explain our data better than does difference aversion, and that subjects clearly behave reciprocally. Our regression analysis indicates that a B who has a higher payoff than A puts great weight on A's payoff. However, if B has a lower payoff than A and no reciprocity is involved, the weight on A's payoff is close to 0. When A has mistreated B, B significantly decreases positive weight or puts negative weight on A's payoff.

While most of our data and our formal tests concern two-player games, in Section VI we discuss results in the five three-player games. These games provide some evidence for a multi-

---

<sup>2</sup> Other recent papers similarly providing data that calls into question the role of inequality reduction in Pareto-damaging behavior include Kagel and Wolfe [1999] and Engelmann and Strobel [2001].

<sup>3</sup> Andreoni and Miller [1998], Charness and Grosskopf [2001], and Kriticos and Bolle [1999] find similar results, with significant numbers of participants opting for inequality-increasing sacrifices to help others.

<sup>4</sup> One exception we find to this pattern is that positive feelings reduces difference aversion when self-interest is not at stake. We return to this finding in our concluding discussion. We also note that McCabe, Rigdon, and Smith (2000) find, in one simple game, significant and statistically significant evidence of positive reciprocity.

person generalization of social-welfare preferences, and further demonstrate 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 demonstrate that subjects are not indifferent to the distribution of material payoffs among other people.

Our experiments add to other recent evidence in providing raw data for developing better models of social preferences. Since there are clearly many forces at work in subjects' behavior, researchers are faced with the decision as to how complex a model to formulate, trading off progress on applicable models against the quest for psychological and empirical accuracy. Our own perspective is that too much will be lost if experimentalists jump too quickly to calibrating highly simplified models that ignore prevalent phenomena such as reciprocity. At this stage, models ought to be developed that help interpret psychologically sound and empirically prevalent patterns of behavior common in a broad array of games.

In this spirit, in Appendix A we develop a multi-person model of *reciprocal-fairness equilibrium* that combines social-welfare preferences and reciprocity. We presume that players are motivated to pursue social welfare, but withdraw the willingness to give others their social-welfare shares when these others are being unfair, and may even sacrifice to punish them. Despite its complexity, this model clearly omits many factors that play a role in laboratory behavior, and hence is unlikely to fit the data tightly. Rather, it is meant to provide incremental conceptual and calibrational progress in understanding the nature of social preferences.

For those who instead feel it is more urgent to develop simpler and more applicable models, our analysis in the body of the paper shows that an equally parsimonious alternative model explains behavior in our data better than difference-aversion models. The alternative essentially reverses the weight that players put on the payoffs of others doing better than them from strongly negative to weakly positive, reflecting a willingness to pursue social efficiency when it comes at a small cost to the worst-off player.<sup>5</sup>

---

<sup>5</sup>. This means, of course, that the *more* parsimonious model that assumes players ignore the payoffs of others who are ahead of them performs as well as difference-aversion models. We don't think our simple alternative will provide a *good* fit for the broad set of games economists should care about. It obviously doesn't explain rejections of unfair offers in ultimatum and bargaining games, which is the primary source of difference-aversion models beginning with Bolton [1991]. Our claim is merely that—once we start examining a broader array of games—it will provide a *better* fit than difference-aversion models. Difference-aversion models predict rejections in the ultimatum game. But they make *worse* predictions than pure self-interest in other games that are simpler, more diagnostic, and (we would contend) more economically relevant than the ultimatum game.

That said, we do not believe this paper establishes definitively that previous interpretations of non-self-interested behavior have been wrong.<sup>6</sup> Rather, our analysis clarifies clear confounds in the previous research supporting those interpretations. More than our specific findings and interpretations, in fact, we hope this paper helps move experimental research away from studying the existing, manifestly misleading, menu of games and towards a wider range of simpler and more diagnostic games. We conclude in Section VII with a discussion of some of the issues raised by this program and with some suggestions for new directions of research.

## II. Social Preferences

In this section we outline a simple conceptual model of social preferences in two-person games that embeds different existing theories of social preferences as different parameter ranges, and allows for the estimation of these parameter values in our empirical analysis below.<sup>7</sup> Letting  $\pi_A$  and  $\pi_B$  be Player A's and B's money payoffs, consider the following simple formulation of Player B's preferences:

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

where

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

---

<sup>6</sup> Indeed, while our analysis stresses that our data contradict difference-aversion models, we do not think we have conclusively disproved these models. This is for (at least) two reasons. 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 subject behavior is heterogeneous, and that there are subjects who exhibit some degree of difference aversion in some circumstances. Fehr and Schmidt [1999], for instance, have argued that only 40 percent of subjects need be difference-averse to explain the phenomena they explain. This is arguably consistent with our data. More generally, insofar as the existing literature has emphasized the *existence* of difference aversion as a force among *some* subjects, our evidence suggesting that it is weaker and rarer than very opposite forces may not contradict what has been found so far.

<sup>7</sup> By “conceptual,” we mean that one major component of preferences—the motive to punish unfair behavior—is left under-specified. The model in Appendix A develops the additional framework needed to fully formalize our assumptions about reciprocity.

This formulation says that B’s utility is a weighted sum of her own material payoff and A’s payoff, where the weight B places on A’s payoff may depend on whether A is getting a higher or lower payoff than B and on whether A has behaved unfairly.<sup>8</sup> The parameters  $\rho$ ,  $\sigma$ , and  $\theta$  capture various aspects of social preferences. The parameter  $\theta$  provides a mechanism for modeling reciprocity, which we shall return to below. The parameters  $\rho$  and  $\sigma$  allow for a range of different “distributional preferences”, that rely solely on the outcomes and not on any notion of reciprocity. We begin by discussing purely distributional preferences, which may be appropriate either in contexts where reciprocity is likely not to motivate subjects or when modelers are looking for a simple proxy for full-fledged preferences that include reciprocity.

One form of distributional preferences (consistent with the psychology of status) is simple competitive preferences. These can be represented by assuming  $\sigma \leq \rho \leq 0$ , 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 like their payoffs to be high relative to others’ payoffs.<sup>9</sup> 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, and prefers that payoffs are equal, including wishing to lower A’s payoff when A does better than B. 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 calling) social-welfare preferences rather than difference aversion. Our notion of social-welfare

---

<sup>8</sup> Another way of writing this utility function that some readers might find more intuitive is to break it down into two cases: When  $\pi_B \geq \pi_A$ ,  $U_B(\pi_A, \pi_B) \equiv (1-\rho-\theta q)\pi_B + (\rho+\theta q)\pi_A$ ; when  $\pi_B \leq \pi_A$ ,  $U_B(\pi_A, \pi_B) \equiv (1-\sigma-\theta q)\pi_B + (\sigma+\theta q)\pi_A$ .

preferences subsumes the different cases examined by Andreoni and Miller [1998], by letting the parameters take on the values  $1 \geq \rho \geq \sigma > 0$ .<sup>10</sup> 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.<sup>11</sup> Social-welfare 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, but are particularly keen to help the person who is worst off.<sup>12</sup>

Since social-welfare preferences assume that people always prefer Pareto-improvements, they cannot explain Pareto-damaging behavior such as rejections in the ultimatum game. 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.<sup>13</sup> Roughly put, these models say that B's values for  $\rho$  and  $\sigma$  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  $\rho$  and  $\sigma$  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

<sup>9</sup> Assuming  $\sigma \leq \rho$  says that the preference for gains relative to the other person is at least as high when behind as when ahead.

