Title
Social Preferences: Some Simple Tests and a New Model

Authors
Charness, Gary
Rabin, Matthew

Publication Date
2000-06-05

Social Preferences:
Some Simple Tests and a New Model

First Draft: October 1999
This Draft: January 2000

Gary Charness
Department of Economics and Business, Universitat Pompeu Fabra—Barcelona
Department of Economics, University of California—Berkeley

Matthew Rabin
Department of Economics, University of California—Berkeley

Abstract: Departures from pure self interest in economic experiments have recently inspired models of "social preferences". We conduct experiments on simple two-person and three-person games with binary choices that test these theories more directly than the array of games conventionally considered. Our experiments show strong support for the prevalence of "quasi-maximin" preferences: People sacrifice to increase the payoffs for all recipients, but especially for the lowest-payoff recipients. People are also motivated by reciprocity: While people are reluctant to sacrifice to reciprocate good or bad behavior beyond what they would sacrifice for neutral parties, they withdraw willingness to sacrifice to achieve a fair outcome when others are themselves unwilling to sacrifice. Some participants are averse to getting different payoffs than others, but based on our experiments and reinterpretation of previous experiments we argue that behavior that has been presented as "difference aversion" in recent papers is actually a combination of reciprocal and quasi-maximin motivations. We formulate a model in which each player is willing to sacrifice to allocate the quasi-maximin allocation only to those players also believed to be pursuing the quasi-maximin allocation, and may sacrifice to punish unfair players.

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

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

Acknowledgments: We thank Jordi Brandts, Antonio Cabrales, Colin Camerer, George Loewenstein, Chris Shannon, and seminar participants at the June 1999 MacArthur Norms and Preferences Network, the 1999 Russell Sage Foundation Summer Institute in Behavioral Economics, Harvard Economics Department, Stanford GSB, Berkeley, and UCSD for helpful comments, and Kitt Carpenter, David Huffman, Chris Meissner, and Ellen Myerson for valuable research assistance. For financial support, Charness thanks the Spanish Ministry (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 / 549 Evans Hall #3880 / University of California, Berkeley / Berkeley, CA 94720-3880. E-mail: charness@econ.berkeley.edu. Web page: http://bonvent.upf.es/home/charness/. Matthew Rabin / Department of Economics / 549 Evans Hall #3880 / University of California, Berkeley / Berkeley, CA 94720-3880. E-mail: rabin@econ.berkeley.edu. Web page: http://elsa.berkeley.edu/rabin/index.html.

1. Introduction

Participants in experiments frequently choose actions that do not maximize their own monetary payoffs when those actions affect the payoffs of others. People sacrifice money in bargaining to punish those who mistreat them, share money with other parties who have no say in allocations, and make voluntary contributions to public goods.

To capture such departures from narrow self interest, several models of *social preferences* have recently been proposed. These models assume that people not only have a self-interested desire to receive high payoffs, but are also concerned about the payoffs of others. In this paper, we report findings from a series of simple experiments that test existing theories more directly than the conventional array of games, and formulate a new model to capture patterns of behavior that previous models don't explain.[1]

Existing models of social preferences fall into two categories: Those that assume people care solely about the distribution of payoffs, and those that assume people are also motivated to reciprocate the intentional actions of others. We review such models and previous experimental evidence of social preferences in Section 2. In the category of distributional preferences, Loewenstein, Thompson, and Bazerman (1989), Bolton (1991), Bolton and Ockenfels (1999), and Fehr and Schmidt (1999) develop models in which a person is motivated to reduce differences in payoffs between himself and others, sacrificing to help others when ahead, but also making Pareto-damaging sacrifices—actions that hurt some and help none—when behind. We label such preferences "difference aversion". An alternative model of distributional preferences, related to the ideas discussed in Yaari and Bar-Hillel (1984) and Andreoni and Miller (1998), assumes that people don't dislike differences in payoffs *per se*, but care more about helping low-payoff people than high-payoff people. Combining the assumption that people are motivated to maximize the payoff to the minimum-payoff person with the desire to increase total payoffs yields what we shall call "quasi-maximin preferences". Such preferences do not induce Pareto-damaging behavior.

---

[1] As we discuss in Section 2, there are other recent papers that construct straightforward and easy-to-interpret tests, including Kagel and Wolfe (1999), Kritikos and Bolle (1999), and Charness and Grosskopf (1999).

In the category of reciprocity preferences, Rabin (1993) developed a model in which one player wishes to increase or decrease another player's payoffs based on her beliefs about whether the other player is treating her fairly, and Dufwenberg and Kirchsteiger (1998) modify and extend that model so as to better explain behavior in sequential games. Falk and Fischbacher (1998) also consider sequential games and combine reciprocity of the sort captured by Rabin (1993) with difference aversion of the sort captured by Fehr and Schmidt (1999). Levine (1998) models reciprocity by assuming that a person's desire to increase or decrease another's payoff depends on his beliefs about the other's inherent degree of altruism.

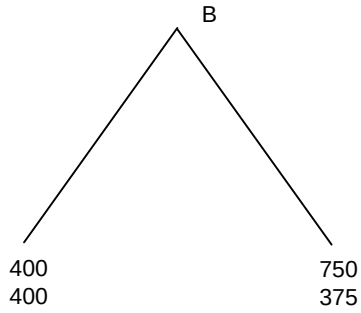
Different researchers have used different models to explain the same data. Consider the divide-a-pie dictator game, where one player is given the unilateral choice on how to split a sum of money with a second player. Andreoni and Miller (1998) explain sharing in dictator games by something akin to quasi-maximin preferences, whereas Bolton and Ockenfels (1999) and Fehr and Schmidt (1999) explain such sharing by difference aversion. Rabin (1993) interprets cooperation in a symmetric prisoners' dilemma as reciprocity, whereas Fehr and Schmidt (1999) and Bolton and Ockenfels (1999) interpret it as difference aversion.

Similarly, consider the ultimatum game: Here, following a proposal by one player on how to split a sum of money, the second player can either accept the proposed split or reject it and thus assign each player a zero material payoff. Rabin (1993) explains rejections as retaliation against unfair treatment, whereas Fehr and Schmidt (1999) and Bolton and Ockenfels (1999) explain such Pareto-damaging behavior by difference aversion. Indeed, one motivation for our research was skepticism about recent models that interpret Pareto-damaging behavior, such as rejections in the ultimatum game, as coming primarily from difference aversion rather than retaliatory preferences. Our intuition was different, and we observed that virtually all evidence in favor of this interpretation was based on the behavior of players in whom retaliatory motivations had been triggered, and allowed only for Pareto-damaging behavior that necessarily involved inequality reduction.

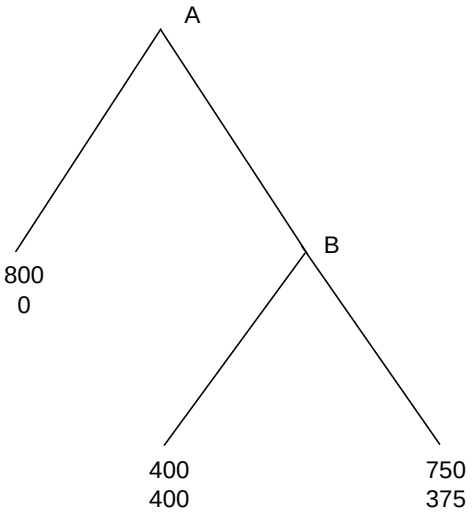
We describe our experimental designs, meant to extend recent efforts by others to differentiate among existing models, in Section 3. We study simple two- and three-player binary-choice games. We study both dictator games, where one person makes a choice that unilaterally determines the distribution of payoffs, and response games, where a first mover

chooses either an outside option or to give the responder a choice between two alternatives. We tested 29 different game forms, with 467 participants, making 1697 decisions. We had responders choose an alternative before knowing whether the first mover made their choice relevant. In most sessions, participants played a series of two to four games, one at a time, knowing that only a randomly-chosen subset of decisions would be used to determine payoffs. Each game was played twice by having players change roles while re-matched with new partners.