<sup>10</sup> 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. 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.

<sup>11</sup> Earlier studies by Frohlich and Oppenheimer [1984, 1992] similarly find that subjects reach agreements that tend to maximize total payoffs, while observing an income floor for individuals in the group. They also find statistical relationships between choices and partisan political preferences.

<sup>12</sup> 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.

<sup>13</sup> Studies demonstrating reciprocity that cannot be explained by distributional models include Kahneman, Knetsch, and Thaler [1986], Blount [1995], Charness [1996], Offerman [1998], 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.

from A's unwillingness to come out behind rather than A's selfishness when ahead. That is, B retaliates against behavior implying that A's  $\rho$  is too small, but not against behavior indicating that  $\sigma$  is small or negative. An alternative hypothesis about reciprocal preferences follows naturally from social-welfare preferences: 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 social-welfare preferences then they decrease these weights.

Reciprocity can be captured simply (and crudely) by assuming  $\theta > 0$ : When  $q = -1$ , indicating that A has "misbehaved" by violating the dictates of social-welfare preferences, this essentially assumes that B lowers both  $\rho$  and  $\sigma$  by amount  $\theta$ . In Appendix A we define the solution concept *reciprocal-fairness equilibrium*, combining social-welfare motivations and intentions-based reciprocity in a model of social-preferences in multi-person games. But the remainder of the paper concentrates on testing the simple model of this section. After explaining our experiments and presenting our results in the next section, in Section IV we discuss the fit of our data with different distributional models by assuming  $\theta = 0$ , and in Section V we explore the role of reciprocity in our data by considering our model when  $\theta$  is not restricted.

### **III. 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. 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 conducted for this project that were played for financial stakes.<sup>14</sup> We designed the Berkeley games after examining the Barcelona results, and modified several games after observing earlier results.<sup>15</sup>

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 students from a broader range of academic disciplines than is common in economics experiments.<sup>16</sup>

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

---

<sup>14</sup> 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. Since the first draft of this paper was circulated, we have run additional related experiments (for another paper) whose analysis we had not intended to and do not include in this paper. Though these are heavily confounded with reciprocity interpretations, we note that in these new data we observed a rise in the percentage of responding subjects exhibiting difference-averse behavior in three of the conditions we report on here. But by and large the new data seem to qualitatively and quantitatively support the conclusions of this paper, and we do not have data in our possession that we believe broadly contradicts any of our interpretations in this paper.

<sup>15</sup> 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.

<sup>16</sup> 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.

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 later be playing in the other role, but (to discourage reputational motivations) were assured that pairings were changed in each period.

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 in Games 1-4, two of four in Games 5-12, and two of eight in Games 13-32 were paid.

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 (the *strategy method* of elicitation), after being told that her 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.<sup>17</sup> We do not believe that the use of either role reversal or the strategy method is an important factor in our results.

---

<sup>17</sup>. See Roth [1995, p. 323] for a hypothesis that 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'

Table I 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.<sup>18</sup>

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  $\rho$  and  $\sigma$ .

#### INSERT TABLE I ABOUT HERE

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 social-welfare 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 forgone. 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 aspects of social preferences discussed in the previous section. In our general analysis, we will gloss over many plausibly important issues and alternative hypotheses about what

---

decisions. Cason and Mui [1998] and Brandts and Charness [2000] conduct tests where it doesn’t seem to matter much; Shafir and Tversky [1992] and Croson [2000] find some difference in the propensity to cooperate in a Prisoner’s Dilemma using the two methods.

<sup>18</sup> 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 (it was actually 500), 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.

explains the behavior we observe in particular games.<sup>19</sup> 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, and our goal in our experimental design was to create a diverse list of games giving scope for the widest array of social motivations to play out, and providing scope for the models to fail.

## IV. Explaining Behavior by Distributional Preferences

In this section, we compare the power of self-interest and distributional models (competitive, difference-averse, and social-welfare) to explain our data. We mostly consider how many observations in our games are consistent with the values of  $\rho$  and  $\sigma$  permitted by the restrictions for each type of social preferences, when excluding reciprocity by imposing the restriction  $\theta = 0$ .<sup>20</sup> 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  $\rho$  and  $\sigma$ .

Table II shows the explanatory power of various models, under the appropriate restrictions for  $\rho$  and  $\sigma$ .<sup>21</sup> As we are not yet considering reciprocity motivations, which may influence

---

<sup>19</sup>. 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.

<sup>20</sup>. 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.

<sup>21</sup>. 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

preferences in response games, it is most appropriate to make comparisons using only the seven dictator games. The first line indicates that social-welfare preferences are far more effective than the others in explaining behavior when reciprocity issues are absent.

#### INSERT TABLE II ABOUT HERE

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 note that an exact tie in B's payoff provides the best possible chance of revealing Pareto-damaging difference aversion, since it eliminates self-interest and everything else as a countervailing motive. In fact, the one third of subjects exhibiting difference aversion here is the most we find in our data.<sup>22</sup> In games where pursuing Pareto-damaging difference aversion would require sacrifice, we see far less such pursuit. In Berk23, for instance, which tests the reciprocity-free willingness of participants to reject offers of the sort rejected in many ultimatum-game experiments, 0 of 36 B's chose (0,0) over (800,200).<sup>23</sup>

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 social-welfare preferences. 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.<sup>24,25</sup>

---

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.

<sup>22</sup> And it should be noted that this one third also includes those people with competitive preferences.

<sup>23</sup> 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.

<sup>24</sup> 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, but we use the two-tailed test (and say so) where there is no directional hypothesis.

In this and the other comparisons in Table II, the proportion of observations explained by social-welfare 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 social-welfare preferences and the other three categories would be statistically significant at  $p \approx .00$  if each observation were treated as independent.<sup>26</sup>

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 of 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 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 III 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 social-welfare preferences substantially outperform the other models.

INSERT TABLE III ABOUT HERE

Line 1 shows that social-welfare preferences 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.<sup>27</sup> We discuss the specific findings on the various types of response games in the next section. But line 2 of both Table II and Table III shows that social-welfare preferences and even narrow self-

---

<sup>25</sup> A higher proportion of B's take a 100 percent share in Berk26 than in traditional dictator experiments. But the 22 percent rate observed for even splits is not unusual in a dictator game, and no intermediate split was available.