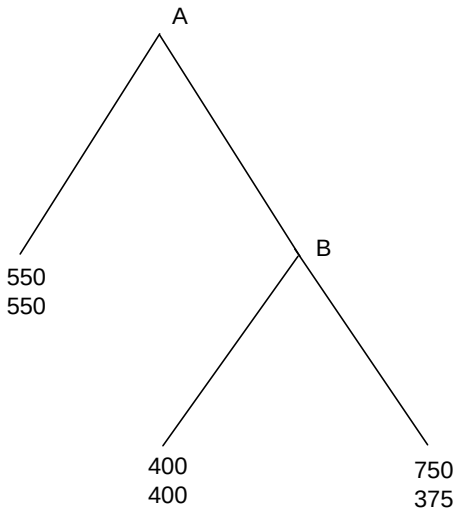
We chose our constellation of games so as to maximize our ability to differentiate among existing hypotheses about players' motivations. To get a sense for our approach and results, consider Games 1-3:



Game 1



Game 2



Game 3

In "Game" 1, Player B unilaterally determines both his own payoff and Player A's payoff by choosing between two actions yielding payoffs (750,375) and (400,400), where the first entry is A's payoff and the second is B's. Difference aversion says B strongly prefers (400,400) over (750,375)—both on self-interest grounds and because he dislikes coming out behind. Depending on the weight placed on self interest and the maximin criterion versus maximizing social surplus, quasi-maximin preferences predict B might choose either (400,400) or (750,375). We find that about 50% of participants chose (750,375). The results from this and all other games are reported in Section 4.

In similarly simple tests, Andreoni and Miller (1998) and Charness and Grosskopf (1999) find similar results, with significant numbers of participants opting for inequality-increasing sacrifices to help others. Our other games yielded similar findings: For instance, we found that 69% of subjects chose (750,400) over (400,400). From such clean tests of difference aversion, we tentatively conclude that no more than a third of people behave at all consistently with difference aversion when evaluating outcomes in which they get lower payoff than others; two thirds of people have *opposite* preferences. We also argue below that quasi-maximin preferences are more "robust" than difference aversion: Those who have quasi-maximin preferences often pursue those preferences when in conflict with other goals—such as self interest or reciprocity—whereas most of those pursuing difference aversion abandon those preferences when in conflict with these same goals.

To test the role of reciprocity, we study simple response games where B's choice follows a move by A to forego an outside option, and compare B's behavior to his behavior given the same choice when A either had foregone a different outside option or had no outside option. Game 2, for instance, involves the same choice by B as in Game 1, but follows an unambiguously kind move by A to forego an (800,0) outcome. Only 38% chose (750,375), which is less than the proportion who chose this outcome in Game 1. That is, B is *less* likely to help A when A has acted kindly than when A did not make a choice. These and our other findings reinforce recent experimental evidence that players are not significantly more willing to sacrifice to help others who have treated them favorably than to help others who have treated them neither favorably nor unfavorably. Behavior such as cooperation in the symmetric

prisoner's dilemma, interpreted by Rabin (1993) and others as positive reciprocity, was more likely an expression of reciprocity-free quasi-maximin preferences.[2]

We studied the determinants of Pareto-damaging behavior by comparing B's propensity to choose (0,0) over (800,200) in different contexts. While 0 out of 36 chose (0,0) when neutral towards A, 10%—6 out of 58—chose it following a decision by A to pose this choice rather than choose an even split. This difference is statistically significant and presumably due to negative reciprocity, though we were surprised how few participants punished others even at little or no cost of doing so.

Participants do, however, quite frequently exhibit a form of reciprocity that we call *concern withdrawal:* They withdraw their willingness to sacrifice to allocate the quasi-maximin share towards somebody who himself is unwilling to sacrifice for the sake of fairness. Consider Game 3. Player A first chooses between payoffs of (550,550) or to allow B the same choice as in Games 1 and 2. Reciprocity predicts that B is more likely to choose (400,400) over (750,375) than in Game 1, since A has been unkind by not choosing (550,550). In fact, about 90% chose (400,400). This suggests that reciprocity is an important component of a player's willingness to sacrifice to help another player.

We close Section 4 by summarizing our findings from the 29 games we studied, and show that—while there is clearly a wide range of motivations among participants in our experiments—our results yield some easy-to-interpret patterns that call into question previous models and provide a foundation for our new model. There seems to be very little positive reciprocity in our data, difference aversion seems to motivate a significant minority of subjects, but to do so weakly, and negative reciprocity is weak but clearly present. Compared to all of these, both quasi-maximin preferences and concern withdrawal are stronger and more common motivations.

In Section 5 we formulate our model of *reciprocal-fairness equilibrium* based on these general results. We assume that each player is motivated by both self interest and a desire to give

---

[2] Positive reciprocity was clearly a determining factor in behavior in one scenario: In the absence of self interest, it overpowers Pareto-damaging difference aversion. When B chooses between (400,400) and (750,400) following the decision by A not to grab a (750,0) allocation, only 6% of participants chose (400,400) compared to 31% choosing (400,400) over (750,400) when neutral. While our research supports the view that positive reciprocity rarely increases willingness to sacrifice, it virtually eliminates Pareto-damaging difference aversion when self interest is not at stake.

each other player his quasi-maximin share. However, she withdraws her desire to allocate this quasi-maximin share to others who are not likewise exhibiting quasi-maximin behavior, and may even sacrifice to punish such players. We show that every *quasi-maximin equilibrium*—a Nash equilibrium with respect to quasi-maximin payoffs—is a reciprocal-fairness equilibrium if each player is as unselfish as the social standard requires. There may also be reciprocal-fairness equilibria involving concern withdrawal or negative reciprocity in which players do not maximize quasi-maximin preferences.

Our model is meant to capture the key aspects of social preferences that we feel previous models have failed to identify, but it is too simple to tightly organize the wealth of data from all laboratory experiments. While our model *is* intended to help improve qualitative and quantitative predictions of social-preferences models in a broad range of experiments, we would be shocked if our specific functional form could serve as a precise explanation of general experimental data, and chagrined if researchers focused too strongly on specific features of our model rather than its usefulness as a building block to improved models.[3] Indeed, our model does not tightly fit our own data. While this is not ideal, we feel that this is not a comment on our model, but rather on our approach: A poor fit inheres in the wide range of games and large numbers of participants we studied, and existing models that fit the data on the range of games upon which they are calibrated do not measure up on simple alternative games that remove confounding factors. We feel that research has not yet reached the stage where we are able to formulate a parsimonious model that closely predicts behavior in a broad range of games.

In Section 6 we discuss various shortcomings with our experiments and model. For example, our results may be misleadingly unsupportive of difference aversion. Some of the differences from earlier research in both our design and in our results—especially the relative lack of retaliatory behavior—demand caution in extrapolating results from our experiments. In addition, we ourselves have gathered some preliminary survey data on hypothetical games that indicate more Pareto-damaging behavior than in games we have played for financial stakes. All said, however, we believe our results are sufficiently intuitive and sufficiently consistent, and the

[3] Note that our skepticism of difference aversion is not targeted at any particular functional form, nor searching far games where additional factors omitted from these models cause the model to fail. Our intuition and empirical tests all address the question of whether the core motivations embedded in these models are providing approximately correct explanations for the experimental phenomena they claim to explain.

confounds in earlier tests that reach different conclusions from us are sufficiently manifest, that we suspect our qualitative results will replicate.