<sup>26</sup> 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.

<sup>27</sup> 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 II and III, and especially Table IV below, our experiments call into question even this weaker case for difference aversion.

interest outperform difference aversion and competitive preferences. 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 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 restriction on preferences if his choice is consistent given any belief about what B might do. A stronger, more common, and more 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 II and III assess A's behavior using both methods.

Referring to Table II, under the liberal interpretation of consistency, few choices by A are entirely inconsistent with any of the models, but clearly difference aversion and social-welfare preferences 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 social-welfare 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 social-welfare equilibrium than with difference aversion.

The behavior by A's in our experiments help shed light on the stylized fact, much-emphasized over the years, that observed generosity by proposers in ultimatum games is not discernibly inconsistent with narrow self-interest, since “generous” offers are an optimizing response to 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 approximately self-interested is (as with many hypotheses) not sustainable when analyzing games besides the ultimatum game. In our data, 27 percent 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 percent sacrificed, whereas

only 15 percent 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 percent (92/276) sacrificed. In the two cases where entry by A could not help either player, 19 percent (10/52) entered.<sup>28</sup> Tables II and III show that, depending on how one measures it, departures from self-interest are just as common for A's as for B's.

The last two rows of Tables II and III 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.<sup>29</sup>

#### INSERT TABLE IV ABOUT HERE

Table IV 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. Overall, B's engage in Pareto-damaging behavior in 17 percent of their opportunities to do so. More interestingly, 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 combined with the overwhelming confound between inequality reduction and Pareto-damaging behavior even in previous research that disentangles Pareto damage from negative reciprocity, this further calls into question the strong link implied by difference-aversion models between sacrifice and inequality-reduction.

B sacrifices to help A 36 percent 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 social-welfare preferences. The fact that, overall, 34 percent of inequality-increasing opportunities to

---

<sup>28</sup>. The eight games where entry could help B are Barc4, Barc6, Barc7, Barc9, Berk14, Berk19, Berk21, and Berk25; the two games where it hurts (given Bs' actual behavior) both are Berk30 and Berk32.

<sup>29</sup>. As the number of participants in each game varied, our percentages in Tables II and III (and elsewhere) could be correspondingly influenced by weighting different games differently. Thus, we also checked these percentages by assigning an equal weight to each game form. We find that the percentages changed very little—with this approach, the penultimate row of Table II becomes 84 percent, 73 percent, 87 percent, 94 percent, and the last row becomes 73 percent, 67 percent, 82 percent, 94 percent.

sacrifice are taken, however, indicates much stronger support for social-welfare preferences than for difference aversion, as reflected in the statistics reported in Tables II and III.

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 when these motivations conflict with self interest. Table V provides such data, and also directly compares social-welfare preferences to difference aversion when the two models make differing predictions about sacrifice.<sup>30</sup>

#### INSERT TABLE V ABOUT HERE

Table V shows that B sacrifices 40 percent of the time when doing so is consistent with social-welfare preferences, but only 8 percent of the time when a sacrifice is inconsistent with social-welfare preferences. The last two rows strongly suggest that social-welfare preferences play a more prominent role in B's decision to sacrifice money although caution must be used, since in our set of games the average sacrifice needed to promote difference aversion is greater than that needed to promote social-welfare preferences.<sup>31</sup>

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 percent of choices can be attributed to social-welfare-maximization, 20 percent to difference aversion, and 10 percent to competitiveness. Results from Charness and Grosskopf [1999], in which a small amount of money was at stake, are perhaps even more telling. While 67 percent of 108 subjects chose (Other,Self) payoffs of (1200,600) over (625,625), only 12 percent chose payoffs of (600,600) over (1200,625). That is, of the two thirds of subjects who had social-welfare rather than difference-averse or competitive preferences, virtually all were willing to sacrifice 25

---

<sup>30</sup> It is clear that competitive preferences do a poor job of explaining sacrifices by B.

<sup>31</sup> Similar evidence from elsewhere also supports our findings about the relative frequency of behavior consistent with social-welfare preferences, difference aversion, and competitiveness. Charness and Grosskopf [1999] found that while about 33 percent of subjects chose (Other,Self) allocations of pesetas of (600,600) over (900,600), about 11 percent of subjects chose (Other,Self) allocations of (400,600) over (600,600). This suggests that about one third 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, 74 percent chose 1200, 7 percent chose a number between 600 and 1200, 10 percent chose 600, and 8 percent chose a number less than 600.

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  $\rho$  and  $\sigma$  in the Section II equations (excluding reciprocity as an explanatory variable by imposing the restriction  $\theta = 0$ ) by performing maximum-likelihood estimation on our binary-response data using the logit regression

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

to determine the values that best match predicted probabilities of play with the observed behavior. The precision parameter  $\gamma$  reflects sensitivity to differences in utility, where the higher the value of  $\gamma$ , the sharper the predictions (see McFadden [1981]). When  $\gamma$  is 0, the probability of either action must be 50 percent; when  $\gamma$  is arbitrarily large, the probability of the action yielding the highest utility approaches 1.

We estimate the values in these equations by imposing further 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.<sup>32</sup> Since we have the same number of observations in each case, in addition to observing the estimated value of  $\gamma$  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

---

<sup>32</sup> We follow an approach similar to that used in Charness and Haruvy [1999].

restrictions they impose. While the allowance for “noise” in maximizing utility provides a proxy of sorts for the heterogeneity that certainly exists among participant’s parameter values, it does so only crudely.<sup>33</sup> 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. Moreover, while 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 is influenced by our choice of games to study.

INSERT TABLE VI ABOUT HERE

Table VI reports the regression results for a variety of different restrictions on the parameter values. 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  $\gamma$  or the log-likelihood, 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 and marginally raising  $\gamma$ . The estimation also confirms that the best-fit single parameter has B putting significant positive weight on A’s payoff.

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  $\sigma$  is positive, rather than negative as

---

<sup>33</sup>. Estimation of separate  $\rho$  and  $\sigma$  values for each individual would be difficult, since the number of observations for each individual is little more than the number of parameters to be estimated.

posited by Bolton [1991].<sup>34</sup> 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 than 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. The estimated value of  $\sigma$  is very small and insignificantly different than 0, so that the estimate of  $\rho$  is virtually identical whether or not  $\sigma$  is included. Along with the small changes in the log-likelihood and  $\gamma$  estimates, we see that neither difference aversion nor social-welfare preferences are significantly better models of distributional preferences than the simpler charity model when reciprocity is ignored.

Overall, the major gains in explanatory power come from allowing  $\rho$  to vary independently of  $\sigma$ . In lines 4-6, the log-likelihood 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)  $\sigma$  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 percent significance level of 3.84. Thus, any explanation for nonpecuniary behavior that relies upon  $\sigma$  being typically negative seems inadequate.

## V. 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

---

<sup>34</sup> But it is only marginally significant, and it seems clear that it is significantly greater than 0 only because of the imposed restriction that  $\rho = 0$ . In some games either parameter could explain behavior, and hence it appears that  $\sigma$  is reflecting the positive value of  $\rho$  in those games.

different choices by A, or no choice by A. By comparing the selection B makes from identical choice sets as a function of the choice A previously did or did not make, we gather very direct evidence on the role of reciprocity in explaining responder behavior. 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, and conclude the section with regression analysis estimating the reciprocity parameter  $\theta$  in the model of Section II.

Our three games in which B chooses between (750,400) and (400,400) are especially informative:

<u>Games With the Choice Between (400,400) and (750,400)</u>		<u>(400,400)</u>	<u>(750,400)</u>
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

Most interesting is the difference between Barc7 and Berk29, which is significant at  $p \approx .00$ . This comparison can be seen as testing the relative strength of positive reciprocity versus difference aversion when self-interest is not implicated. In contrast to the 31 percent of B's who choose (400,400) in the dictator game Berk29, only 6 percent do so following a generous move by A.<sup>35</sup> Because B's choice between (750,400) and (400,400) is a strong invitation to B to pursue difference aversion, and (as we show below) positive reciprocity is nowhere else a strong motivation in our data, the weakness of difference aversion here indicates that it is not a strong factor when in conflict with other social motivations. The results from Barc5 surprised us, as B's 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.

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

<u>Games With the Choice Between (400,400) and (750,375)</u>		<u>(400,400)</u>	<u>(750,375)</u>
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

---

<sup>35</sup> 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 percent vs. 66 percent in the (600,600) vs. (900,600) dictator game in Barcelona was quite similar to the 31 percent vs. 69 percent result in Berk29.

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 social-welfare-maximizing 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.<sup>36</sup> B is in fact *less* likely to sacrifice in pursuit of the social-welfare-maximizing outcome following kind behavior by A than in the dictator context (the difference is collectively significant in a two-tailed test at  $p \approx .14$ ). However, we see evidence of *concern withdrawal*: B is likely to withdraw his willingness to sacrifice to give the social-welfare-maximizing 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 percent to 7 percent (from Barc2 to Barc1) and from 50 percent to 18 percent (from Berk17 to Berk13). These differences are both significant at  $p < .01$ .

The lack of positive reciprocity also shows up when comparing Barc6 to Barc8, the games where B chooses (300,600) vs. (700,500) and (200,700) vs. (600,600), and Berk15 to Berk19.

<b><u>Games Where B Chooses Between (300,600) and (700,500)</u></b>		<b><u>(300,600)</u></b>	<b><u>(700,500)</u></b>
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
<b><u>Games Where B Chooses Between (200,700) and (600,600)</u></b>		<b><u>(200,700)</u></b>	<b><u>(600,600)</u></b>
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:

<b><u>Games Where B Chooses Between (0,800) and (400,400)</u></b>		<b><u>(0,800)</u></b>	<b><u>(400,400)</u></b>
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

---

<sup>36</sup> We note in passing that this lack of positive reciprocity is consistent with results from trust games (e.g., Berg, Dickhaut, and McCabe [1995]) and gift-exchange games, which are often interpreted as positive reciprocity. The decision by responders to "return" some money given to them seems typically consistent with the type of sharing we find in dictator games.

The results from Berk14, where 55 percent choose (400,400) over (0,800) in contrast to the 22 percent 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  $p \approx .01$ .<sup>37</sup>

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:

<u>Games Where B Chooses Between (800,200) and (0,0)</u>		<u>(800,200)</u>	<u>(0,0)</u>
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

Zero of 36 subjects 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  $p < .06$ . But together with Barc11 and Barc22, the other games where B can sacrifice to hurt A, we find relatively little negative 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 percent of Bs do so. Doing so contradicts social-welfare preferences in Barc5 and both social-welfare 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.<sup>38</sup>

---

<sup>37</sup>. 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.

<sup>38</sup>. 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.

As a first pass at summarizing the evidence on reciprocity, Table VII specifies a distributional parsing of Pareto damage and positive sacrifice in terms of how A has treated B. It 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 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  $p \approx .02$ .

INSERT TABLE VII ABOUT HERE

B sacrifices to help A 36 percent 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, particularly concern withdrawal, 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 percent to 36 percent. By contrast, when A has hurt B, helping behavior reduces to 11 percent. Hence, we see that violation of social-welfare 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 formal model developed in Appendix A.

To give a more precise analysis of the role of reciprocity, consider the bottom two lines of Table VI, which remove the constraint on our regression analysis that  $\theta = 0$ , the parameter measuring how “social-welfare misbehavior” by A affects B’s weight on A’s payoff. The level of precision ( $\gamma$ ) is much higher for each of these reciprocity regressions than for the self-interest model regression. More importantly, lines 5, 6, and 7 together show that reciprocal motivations play a greater role in behavior than do non-reciprocal preferences to either help or hurt those who are ahead.<sup>39</sup> Comparing lines 5 and 7 shows that the estimate of  $\theta$  is significantly negative; the

---

<sup>39</sup>. Although our definition of “misbehaved” builds on social-welfare 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]. While games such as Berk28 and Berk32 look highly suggestive to us as indications that it is violations of social-welfare preferences that trigger retaliation, our formal analysis does not support either model against the other.

likelihood-ratio test gives  $\chi^2 = 9.18$  ( $p \approx .00$ ).<sup>40</sup> Note that this is much stronger than allowing  $\sigma$  to vary; comparing lines 6 and 7, for instance, 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,  $p \approx .27$ ). Once one includes reciprocity in the regressions, allowing for players with lower payoffs to care (*positively*) intrinsically about the payoff of the other player has some, but not much, explanatory power.

All said, an analysis over a broad range of games indicates that reciprocity considerations are an important component of behavior.

## VI. Multi-Person Games

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 II 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. 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

---

<sup>40</sup>. Although the multiple-observation caveat to statistical significance may still apply, a comparison between the likelihood-ratio tests nevertheless indicates that allowing  $\theta$  to be nonzero has a much greater impact than allowing  $\sigma$  to be nonzero.

interest in light of our two-player results. Our results above show that about 50 percent of B's choose (400,400) over (750,375), consistent with those subjects being difference-averse, self-interested, or competitive. None of these motivations would explain the choice by C's to choose (400,400), suggesting that a good proportion of Bs are choosing (400,400) for "disinterested", social-welfare reasons rather than just to get more money. Barc10 and Barc12 together show that social efficiency is not the only distributional driving force, as (1200,0,x) is more socially efficient than (750,375,x), but is chosen much less frequently ( $p \approx .01$ ).