Regardless of our specific findings, we hope this paper helps move experimental research away from testing hypotheses solely on variants of the existing, familiar menu of experimental games. The prisoners' dilemma, public-goods games, and especially the ultimatum game are bad experimental designs for differentiating among social-preferences models.[4] A willingness to move away from such games, and an eagerness to conduct direct, simple, and unconfounded tests of models, will accelerate understanding of social motivations in experimental behavior.[5] In fact, we found more difference aversion than at least one of us expected. Precisely because we design games to isolate difference aversion from confounding explanations, our experiments may be, for skeptics, among the strongest evidence *for* difference aversion. We conclude Section 6 and the paper with a discussion of ideas for further research, emphasizing some specific ways that our model falls short, and proposing some experiments to find further faults with it.

2. Previous Models and Evidence

Models of difference aversion are exemplified by Loewenstein, Bazerman, and Thompson (1989), Bolton and Ockenfels (1999), and Fehr and Schmidt (1999).[6] These models assume that people prefer to minimize differences between their own monetary payoffs and those of other people.[7] Fehr and Schmidt's (1999) model, the most readily applicable variant of difference aversion, says that Player *i* has preferences of the form:

$$U_i(\pi) = \pi_i - \alpha_i[\tfrac{1}{n-1}]\sum_{j\neq i}\max[\pi_j - \pi_i, 0] - \beta_i[\tfrac{1}{n-1}]\sum_{j\neq i}\max[\pi_i - \pi_j, 0]$$

---

[4] Nor do we think they are sufficiently representative of economic situations to justify a decision to concentrate experimental research solely on them, or to care primarily whether models do well at explaining their data.

[5] Using simple games has the additional benefit of discouraging attempts to interpret behavior that seems motivated by social preferences as failed attempts at money-grabbing.

[6] Loewenstein, Bazerman, and Thompson's (1989) evidence and discussion highlight something akin to reciprocity, but omit it from their formal model.

[7] And, it goes without saying, that people are self-interested. Since all models presented assume that a major component of preferences is narrow self interest, for the most part it *will* go without saying.

where there are *n* players, $\pi = (\pi_1,..., \pi_n)$ is the vector of monetary payoffs, and $\alpha_i \geq 0$ and $\beta_i \geq 0$ are parameters of the model measuring how much Player *i* dislikes having less money than others or more money than others, respectively. Fehr and Schmidt (1999) assume that $\beta_i \leq \min[\alpha_i,1]$, which means that getting less than others bothers a person at least as much as getting more, and that she is never so bothered by getting more than others as to want to throw out her own money without benefiting others.

While Fehr and Schmidt (1999) and Bolton and Ockenfels (1999) show that difference aversion can match experimental data in ultimatum games, public-goods games, and some other games, there are considerable experimental data that do not match these models. Andreoni and Miller (1998) test a menu of dictator games that allow one player to decide how to split a fixed number of units with a second player, where the dollar value of the units being divided may differ for the two players. They find that some players grab money and others equalize payoffs. Just as in the classical dictator game, such behavior can be explained qualitatively by difference-aversion models. But Andreoni and Miller (1998) also find that many players sacrifice money to increase total surplus by giving away all or most to the other player, which is the opposite of difference aversion, and interpret these participants who equalize payoffs to be pursuing (what we are calling) maximin preferences rather than difference aversion. They interpret those who give away all their units when these are more valuable to the other player as surplus-maximizers. The following utility function subsumes both of these types of preferences, as well as self interest, as "quasi-maximin" preferences. It represents a sort of reinterpretation of Andreoni and Miller (1998) and previous literature, and is the basis for the model we will employ below.

$$U_i(\pi) \equiv (1-\gamma)\pi_i + \gamma[\delta Min\{\pi_k\} + (1-\delta)\sum_{k=1}^{K}\pi_k],$$

where $\gamma, \delta \in [0,1]$ are parameters measuring the degree of concern for self interest and surplus maximization. Quasi-maximin preferences can also account for sharing in public-goods and prisoners' dilemma games, and better explain dictator sharing games that allow for different exchange rates. Since quasi-maximin preferences assume that people always prefer Pareto-

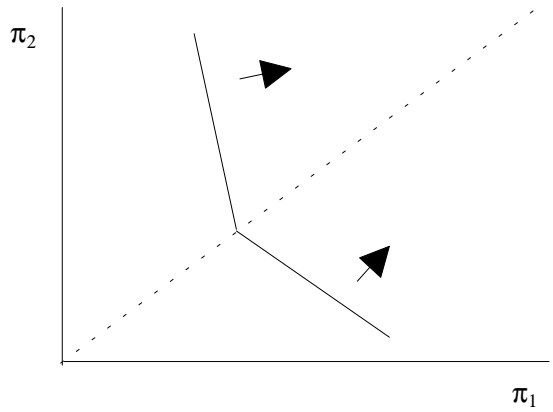
improvements, they cannot explain rejections in the ultimatum game, whereas difference aversion can.

An alternative form of distributional preferences, consistent with the psychology of status, is more rarely discussed: That people always like their payoffs to be as high relative to others' as possible. Such competitive preferences can be represented in simple linear form as follows:

$$U_i(\pi) \equiv (1 - \theta)\pi_i + \theta(\pi_i - m_i) = \pi_i - \theta m_i,$$

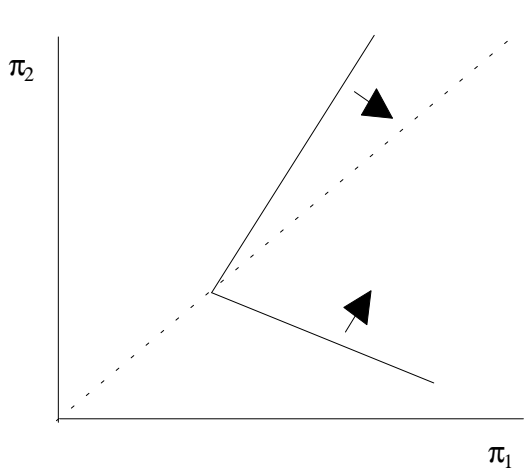
where m_i is the average payoffs for other players besides Player i, and $\theta \in (0,1)$, is a parameter measuring how much they enjoy outdoing others. While we suspect difference aversion is more common, difference aversion is also confounded with competitiveness—some people who decrease others' payoffs when they are getting a lower payoff than the others may also prefer to hurt others even when they are ahead.

We can represent the simple linear forms of the three types of distributional preferences as follows:



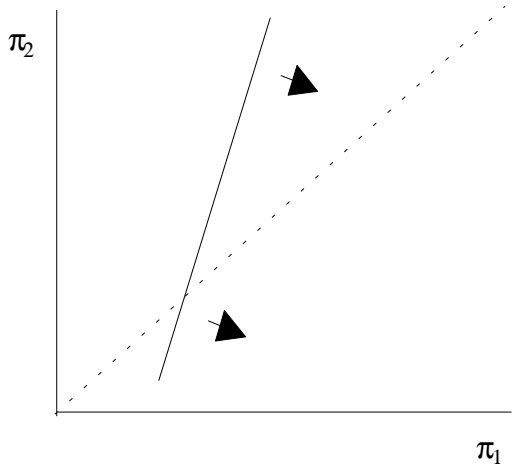
Player 1's Quasi-Maximin

Figure 1.1



Player 1's Difference Aversion

Figure 1.2



Player 1's Competitiveness

Figure 1.3

Figures 1.1-1.3: Player 1's Preferences over (π_1, π_2)

These distributional preferences can also be represented together as one formula, with each of the different preferences embedded in the formula as a special case:

$$U_i(\pi) \equiv \pi_i + \rho(\pi_j - \pi_i) \text{ when } \pi_i \geq \pi_j,$$
$$U_i(\pi) \equiv \pi_i + \sigma(\pi_j - \pi_i) \text{ when } \pi_i \leq \pi_j.$$

In the above formula, competitive preferences corresponds to $\rho < 0$, $\sigma < 0$; difference aversion corresponds to $1 > \rho > 0$, $\sigma < 0$; and quasi-maximin preferences corresponds to $1 > \rho > \sigma > 0$. We shall return to a discussion of this issue in Section 4.