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 surprising to posit that subjects were indifferent to the allocations among others, we designed Berk24 as a simple and direct test of their hypothesis.<sup>41</sup>

Berk24 demonstrates that subjects care about the allocation among other parties: 54 percent 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 are consistent with both social-welfare preferences and Fehr and Schmidt's [1999] difference aversion, but are inconsistent with Bolton and Ockenfels's [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 percent is a higher proportion than we found are inclined to sacrifice *nothing* to eliminate disadvantageous inequality against themselves. Hence, our results suggest that

---

<sup>41</sup> 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 consequence of a rejection) low payoffs (in their experiment 2) to innocent third parties also goes against social-welfare preferences, and can only be rationalized as a strong willingness to punish the proposer.

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 percent to 80 percent 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 percent of the time overall. This difference in choices is significant at  $p \approx .00$ . Because the differences in distributional consequences of behavior were minor, we do not consider this a good test of the relative strength of distributional vs. reciprocity motivations. Rather, it shows that reciprocity can overwhelm distributional concerns in some circumstances.

## VII. Summary and Conclusion

This paper continues recent research delineating the nature of social preferences in laboratory behavior. Our results suggest that the apparent adequacy of recent difference-aversion models has likely been an artifact of powerful and decisive confounds in the games used to construct these models.<sup>42</sup> We find a strong degree of respect for social efficiency, tempered by concern for those less well off.

Our data are rich and complicated. We have not analyzed them exhaustively, nor incorporated all observed patterns into our formal models. We have not tested for individual differences and correlation across games, and neither our analysis nor our model deals with heterogeneity of subject preferences. Nor does our model capture evidence in the data for what might be called a *complicity effect*: The mere fact of one player being involved in a decision seems to make the other player more self-interested. Perhaps impulses towards pro-social behavior are diminished when an agent does not feel the full responsibility for an outcome.<sup>43</sup>

One benefit of the sort of simple games we study 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 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 experimental research would be to design ways to more directly discern participants’ beliefs about the intentions or likely behavior of other subjects.<sup>44</sup>

Most of our evidence strongly replicates others’ findings that positive reciprocity has virtually no explanatory power in many of the conventional games studied. Yet the data from

---

<sup>42</sup> 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.

<sup>43</sup> See Charness [2000] for a discussion of *responsibility alleviation*, and a review of papers with evidence related to the phenomenon.

<sup>44</sup> 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.

one game call into question the generality of this conclusion. In the game Barc7, the reader will surely recall, A can forego a (750,0) outcome to give B the choice between (750,400) and (400,400). Only 6 percent of Bs choose (400,400). This is only one game in one session with 36 subjects, but the findings are provocative: Together with the fact that 30 percent choose (400,400) following either no move or a nasty move by A, the 6 percent suggests a possible form of positive reciprocity that may be very strong compared to difference aversion. 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”. Even if researchers eventually conclude that many subjects are difference-averse when neutral, it may be necessary to develop models where positive feelings towards another subject can lead them to be unwilling to harm that subject in pursuit of difference aversion.

We are especially keen to understand the behavior of subjects in Barc7 and games like it because we think they are a simplified form of very common social and economic situations: 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 her. Indeed, we suspect situations resembling this game are far more common in the real world than situations resembling the ultimatum game. Such games capture phenomena such as employer-employee bargaining, where any accepted take-it-or-leave-it wage offer by an employer will be followed by opportunities for employees to undermine or to enhance the employer’s profits. More generally, opportunities to affect another’s payoff at small cost to oneself is important economically, and suggests that reciprocity motives are likely to loom large.

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.

Department of Economics, University of California, Santa Barbara

Department of Economics, University of California, Berkeley

## APPENDIX A: A MORE GENERAL MODEL

In this Appendix, we construct a model that integrates social-welfare preferences with reciprocity—both concern withdrawal and negative reciprocity—into a multi-person model of social preferences. This model gives an interpretation of the underlying motives that play a role in the two-player games, as well as extending our analysis to more than two players. The model we develop omits the many other factors that seem to play a role in some subjects’ choices, and presents general functional forms allowing many degrees of freedom. Moreover, the model is complicated, as it formulates reciprocity as a *de facto* psychological Nash equilibrium as defined in the framework developed by Geanakoplos, Pearce, and Stacchetti (1989). All said, we do not see it as being primarily useful in its current form for calibrating experimental data, but rather as providing progress in conceptualizing what we observe in experiments.

We first define reciprocity-free preferences in two steps, and add the reciprocity component later. First consider a “disinterested social-welfare criterion:

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

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.<sup>45</sup> Setting  $\delta = 1$  corresponds to a pure “maximin” or “Rawlsian” criterion, whereby social welfare is measured solely according to how well off the least well off is. Setting  $\delta = 0$  corresponds to total-surplus maximization.

Now consider Player  $i$ ’s payoffs as a weighted sum of this disinterested social-welfare criterion (which includes his own payoff) and his own payoff:

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

where  $\lambda \in [0,1]$  measures how much Person  $i$  cares about pursuing the social welfare versus his own self-interest. Setting  $\lambda = 1$  corresponds to purely “disinterested” preferences, in which

---

<sup>45</sup> It would be more realistic (but 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.

players care no more (or less) about her own payoffs than others' payoffs, and setting  $\lambda = 0$  corresponds to pure self interest.

To see the connection between these preferences and the two-player specification of Section II when we ignore reciprocity, note that it reduces to:

$$\begin{aligned} V_B(\pi_A, \pi_B) &\equiv (1-\lambda\delta)\pi_B + \lambda\delta\pi_A \quad \text{when } \pi_B \geq \pi_A, \\ V_B(\pi_A, \pi_B) &\equiv \pi_B + \lambda(1-\delta)\pi_A \quad \text{when } \pi_B \leq \pi_A. \end{aligned}$$

If we normalize these two equations by dividing by  $1 + \lambda(1-\delta)$ , so that as in the Section II model

$$V_B = \pi_B \text{ when } \pi_B = \pi_A, \text{ we see that } \rho = \frac{\lambda}{1 + \lambda(1-\delta)} \text{ and } \sigma = \frac{\lambda(1-\delta)}{1 + \lambda(1-\delta)}.$$

These equations have a natural interpretation. When  $\lambda$  increases (meaning B puts more weight on the social good and less on his own material payoffs), both  $\rho$  and  $\sigma$  increase. When  $\delta$  increases (so that B puts relatively more weight on the maximin component and less on total surplus), then  $\rho$  increases and  $\sigma$  decreases. That is, both parameters move in the direction of more concern for the person who has a lower payoff, whether this is A or B. Indeed, looking at  $\frac{\rho}{\sigma} = \frac{1}{1-\delta}$  makes this even clearer. Increasing  $\rho$  and  $\sigma$  by the same proportion indicates an decrease in self-interestedness,  $\lambda$ , whereas increasing the ratio indicates an increase in  $\delta$ .

Before defining the full model that incorporates reciprocity, we first define an equilibrium notion based just on these social-welfare preferences. 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  $\pi_i(a_1, \dots, a_N)$  represents Player  $i$ 's payoffs given actions  $(a_1, \dots, a_N)$ .

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

Because  $\pi_1, \dots, \pi_n$  are continuous in the players' actions, the functions  $V_i(\pi)$  are well-defined and continuous in the players' actions. Hence, a SWE always exists. SWE 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—SWE can provide more explanatory power than other distributional models.<sup>46</sup> But SWE also serves as a foundation for our reciprocity model. Indeed, with an important restriction placed on the parameters of our model, every SWE will be an equilibrium in our full reciprocity model.

To begin to incorporate reciprocity, consider a strategy profile  $s \equiv (s_1, s_2, \dots, s_n)$ , as well as a *demerit profile*,  $d \equiv (d_1, \dots, d_n)$ , where  $d_k \in [0, 1]$  for all  $k$ . Below  $d$  will be determined endogenously. For now,  $d_k$  can be interpreted roughly as a measure of how much Player  $k$  deserves, where the higher the value of  $d_k$ , the less others think Player  $k$  deserves. With this interpretation, we define players’ preferences as a function of both their underlying social-welfare preferences and how they feel about other players:

$$U_i(s, d) \equiv (1-\lambda) \cdot \pi_i + \lambda \cdot [\delta \text{Min}[\pi_i, \text{Min}_{m \neq i} \{ \pi_m + b d_m \}] + (1-\delta) \cdot (\pi_i + \sum_{m \neq i} \max[1 - k d_m, 0] \pi_m) - f \sum_{m \neq i} d_m \cdot \pi_m],$$

where  $b$ ,  $k$ , and  $f$  are non-negative parameters of the model. The key new aspect to these preferences is that the greater is  $d_j$  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 can be seen most starkly by setting  $f = 0$  and assuming that  $b$  and  $k$  are very large. Then the preferences  $U_i(s, d)$  imply that Player  $i$  maximizes the disinterested social-welfare-maximizing allocation among all those other players for which  $d_j = 0$ —that is, among all the deserving others—and ignores the payoffs among those who are undeserving.

We begin endogenizing the demerits  $d$  by defining, for every profile of strategies  $s_i$  and demerits  $d_i$  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:

---

<sup>46</sup> As with other distributional models, one could readily define a range of solution concepts with respect to social-welfare preferences. 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.

$$S_i^*(s_{-i}, d_i; g) \equiv \{s_i \in S_i \mid s_i \in \operatorname{argmax} \{(1-g) \pi_i + g[\delta \operatorname{Min}[\pi_i, \operatorname{Min}_{m \neq i} \{\pi_m + b d_m\}] + (1-\delta) [\sum_{j=1 \dots n} \pi_j - k \sum_{m \neq i} d_m \cdot \pi_m] - f \sum_{m \neq i} d_m \cdot \pi_m]\}\},$$

where  $\pi$  is the profile of material payoffs. “Typically”,  $S_i^*(s_{-i}, d_i, g)$  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  $g_i(s, d)$  be some upper hemi-continuous and convex-valued correspondence from  $(s, d)$  into the set  $[0, 1]$  such that  $g_i(s, d) \approx \{g \mid s_i \in S_i^*(s_{-i}, d_{-i}, g)\}$ .<sup>47</sup> The function  $g_i(s, d)$  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 social-welfare criterion) by choosing  $s_i$  in response to  $s_{-i}$ , given that she has disposition  $d_{-i}$  towards the other players. Except for a technical fix to assure that  $g_i(s, d)$  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 all other players are behaving well. In such cases, our model does not pin down a particular functional form, and hence is quite unrestrictive.

The unrestrictiveness of our model in such cases is partly for technical convenience and because it doesn’t matter much.<sup>48</sup> But we don’t restrict  $g_i(s, d)$  when  $\{g \mid s_i \in S_i^*(s_{-i}, d_{-i}, g)\}$  is empty also because we don’t feel we know the right psychology for how people interpret

---

<sup>47</sup>. More exactly, for values  $(s, d)$  where  $\{g \mid s'_i \in S_i^*(s'_{-i}, d'_{-i}, g)\}$  is non-empty for all  $(s', d')$  in a large-enough neighborhood of  $(s, d)$ , then  $g_i(s, d) = \{g \mid s_i \in S_i^*(s_{-i}, d_{-i}, g)\}$ . The full definition of  $g_i(s, d)$  is as follows. Let  $\epsilon(s, d)$  be the neighborhood around  $(s, d)$  with all components within  $\epsilon > 0$  of  $(s, d)$ . We then let  $g_i(s, d)$  be any upper hemi-continuous and convex-valued correspondence such that  $\{g \mid s_i \in S_i^*(s_{-i}, d_{-i}, g)\} \subseteq g_i(s, d) \subseteq G(\epsilon, s, d)$ , where  $G(\epsilon, s, d)$  is the convex hull of  $\{g \mid t_i \in S_i^*(t_{-i}, \chi_{-i}, g) \text{ for some } (t, \chi) \in \epsilon(s, d)\}$  if  $\{g \mid t_i \in S_i^*(t_{-i}, \chi_{-i}, g) \text{ for some } (t, \chi) \in \epsilon(s, d)\}$  is non-empty, and  $G(\epsilon, s, d) = [0, 1]$  if  $\{g \mid t_i \in S_i^*(t_{-i}, \chi_{-i}, g) \text{ for some } (t, \chi) \in \epsilon(s, d)\}$  is empty. This is entirely unrestrictive when  $\{g \mid t_i \in S_i^*(t_{-i}, \chi_{-i}, g) \text{ for some } (t, \chi) \in \epsilon(s, d)\}$  is empty. But, assuming as we do that  $\epsilon$  is small,  $g_i(s, d) \approx \{g \mid s_i \in S_i^*(s_{-i}, d_{-i}, g)\}$  when  $\{g \mid s_i \in S_i^*(s_{-i}, d_{-i}, g)\}$  is non-empty. 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.

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  $g_i(s,d)$  to some selflessness standard,  $\lambda^*$ —the weight they feel a decent person *should* put on social welfare. Specifically, we assume that other players' level of animosity towards Player  $i$  corresponds to  $r_i(s,d, \lambda^*) \in \{\text{Min}[g - \lambda^*, 0] \mid g \in g_i(s,d)\}$ . That is, whenever  $\text{Max}\{g \mid g \in g_i(s,d)\} < \lambda^*$ , 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 social-welfare preferences with  $\lambda = \lambda^*$ . When  $\text{Min}\{g \mid g \in g_i(s,d)\} \geq \lambda^*$ , others will feel no animosity towards Player  $i$ . Requiring elements of  $r_i(s,d, \lambda^*)$  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  $\lambda, \lambda^*, \delta, b, k, f$  and correspondence  $g_i(s,d)$  if there exists  $d$  where, for all  $i$ , there exists  $g_i \in g_i(s,d)$  such that

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

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.<sup>49</sup>

---

<sup>48.</sup> It would be more problematic if we were to use it to predict non-equilibrium outcomes, or outcomes for heterogeneous preferences.