Other studies have shed light on the relative prevalence of these types of distributional preferences.[8] In much the same spirit as the tests we develop in this paper, Kritikos and Bolle

[8] Of course, other conceptions of social preferences exist. For instance, Liebrand (1984), McClintock and Liebrand (1988), and Offerman, Sonnemans, and Schram (1996) deployed a "ring test" of social-value orientations to classify "types" of preferences: people make a series of 24 pairwise choices between alternatives, with the sum of the squares of the payoffs for the chooser and another person held constant. Liebrand (p. 245) describes these categories as:

(1999) conducted a series of simple binary-choice dictator game experiments that shed light on the nature of distributional preferences. They find that in choosing between (Other,Self) payoffs of (4,1) and (0,0), 70 of 80 participants (88%) choose (4,1); when choosing between (10,20) and (40,10), 58% chose (40,10). Even more closely to the examples we develop, they find that 75% chose (5,0) over (0,0), and 74% chose (35,0) over (15,0).[9] Kritikos and Bolle (1999) conclude that inequality aversion is not an important variable compared to a combination of alturism and reciprocity.

Charness and Grosskopf (1999) test variants of dictator games where a person makes a decision that has little or no effect on his own payoff. but substantial effects on a second person's payoff. While about 33% of subjects chose (Other,Self) allocations of pesetas of (600,600) over (900,600), only about 11% of subjects chose (Other,Self) allocations of (400,600) over (600,600). This suggests that about 1/3 of subjects who chose to equalize payoffs when behind are competitive rather than difference averse. In a variant where the chooser receives 600 but can choose any payoff for the other person between 300 and 1200, 74% (80/108) chose 1200, 10% (11/108) chose 600, and 8% (9/108) chose a number less than 600. These experiments, which test distributional preferences when no self interest is at stake, indicate that something like 70% of people are quasi-maximin, 20% difference averse, and 10% competitive.

Other results from Charness and Grosskopf (1999) in which a small amount of money was at stake are perhaps even more telling. They found that 67% (72/108) of subjects chose (Other,Self) payoffs of (1200,600) over (625,625), whereas only 12% (13/108) chose payoffs of (600,600) over (1200,625). That is, of the two thirds of subjects who had quasi-maximin rather than difference-averse or competitive preferences, virtually all were willing to sacrifice 25 pesetas to implement those preferences. Of the one third of subjects who had either difference-

---

"*altruism*: the motivation to maximize other's outcomes; *cooperation*: the motivation to maximize own and other's outcomes; *individualism*: the motivation to maximize own outcomes; and *competition*: the motivation to maximize the difference between own and other's outcomes." Aggregating the results from these studies (which vary across subject pools), about 48% are individualists, 40% are cooperators, 8% are competitors, and 4% are altruists. This scheme for organizing social preferences does not correspond to the array of preferences we've discussed. Besides ignoring reciprocity, the ring-test approach does not lend itself to identifying difference-averse subjects, nor to differentiating between surplus-maximizing and maximin preferences.

[9] All payoffs are (Other, Self) and are in German marks. Some choices were implemented stochastically, as we do in our design.

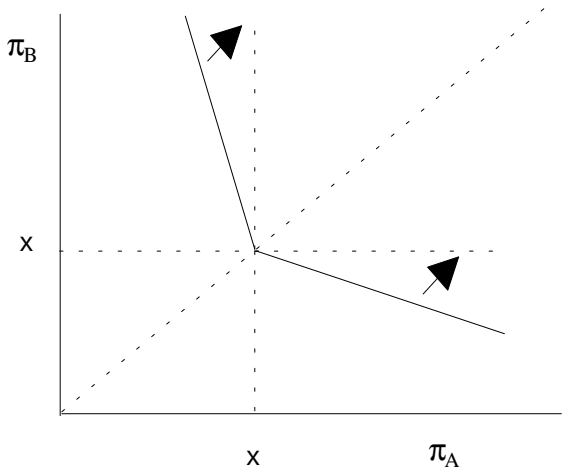
averse or competitive preferences, two thirds were unwilling to sacrifice 25 pesetas to implement those preferences.

Though we emphasize two-player distributional preferences throughout the paper, these previous models and the new one we propose below all relate to multi-person models as well. Of special interest are questions about how players feel about changes in the distribution among others' payoffs given their own payoffs. Andreoni and Miller (1998) do not address this question in the context of quasi-maximin preferences, but Yaari and Bar-Hillel (1984) implicitly do, because their data concern a person's judgment of just division between two other parties. The two major papers on difference aversion, Fehr and Schmidt (1999) and Bolton and Ockenfels (1999), propose different hypotheses. Bolton and Ockenfels (1999) assume that people only care about the average payoff of all other players, and are unconcerned with the distribution of those payoffs. Bolton and Ockenfels (1998, 1999) provide examples where responders in a variant of the ultimatum game studied by Güth and van Damme (1998) seem relatively unconcerned with the distribution of payoffs among other parties. In the simplest form of these games, a proposal was made by one person on a three-way split of a sum of money. A second person could accept or reject this, where a rejection meant that all players get zero, and an acceptance meant they all got the proposed allocation. Responders' propensity to reject the proposed allocation was based only on how much the responders would receive by accepting and was unrelated to how much the third party would receive. This behavior matches the Bolton and Ockenfels (1999) model.

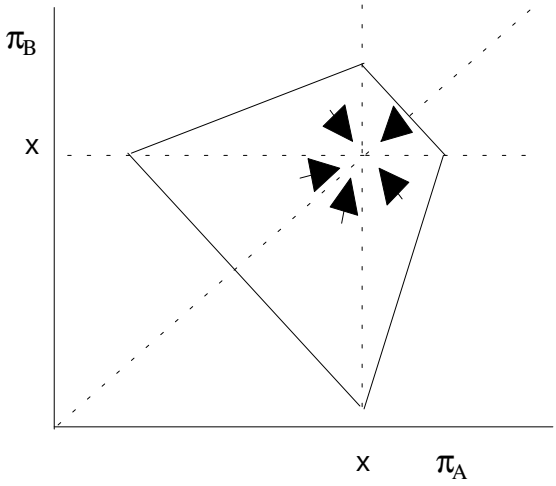
On the other hand, as we shall argue below, we do not believe that most rejections in the ultimatum game are distribution-based rather the reciprocity-based, so an alternative interpretation of at least some (but not all) of their data is that responders were insensitive to the third-party allocation in large part because they were insensitive to allocations *per se*. Only under the maintained hypothesis that rejections are induced by difference aversion rather than retaliation is it clear that we should infer that responders' concern (or lack thereof) for allocations among other parties sheds light on the functional form of difference aversion. Indeed, Kagel and Wolfe (1999) designed a different variant of a three-person ultimatum game and find a form of insensitivity to third-party allocations when all variants of difference aversion predict high sensitivity to these allocations. Their games were similar to those of Güth and van Damme (1998), but involved a "consolation prize" to the third party if a proposal was rejected. Hence,

those who reject an unfair offer might increase inequality rather than decrease it. Hence, the insensitivity to the size of these consolation prizes strongly suggests that insensitivity to a third party sheds light on reciprocity, not distributional preferences. Third-party payoffs don't factor into responders' behavior any more when Bolton and Ockenfels's model predicts they should then when it predicts they shouldn't.

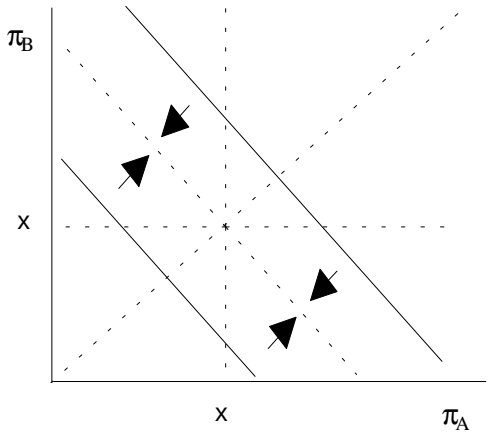
To see how the various distributional preferences differ in their predictions about the distribution of payoffs among two or more other parties, consider Figures 2.1 - 2.4, which represent indifference curves of simplified forms of quasi-maximin, the two types of difference aversion, and competitive preferences over Players A and B's payoffs of a Player C who has a fixed payoff *x*.



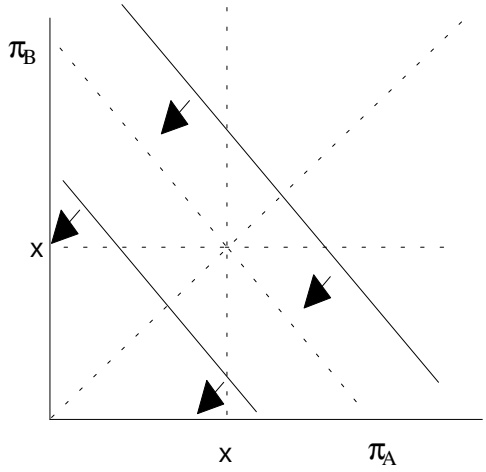
2.1 -- Quasi-Maximin



2.2 -- Fehr and Schmidt



2.3 -- Difference Aversion (Bolton and Ockenfels)



2.4 -- Competitive Preferences

Figures 2.1-2.4: Player C's Preferences Over π_A and π_B Given π_C,

Our contention in the introduction that quasi-maximin preferences may explain experimental data better than does difference aversion requires that we explain rejections in ultimatum games and related behavior—which is consistent with difference aversion and competitiveness, but not with quasi-maximin preferences. One possibility is that the behavior observed in ultimatum games is driven by the minority who have difference-averse or competitive preferences. While we do not claim that the evidence is conclusive on this point, our intuition, our interpretation of previous experiments, and our new experiments make us suspect that this is not right. Rather, we believe that there is a natural alternative explanation for Pareto-damaging behavior: Reciprocity. Several models beginning with Rabin (1993) have assumed that players derive utility from reciprocal behavior, so that they are motivated to sacrifice money either to help those who have been kind or to hurt those who have been unkind. The models in Rabin (1993), Dufwenberg and Kirchsteiger (1998) and Falk and Fischbacher (1998) use the formal framework of psychological games, as developed by Geanakoplos, Pearce, and Stacchetti (1989). Without presenting the formal apparatus of psychological games, these reciprocity models can be represented in vastly over-simplified form as:

$$U_i(\pi) \equiv \pi_i + f_j \cdot f_i ,$$

where f_j is a measure of Player *i*'s beliefs about whether Player *j* is treating Player *i* kindly, and f_i is a measure of how kindly Player *i* is treating Player *j*. A positive value for each of these terms indicates kind behavior, and a negative value indicates mean behavior. This specification assumes a tendency to reciprocate both good and bad intentions of others—if Player *i* believes Player *j* is trying to be kind, she will wish to do so by helping Player *j*. If she believes Player *j* is being selfish or mean, she will lower Player *j*'s payoff.

There are many studies showing reciprocity that cannot be explained by distributional models. Kahneman, Knetsch, and Thaler (1986) instruct people to make the binary choice of splitting $20 (10,10) or (18,2) with an anonymous second party. After a fraction of these choices were randomly implemented, people whose choices were not implemented were arranged in three-person groups, and one person in each group was informed about the earlier choices of the other two people in the group. If these other people had made different choices, the decider was then asked to choose between (Self, Even Chooser, Uneven Chooser) payoffs of (6,0,6) or (5,5,0). 74% of participants made the latter choice, sacrificing $1 to punish an unfair allocator. Moreover, there was a substantial correlation between the choices made in the two stages - 88% of those who had split evenly in the first stage chose to make the $1 sacrifice, while only 31% of those who allocated (18,2) elected to punish. Note that there is clearly no distributional explanation for this phenomenon, since the second-round experiments are among only those whose first-round choices were not actually implemented.

Blount (1995) elicits the minimum acceptable offer in variants of the ultimatum game. She shows that people were more likely to accept a lesser share of a sum of money when they knew the proposed split was generated by a random mechanism or a third party than when generated by the (self-interested) party with whom she would split. In one treatment, the average minimal acceptable offer from a $10.00 pie was $2.91 when made by the self-interested party, $2.08 when made by the third party, and $1.20 when generated at random.[10] Blount presents two other treatments that also indicate reciprocity played a role, but does not report the data in enough detail to fully determine the role of reciprocity..

[10] The $1.20 is significantly different from the other two, but the $2.08 is not significantly different from $2.91.

Offerman (1998) studies the effects of random choice mechanisms while allowing for both positive and negative reciprocity. He considers players' responses to a helpful or hurtful choice, as a function of whether the "choice" was made by an interested party or generated at random. The helpful choice gave $(\pi_1,\pi_2) = (8,14)$, measured in Dutch guilders; the hurtful choice gave them (11,6). Following the choice of either (8,14) or (11,6), Player 2 could either let the choice stand or sacrifice 1 guilder to either increase Player 1's payoff by 4 or decrease it by 4.

Following the helpful choice, Player 2 never paid to lower Player 1's payoff, but paid to help Player 1—changing the payoffs from (8,14) to (12,13)—50% of the time when the choice was random, and 75% of the time when the choice was made by Player 1. This indicates the distributional preference to help Player 1, but also suggests some positive reciprocity, since Player 2 was more likely to help Player 1 when the helpful choice was intentional. The effect on Player 2's response of Player 1's intentions was more dramatic following a hurtful choice. After a randomly-generated hurtful choice, 17% of subjects paid to lower Player 1's payoffs—changing payoffs from (11,6) to (7,5)—58% stood pat, and 25% paid to increase Player 1's payoff. Following an intentional hurtful choice, however, 83% paid to lower Player 1's payoffs, 17% stood pat, and 0% paid to increase Player 1's payoffs.[11] The following chart summarizes Offerman's (1998) results:

	Player 2 decides to:	**Hurt**	**Pat**	**Help**
When Player 1 is helpful:	Intentionally	0%	25%	75%
	Randomly	0%	50%	50%
When Player 1 is hurtful:	Intentionally	83%	17%	0%
	Randomly	17%	58%	25%

Results from Offerman (1998)

Charness (1996a) examines the role of intentions by studying the effect of altering the source of wage generation in a variant of the labor-market experiment developed by Fehr, Kirchler, Weichbold, and Gächter (1998). He finds evidence for both distributive concerns and

[11] The effect of intentions was statistically significant at conventional levels for negative reciprocity, but not so for positive reciprocity.

negative reciprocity. In an ultimatum game, Kagel, Kim, and Moser (1995) vary the exchange rates for payoff chips and the information provided about these exchange rates. They find that ultimatum rejection rates depend on responder beliefs about proposer knowledge of the exchange rates, as knowingly unequal proposals were rejected at substantially higher rates than unintentional unequal proposals. Gibbons and Van Boven (1999) manipulate participants' impressions of the other player's preference in a prisoners' dilemma and observe that rates of cooperation are influenced by these impressions.