<sup>49.</sup> 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

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. 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  $(s,d)$  into itself defined by the best-response correspondences  $s_i \in \text{Argmax } U_i(s,d)$  and the demerit functions  $d_i(s,d) \in \{r \mid \exists g \in g_i(s,d) \text{ such that } r = \text{Max}[\lambda^* - g_i, 0]\}$ . 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  $U_i(s,d)$  and the expected-utility structure,  $\text{Argmax } U_i(s,d)$  is upper hemi-continuous and convex-valued. The component  $d_i(s,d)$  is upper hemi-continuous and convex-valued because  $g_i(s,d)$  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 social-welfare equilibria would play a prominent role in our model. Because of the reciprocity component in preferences (operative when  $d_k > 0$  for some  $k$ ), reciprocal-fairness equilibria might not correspond to social-welfare 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  $\lambda^*$  is very high—it may be that such negative outcomes are the only RFE. But if all players' intrinsic desire,  $\lambda$ , to pursue the social good rather than self interest is at least as great as the standard,  $\lambda^*$ , to which people hold each other, then all social-welfare equilibria will be reciprocal-fairness equilibria:

---

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.

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

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

Theorem 2 indicates that SWE 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 heterogeneous preferences, where  $\lambda < \lambda^*$  for some of the participants, so that some bad behavior, and corresponding retaliation, may be observed.

Despite the unrestrictiveness of reciprocal-fairness equilibrium in some ways, it is clearly too restrictive in other respects. 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. And the model clearly omits patterns of behavior that seem apparent in the data, such as complicity effects. By assuming homogeneous preferences, it rules out even a minority of subjects being motivated by preferences such as difference aversion. We think any prospects for good-fitting models will eventually have to account better for such heterogeneity than we have done in this paper. We can also think of specific examples—such as Barc5, where we seem to observe no negative reciprocity (when compared to Berk29, say) even when it is free—that, if they turn out to be consistent patterns, would raise problems for our paper.

## 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.

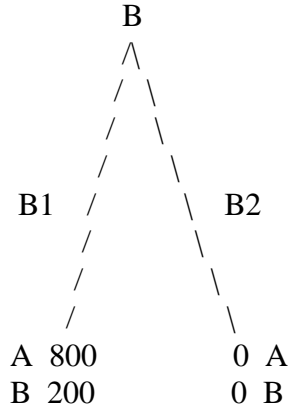
Please feel free to ask questions at any point if you feel you need clarification. Please do so by raising your hand. Please DO NOT attempt to communicate with any other participants in the session until the session is concluded.

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

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.



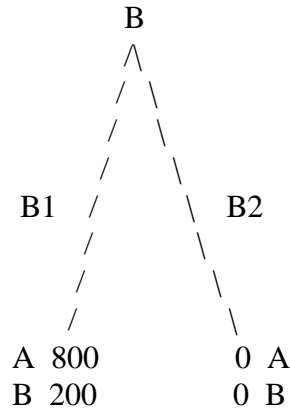
DECISION

I understand I have no choice in this game \_\_\_\_\_

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:

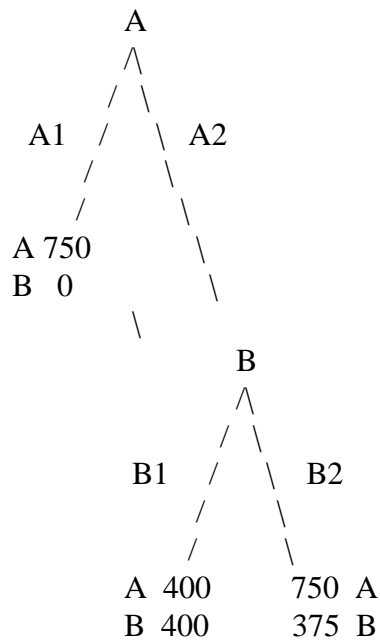
B1

B2

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.



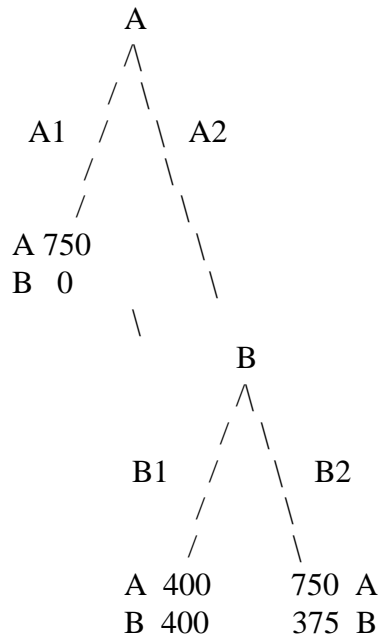
DECISION

I choose:                    A1                    A2

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

The role-reversal data for each of the 19 games is shown below. The (two-tailed) p-value used reflects the percentage of time that a difference in rates as large as the one observed would occur randomly:

<b><u>For each type of behavior as A, did the person help A as B?</u></b>		<b><u>if Out</u></b>	<b><u>if Enter</u></b>	<b><u>p-value</u></b>
<b><u>Helping A Doesn't Affect B's Payoff</u></b>				
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
<b><u>Helping A is Costly to B</u></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	
<b><u>Helping A is Beneficial to B</u></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	
	<i>N</i>	<i>Prefs.</i>	<i>N</i>	<i>Prefs.</i>	<i>N</i>	<i>Prefs.</i>	<i>N</i>	<i>Prefs.</i>
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	
	<i>N</i>	<i>Prefs.</i>	<i>N</i>	<i>Prefs.</i>	<i>N</i>	<i>Prefs.</i>	<i>N</i>	<i>Prefs.</i>
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**

				<u>Maximize</u>	<u>Sacrifice</u>	
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**

				<u>Maximize</u>	<u>Sacrifice</u>	
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

**Table I. Game-by-Game Results**

<b>Two-Person Dictator Games</b>		<b>Left</b>	<b>Right</b>				
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				
Berk26 (32)	B chooses (0,800) vs. (400,400)	.78	.22				
<b>Two-Person Response Games—B's Payoffs Identical</b>		<b>Out</b>	<b>Enter</b>	<b>Left</b>	<b>Right</b>		
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		
<b>Two-Person Response Games—B's Sacrifice Helps A</b>		<b>Out</b>	<b>Enter</b>	<b>Left</b>	<b>Right</b>		
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		
<b>Two-Person Response Games—B's Sacrifice Hurts A</b>		<b>Out</b>	<b>Enter</b>	<b>Left</b>	<b>Right</b>		
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		
<b>Three-Person Dictator Games</b>		<b>Left</b>	<b>Right</b>				
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				
<b>Three-Person Response Games</b>		<b>Out</b>	<b>In</b>	<b>Left</b>	<b>Right</b>		
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		