Brandts and Charness (1999) test for punishment and reward in a cheap-talk game and find that intention is a critical issue, finding substantial negative reciprocity and significant, but limited, positive reciprocity. One player sends a message about her intended play to another player; after play takes place, the other player is then given an opportunity to punish or reward the first player. They found that this other player was much more likely to punish unfavorable play by the first player if that first player had lied about his play than if he had told the truth. Also, nineteen of 111 subjects (17%) chose to reward a favorable play by the first mover.[12] Andreoni, Brown, and Vesterlund (1999) also show that the difference-aversion models do not explain behavior in their experiments on public goods and best-shot games. They find that "fairness is a function of more than just the final allocations of subjects, but depends on the actions that were not chosen as well as those that are." Similarly, the results in Kagel and Wolfe (1999) discussed above lead Kagel and Wolfe (1999) to conclude reciprocity is at play in ultimatum-game rejections, and to conclude that "both strong and weak versions of Bolton and Ockenfels and Fehr and Schmidt fail to organize the data."

Some other studies yield more equivocal or negative evidence regarding reciprocity. Bolton, Brandts, and Katok (1997) find no evidence of positive reciprocity. Bolton, Brandts, and Ockenfels (1997) find no evidence of positive reciprocity, and only statistically insignificant evidence of small levels of negative reciprocity.[13]

---

[12] B's decision to reward changed the payoffs from (6,9) to (8,7). It is easy to prove that such a decision is inconsistent with the constraints on the parameters that Fehr and Schmidt (1999) impose on their model. The Bolton and Ockenfels (1999) model does not provide a functional form, so we only know that an individual would compare a mild improvement in the equality of the payoff ratio [from (6/15, 9/15) to (8/15, 7/15)] with the loss in payoff of 2. A reward choice is consistent with quasi-maximin preferences, but only if the increase in the minimum payoff from 6 to 7 outweighs the cost of two own payoff units.

[13] The potential negative reciprocity in their experiment was that a responder could retaliate against an unwillingness by a proposer to come out behind. Falk and Fischbacher's (1998) model of reciprocity would predict no reciprocal

Reciprocity clearly matters only in combination with distributional concerns, since a person's generosity or nastiness can only be defined with respect to some norm of fairness. In emphasizing the role of intentions, the simple specification used in the text of Rabin (1993) incorporates an unrealistic notion of what the player might consider the fair division. Most notably, the model assumes that players do not measure fairness with respect to any "external" norms, but rather by splitting the difference among the set of available payoffs. In that formula, Player 1 is considered just as mean for choosing (400,400) over (0,800) as for choosing (800,0) over (400,400)—since in both cases he is grabbing as much as possible for himself.

Such a simplistic split-the-difference functional form cannot explain differences between pairs of games such as the ultimatum and best-shot games. In simplified form, B would be more likely to reject an (800,200) allocation in favor of (0,0) if A has offered this instead of (500,500) than if A has offered this instead of (200,800). The model assumes that the more a first player lowers the payoff to a second player relative to what she could have done, the angrier the second player gets. This is incorrect empirically and implausible psychologically. Punishment is rarer in the best-shot game, presumably because the responder is more likely to forgive a first mover's "unfair" offer if the alternative was unfair for the first nover than if the alternative was to share the pie fifty-fifty. A similar result is found by Brandts and Solà (1998). While B rejected the unfavorable split of (320,80) 33% of the time when the alternative proposal was (100,300), they rejected (320,80) only 16% when the alternative proposal was (50,350)—presumably because the responders are less angered by an unwillingness by the proposer to accept the short-end of division (50,350) that is less fair than what they are proposing than they are when the proposer is unwilling to accept a division that is fairer (100,300) than what they are proposing.[14]

Falk and Fischbacher (1998) combine difference aversion and reciprocity into a model that rectifies this misprediction in an intuitive way. Their model assumes that 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. However, we shall argue below that there is a somewhat different intuition, based on

---

behavior in this game. Our model below, however, can predict some negative reciprocity for some parameter values, since it predicts responders can be angered by an unwillingness by the proposer to come out behind when doing so is, by a legitimate quasi-maximin criterion, the right thing to do.

[14] These results are highly significant if different plays of the same players are considered independent, but their true statistical significance is probably lower given that the data reflected repeated choices of a relatively small number of participants.

quasi-maximin preferences rather than difference aversion, that better explains data in a broad array of situations when negative reciprocity is triggered. It is not that responders forgive proposers for not wanting to come out on the short end; rather it is that they forgive proposers for behavior that is not much (or no) more selfish than what a disinterested person would do. The outcome (200,800) is obviously no more socially attractive than (800,200), so the proposer is under no particular obligation to pursue it. Suppose instead the proposer's alternative to offering the responder a (800,200) versus (0,0) choice were to unilaterally allocate (700,1300). This allocation *would* certainly be much preferred by a disinterested person to (800,200), and hence we strongly suspect that responders may react unfavorably in this case, in contradiction to the Falk and Fischbacher (1998) model.

The main specification discussed in Rabin (1993) is also unrealistic in a more fundamental way: It assumes that when others neither help nor harm them, people are purely self-interested—motivated neither to help nor to hurt these others. As noted by many papers, such as Rabin (1993), this is clearly wrong, and is contradicted by most of the experiments illustrated above. Our model emphasizes social concerns in the absence of reciprocity concerns. Indeed, to develop a simple and tractable model which matches our lack of evidence for positive reciprocity, we will assume that players are no more prone to sacrifice for others when the others have merely not done any harm than when they have been nice.

3. Experimental Procedures

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 29 distinct games and 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.[15] There were 319 participants in the

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

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 A.

We conducted no pilot studies and report all data from experiments played for financial stakes. We also collected survey responses from Barcelona students about how they would behave in hypothetical games. Some of these results appear to suggest greater difference aversion for larger stakes differentials, and hence to contradict our results. We do not report these data in detail, but discuss them briefly in Section 6. We designed the Berkeley games after examining the Barcelona results, and modified a few of the games after observing earlier results.[16]

Students at UPF were recruited by posting notices on campus; most participants were undergraduates majoring in either economics or business. Recruiting at Berkeley was done primarily through the use of campus e-mail lists. Because an e-mail sent to randomly-selected people through the Colleges of Letters, Arts, and Sciences provided most of our participants, the Berkeley sessions included people from a broader range of academic disciplines than is common in economics experiments.[17]

Some effort was made to make different treatments as comparable as possible, reflecting a concern that the selection of people who show up—and the moods they show up with—might (for instance) be different in a Tuesday-morning session than in a Friday-afternoon session. 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. While parallel sessions were impractical in

---

[16] 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 made two substitutions to the games originally planned for the last session: We replaced the games of A choosing (375,1000) or giving B a (350,350) vs. (400,400) choice and A choosing (1000,0) or giving B a (800,200) vs. (0,0) choice with the games of A choosing (750,750) or giving B a (800,200) vs. (0,0) choice and A choosing (450,900) or giving 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.

Berkeley, we hope comparability was enhanced by the fact that all sessions were held on Wednesday or Thursday afternoons and 4 of the 5 sessions were held at the same time of day.

In all games, either one or two participants made decisions, and decisions affected the allocation to either two or three players. In two-player games, money was allocated to players A and B based either solely on a decision by B, or on decisions of both A and B. In three-player games, money was allocated to players A, B, and C, based either solely on a decision by C, or on decisions by both A and C. Participants were divided into two groups seated at opposite sides of a large room and were given instruction and decision sheets. The instructions were read aloud to the group. Prior to decisions being made in each game, the outcome for every combination of choices was publicly described to the players.[18]

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 Katok (1997), Bolton, Brandts, and Ockenfels (1997), and Brandts and Charness (1998), each game was played twice and each participant's role differed across the two plays. Participants were told before their first play that they would be playing in the other role as well, but to discourage reputational motivations, they were assured that pairings were changed in each period.

To maximize the amount of data in response games, responders (B or C) were not told before they made their own decision about the decisions of the first mover (A). The responder instead designated a contingent choice, after being told that his decision only affected the outcome if A opted to give the responder the choice, so that he should consider his choice as if A's decision made it relevant for material payoffs. This *strategy method* plausibly induces different behavior than does a *direct-response method* in which players make decisions solely in response (when necessary) to other players' decisions. Roth (1995, p. 323) notes that "having to submit entire strategies forces participants to think about each information set in a different way than if they could primarily concentrate on those information sets that may arise in the course of

---

[17] 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.

[18] To facilitate presentation, we will present our results with different labels than the ones we actually used, which are reported in Appendix B.

the game." This statement has appeal in complex or unfamiliar environments, but we are unaware of evidence of a significant difference between the two methods in simple games.[19] While at least one of us conjectures that differences in the two methods will emerge as more evidence is gathered, we both suspect that the use of the strategy method is not an important factor in our results.

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

Some aspects of our experimental design may discourage comparing our results to those of other experiments. Our use of role reversal and multiple games in sessions may have generated different behavior than had each participant played just one role in one game.[20] On the other hand, we had each participant make each type of decision only once. Many experiments have players make the same decision repeatedly. While this additional difference from standard procedure might make our results even less comparable, we also suspect that having players play the same role in the same game more than once may have similar effects on behavior as does serial play of different games.

---

[19] Cason and Mui (1998) find that the strategy method does not induce choices that differ significantly from choices made using the direct-response method. Brandts and Charness (1998) gives participants binary choices and finds that the percentage of responders who sacrifice to reciprocate generous behavior is 47% for the strategy method versus 37% for the direct response method in the Prisoner's Dilemma, it is 42% rather than 55% in the Chicken. The difference between the strategy method and the direct-response method in the proportion of subjects who sacrificed was not statistically significant in either game or when pooled together. Güth, Huck and Müller (1999) study mini-ultimatum rejection rates and also test for differences between the strategy method and direct-response method. Some substantial differences were found, although because of the small numbers involved these were not statistically significant.

4. Results

We present our results by classes of games, categorizing our games in two ways: Tables 1.1-1.6 organize the games by their strategic structure and the general nature of the tradeoffs involved, while Tables 2.1-2.6 organize the games by the specific choice B is making. We focus much more on Player B's behavior than Player A's behavior, discussing A behavior only when particularly noteworthy. We return at the end of the section to a more detailed discussion of A behavior, and to the relationship between how people behave in the A role versus the B role. After we present our formal model, in Section 6 we provide some summary statistics on how well the results fit our model compared to other models. In this section, the emphasis is on presenting the tests and results in their full complexity.

We first discuss the behavior of participants in "dictator" games, which reveal reciprocity-free preferences. Table 1.1 shows results from our three three-person dictator games:

Table 1.1:	**Three-Person Dictator Games**	**Left**	**Right**
Barc10 (24)	C chooses (400,400,x) vs. (750,375,x)	.46	.54
Barc12 (22)	C chooses (400,400,x) vs. (1200,0,x)	.82	.18
Berk24 (24)	C chooses (575,575,575) vs. (900,300,600)	.54	.46

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.[21]

While results from other dictator games reflect people's self interest, envy, and other self-involved motivations, Barc10 and Barc12 offer a test of people's "disinterested" views of fairness. Both show that disinterested parties will often choose a more equal outcome over one that maximizes total surplus. The distributional models we know of make no predictions when a person is not comparing others' allocation to her own. But such behavior is consistent with

---

[20] We do not have a good intuition for the size or direction of such a difference.

[21] We made sure that participants did not think that their behavior influenced x. Participants were told that the actual value of "x", to be revealed at the end of the experiment, was written on the back of a piece of paper that was visibly placed on a table and left untouched until the end of the experiment.

disinterested variants of both difference aversion and quasi-maximin preferences; the difference in the proportion choosing (400,400) is statistically significant at $p \approx .01$ in the direction that the forms of such models represented by Figures 3.1 - 3.3 would predict. (Here and throughout the paper, the p-value is approximated to two decimal places and is calculated from the test of the equality of proportions [normal approximation to the binomial distribution; see Glasnapp and Poggio, 1985]. As we generally have a directional hypothesis, the p-value given reflects a one-tailed test. Where there is no directional hypothesis, we use a two-tailed test and state that we do so.) The fact that 46% of disinterested subjects chose (400,400) will prove to be of great interest given our results below, because it is not much smaller than the number of non-reciprocating "interested" participants who choose (400,400) over being on the short end of the (750,375) allocation. But in both games, significant numbers of participants are also inclined to maximize total surplus rather than equalize outcomes.

We designed Berk24 as a direct test of the Bolton and Ockenfels (1998, 1999) hypothesis that players don't care much about the distribution of payoffs among other players. They found that subjects did not appear concerned about such matters in the experiments reported by Güth and van Damme (1998). Here they are concerned: More than 50% of the participants sacrificed 25 to equalize payoffs with each of the other players, without changing the difference (zero) between a player's own payoff and the average of other players. Under the assumption that virtually no participants would (without reciprocal motivations) choose (575,575,575) over (600,600,600), these results support either the Fehr and Schmidt (1999) difference aversion model or quasi-maximin preferences, and reject the Bolton and Ockenfels (1999) model. 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: As we report below, the 50% who sacrifice 25 to equalize payoffs among others is a higher proportion than we found who are inclined to sacrifice *nothing* to eliminate disadvantageous inequality against themselves.[22] Hence, our results suggest that people are more concerned about (this aspect of) the distribution among other players' payoffs than they are concerned about equalizing the self-other payoffs in the sense captured by difference-aversion models.

We continue with two-person dictator results:

Table 1.2:	**Two-Person Dictator Games**	**Left**	**Right**
Barc2 (48)	B chooses (400,400) vs. (750,375)	.52	.48
Berk17 (32)	B chooses (400,400) vs. (750,375)	.50	.50
Berk29 (26)	B chooses (400,400) vs. (750,400)	.31	.69
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

In great contrast to the predictions of difference-aversion models, but consistent with quasi-maximin preferences, about one half of B's sacrifice money to *increase* their deficit with respect to A in Barc2 and Berk17. While a substantial number of people don't like receiving less than another person, in Berk29 and elsewhere we never observe more than 1/3 of people exhibiting *any* degree of difference aversion. Note that our use of exact ties in B's payoff provides the best possible chance of revealing any degree of difference aversion, since it eliminates self interest as a countervailing motive. Furthermore, presumably some of the (400,400) choices reflect competitiveness rather than difference aversion. From the results in Charness and Grosskopf (1999) and elsewhere, we suspect that perhaps 10% of people have competitive preferences, having a taste for lowering others' payoffs irrespective of the implications for inequality. Hence, our best guess from these results is that 20% of people exhibit difference aversion when no sacrifice of money is involved.

Berk23 was an attempt to test the willingness of participants to reject offers of the sort rejected in many ultimatum-game experiments, but in a reciprocity-free context. There is obviously no support for difference aversion in this experiment. As we show below, however, inducing negative reciprocity motives for B making the same choice as here did not lead to very high rejection rates. Hence, Berk23 provides only limited evidence that punishment in the ultimatum games doesn't come from difference aversion.

---

[22] Another useful comparison is the results reported in Section 2 from Charness and Grosskopf (1999), where only 12% (13/108) of subjects were willing to sacrifice 25 to reduce the disparity between their own payoffs and others' payoffs by choosing (600,600) over (1200,625).

The contrast in behavior between Barc8 and Berk15 is intuitive. 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. The 4:1 trade-off in payoffs for choosing to sacrifice is the same in both games, yet B more often makes the choice for higher own payoffs in Barc8, when no even split is available.[23] A higher proportion of B's take a 100% share in Berk26 than in traditional dictator experiments. On the other hand, the standard dictator game offers intermediate choices where the allocator receives most, but not all, of the money. The 22% rate observed for even splits is not unusual in a dictator game. Unsurprisingly, we see by comparing Berk26 to Berk15 that far fewer participants sacrifice 400 to help A by 400 and achieve equality than are willing to sacrifice just 100 to help A by 400 and achieve equality.