*Numbers in parentheses show (A,B) or (A,B,C) money payoffs*

**Table II. Consistency of Behavior with Distributional Models**

	Total # Observations	Narrow Self interest	Competitive	Difference Aversion	Social- welfare
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 III. Consistency of Behavior with Distributional Models  
When the Prediction is Unique**  
(Entries are chances taken over total chances)

Class of Games	Narrow Self Interest	Competitive	Difference Aversion	Social Welfare
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 IV. B's Sacrifice Rate by Effect on Inequality**

Class of Games	Sacrifices/Chances	Probability of Sacrifice
<b>Games allowing Pareto-damage</b>	<b>59/357</b>	<b>17%</b>
Decreases inequality	34/228	15%
No effect on inequality	4/35	11%
Increases inequality	21/94	22%
<b>Games where sacrifice helps A</b>	<b>199/546</b>	<b>36%</b>
Decreases inequality	99/212	47%
No effect on inequality	8/68	12%
Increases inequality	92/266	35%
<b>All Games</b>	<b>268/903</b>	<b>30%</b>
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 V. Distributional Models as Explanations for B's 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 SWP	191/478	40%
Inconsistent with SWP	22/259	8%
Consistent with DA but not SWP	9/120	8%
Consistent with SWP 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 VI. Regression estimates for B behavior (N=903)**

Model	Restrictions	$\rho$	$\sigma$	$\theta$	$\gamma$	LL
Self-interest	$\rho = \sigma = \theta = 0$	-	-	-	.004 (9.07)	-593.4
Single parameter— “altruism”	$\rho = \sigma, \theta = 0$	.212 (7.20)	.212 (7.20)	-	.005 (8.65)	-574.5
Single parameter— “behindness aversion”	$\rho = \theta = 0$	-	.118 (1.76)	-	.004 (8.53)	-591.5
Single parameter— “charity”	$\sigma = \theta = 0$	.422 (25.5)	-	-	.014 (11.6)	-527.9
$\rho, \sigma$ model without reciprocity	$\theta = 0$	.423 (25.5)	-.014 (-0.73)	-	.014 (11.6)	-527.7
“Reciprocal charity”	$\sigma = 0$	.425 (27.9)	-	-.089 (-2.98)	.015 (11.3)	-523.7
$\rho, \sigma$ model with reciprocity	none	.424 (28.3)	.023 (1.10)	-.111 (-3.11)	.015 (11.6)	-523.1

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

**Table VII. B's Response as a function of A's help or harm**

Class of Games	Sacrifices/Chances	Probability of Sacrifice
All games allowing Pareto-damage	<b>59/357</b>	<b>17%</b>
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	<b>199/546</b>	<b>36%</b>
A helped B	100/278	36%
A had no play*	88/202	44%
A hurt B in violation of SWP	7/66	11%

\*We include Berk18 in this classification, since A's decision to enter was obvious and universal.

## REFERENCES

- Andreoni, James and John Miller, "Giving According to GARP: An Experimental Study of Rationality and Altruism," mimeo, 1998, *Econometrica*, forthcoming.
- Andreoni, James, Paul Brown, and Lise Vesterlund, "What Makes an Allocation Fair? Some Experimental Evidence," mimeo, 1999.
- Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, X (1995), 122-42.
- Blount, Sally, "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior and Human Decision Processes*, LXIII (1995), 131-144.
- Bolton, Gary and Axel Ockenfels, "Strategy and Equity: An ERC-analysis of the Güth-van Damme game," *Journal of Mathematical Psychology*, XLII (1998), 215-226.
- \_\_\_ and \_\_\_, "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, XC (2000), 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*, XV (2000), 367-387.
- Bolton, Gary, Jordi Brandts, and Axel Ockenfels, "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game," *Experimental Economics*, I (1998), 207-219.
- Brandts, Jordi and Gary Charness, "Hot vs. Cold: Sequential Responses in Simple Experimental Games," *Experimental Economics*, II (2000), 227-238.
- \_\_\_ and \_\_\_, "Retribution in a Cheap-talk Game," mimeo, 1999.
- Brandts, Jordi and Carles Solà, "Reference Points and Negative Reciprocity in Simple Sequential Games," mimeo, 1998, *Games and Economic Behavior* forthcoming.
- Cason, Timothy and Vai-Lam Mui, "Social Influence in the Sequential Dictator Game," *Journal of Mathematical Psychology*, XLII (1998), 248-265.
- Charness, Gary, "Attribution and Reciprocity in an Experimental Labor Market," mimeo, 1996.
- \_\_\_, "Responsibility and Effort in an Experimental Labor Market," *Journal of Economic Behavior and Organization*, XLII (2000), 375-384.
- Charness, Gary and Brit Grosskopf, "Relative Payoffs and Happiness: An Experimental Study," *Journal of Economic Behavior and Organization*, XLV (2001), 301-328.
- Charness, Gary and Ernan Haruvy, "Altruism, Fairness, and Reciprocity in a Gift-exchange Experiment: An Encompassing Approach" mimeo, 1999, *Games and Economic Behavior*, forthcoming.

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

\_\_\_ and \_\_\_, "Some Simple Tests of Social Preferences," University of California at Berkeley, mimeo, 2000.

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

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

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

Engelmann, Dirk and Martin Strobel, "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," mimeo, 2001.

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

Falk, Armin, Ernst Fehr, and Urs Fischbacher, "On the Nature of Fair Behavior," mimeo, 1999, *Economic Inquiry*, forthcoming.

Fehr, Ernst and Klaus Schmidt, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, CXIV (1999), 817-868.

Frohlich, Norman and Joseph Oppenheimer, "Beyond Economic Man: Altruism, Egalitarianism, and Difference Maximizing," *Journal of Conflict Resolution*, XXVIII (1984), 3-24.

\_\_\_ and \_\_\_, *Choosing Justice* (Berkeley: University of California Press 1992).

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

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

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

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

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

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

McCabe, Kevin, Mary Rigdon, and Vernon Smith, "Positive Reciprocity and Intentions in Trust Games," mimeo, 2000.

McFadden, Daniel, "Econometric Models of Probabilistic Choice," in *Structural Analysis of Discrete Data with Econometric Applications*, Charles Manski and Daniel McFadden, eds. (Cambridge, MA: M.I.T. Press, 1981).

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

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

Roth, Alvin, "Bargaining Experiments," in *Handbook of Experimental Economics*, J. Kagel and A. Roth, eds. (Princeton, NJ: Princeton University Press, 1995).

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

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

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