We turn to two-person response games. We begin with games where B's choices do not affect her own payoffs:

Table 1.3:	**Two-Person Response Games—B's Payoffs Identical**	**Out**	**Enter**	**Left**	**Right**
Barc7 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	.47	.53	.06	.94
Barc5 (36)	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	.39	.61	.33	.67
Berk28 (32)	A chooses (100,1000) or lets B choose (75,125) vs. (125,125)	.50	.50	.34	.66
Berk32 (26)	A chooses (450,900) or lets B choose (200,400) vs. (400,400)	.85	.15	.35	.65

We designed Barc7 to test the relative strength of positive reciprocity versus difference aversion when self interest is not implicated. In contrast to the 31% of B's who choose (400,400) in the dictator game Berk29, only 6% do so following a generous move by A.[24] The difference between Barc7 and Berk29 is significant at $p \approx .00$.[25] We again wish to emphasize that there is no reason to consider B's choice between (750,400) and (400,400) anything but a strong invitation to B to pursue difference aversion. We show below that positive reciprocity is nowhere else a strong motivation in our data, so that its dominance here over difference aversion

[23] It may be worth noting that there is no combination of α and β in the Fehr and Schmidt (1999) model that can explain the behavior of the 33% who choose (700,500) in Barc8. The choice of (700,500) implies that $500 - 200\alpha > 600 - 300\beta$, which is inconsistent with Fehr and Schmidt's assumption that $\alpha \geq \beta$ and $\beta < 1$.

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

[25] However, only 53% of A's entered, suggesting that either some of them were competitive, or that they did not anticipate such positive behavior by B's.

seems to indicate that difference aversion is weak even among the 1/3 of the population who are motivated by it.

We were surprised by our findings in Barc5, Berk28, and Berk32. In each case, an apparent "mean" action by A was punished by only about 35% of B's. It costs B nothing to punish A. But doing so contradicts quasi-maximin preferences in Barc5 and both quasi-maximin preferences and difference aversion in Berk28 and Berk32. These are indicative of many of our results: For whatever reason, we observed relatively few instances of retaliatory decreases in others' payoffs unless they benefited the retaliators materially.

An interpretation of difference-aversion models, seemingly promoted by Fehr and Schmidt (1999), is that they work well when a person's negative reciprocity is triggered. It is far from clear that the data strongly support even this more limited applicability of the model. Again urging caution about comparing across subject pools, note the contrast between Barc5 and Berk28/Berk32. In Berk28 and Berk32, we observe the same proportion of B's *increasing* inequality to punish A's attempt to *decrease* inequality as we see in Barc5 B's decreasing inequality to punish A's for increasing inequality; we would conclude from the equal propensity to punish in these three games that difference aversion has no explanatory power in predicting retaliatory behavior. Other experiments we report on below, especially game Berk22, strongly contrast with these results, and lend more credence to difference aversion models as a factor in retaliation. Moreover, observing the lack of difference between Berk29 and Barc5, difference aversion explains *all* of what seems to be retaliation.

While these three responder games offer a somewhat confusing picture, Table 2.1, which shows all of the games in which B chooses between (750,400) and (400,400), offers a clearer picture about how reciprocity is implicated in responder behavior:

Table 2.1:	**Games With the Choice Between (400,400) and (750,400)**	**(400,400)**	**(750,400)**
Berk29 (26)	B chooses (400,400) vs. (750,400)	.31	.69
Barc5 (36)	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	.33	.67
Barc7 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	.06	.94

We believe that a majority of laboratory participants are not at all difference averse when they get lower payoffs than others. Rather than wanting to lower others' payoffs, they want to

raise them. And we strongly suspect that difference-averse behavior by the significant minority of people who are difference averse is not robust to positive reciprocity.

Our two three-person response games also offer strong evidence of reciprocity in responder behavior:

Table 1.4:	**Three-Person Response Games**	**Out**	**In**	**Left**	**Right**
Berk16 (15)	A chooses (800,800,800) or lets C choose (100,1200,400) or (1200,200,400)	.93	.07	.80	.20
Berk20 (21)	A chooses (800,800,800) or lets C choose (200,1200,400) or (1200,100,400)	.95	.05	.86	.14

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. In both games, Loewenstein, Thompson, and Bazerman (1989) and Fehr and Schmidt (1999) predict that C will prefer to give the others 1200 and 200 between them, and Bolton and Ockenfels (1999) predict she will prefer 1200 and 100. Reciprocity models predict that the question of *who* gets the 1200 and who gets the low payoff would likely dominate C's choice. Reciprocity clearly explains at least two thirds of the behavior here, since the proportion of C's choosing the 1200/400/100 combination over the 1200/400/200 combination jumped from 14% to 80% when doing so meant A rather than B who would get the low payoff. C's were unhappy with A's greed, and chose to give A the lower payoff irrespective of the distributional consequences, punishing A's 83% of the time overall. This difference in distributional preferences is significant at $p \approx .00$. Because the differences in distributional consequences of behavior were minor, we do not consider this a very discerning test of the general relative strength of distributional vs. reciprocity motivations. Rather, it shows that reciprocity can overwhelm distributional concerns in some circumstances.

We now turn to games in which following an entry decision by A, B has the opportunity to sacrifice to help A:

Table 1.5:	**Two-Person Response Games—B's Sacrifice Helps A**	**Out**	**Enter**	**Left**	**Right**
Barc3 (42)	A chooses (725,0) or lets B choose (400,400) vs. (750,375)	.74	.26	.62	.38
Barc4 (42)	A chooses (800,0) or lets B choose (400,400) vs. (750,375)	.83	.17	.62	.38
Berk21 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,375)	.47	.53	.61	.39
Barc6 (36)	A chooses (750,100) or lets B choose (300,600) vs. (700,500)	.92	.08	.75	.25

Barc9 (36)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	.69	.31	.94	.06
Berk25 (32)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	.62	.38	.81	.19
Berk19 (32)	A chooses (700,200) or lets B choose (200,700) vs. (600,600)	.56	.44	.22	.78
Berk14 (22)	A chooses (800,0) or lets B choose (0,800) vs. (400,400)	.68	.32	.45	.55
Barc1 (44)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.96	.04	.93	.07
Berk13 (22)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.86	.14	.82	.18
Berk18 (32)	A chooses (0,800) or lets B choose (0,800) vs. (400,400)	.00	1.00	.44	.56

The games Barc3, Barc4, and Berk21 all involve a situation where A can either leave B with 0 by choosing "Out", or give B the opportunity to either split evenly an amount that totals approximately what A could have had for himself by choosing Out, or sacrifice a little to give A approximately his foregone payoff. Results in these three games were quite consistent, and consistently surprised us. Rather than observing positive reciprocity, the rate at which B's sacrificed the even split to help A was actually a bit *lower* than in Barc2 and Berk17, the dictator versions of this same decision by B. The lower sacrifice rate between Barc3, Barc4, and Berk21 collectively and Berk2 and Berk17 collectively is significant at $p \approx .14$ using a two-tailed test.

The lack of positive reciprocity is a pattern that also holds for comparing the next game, Barc6, to Barc8, the dictator (300,600) vs. (700,500) choice. For direct comparison:

Table 2.2:	**Games Where B Chooses Between (300,600) and (700,500)**	**(300,600)**	**(700,500)**
Barc8 (36)	B chooses (300,600) vs. (700,500)	.67	.33
Barc6 (36)	A chooses (750,100) or lets B choose (300,600) vs. (700,500)	.75	.25

In Barc9 and Berk25, 8 of 68 B's choose (450,350) over (350,450). We did not run a dictator control for this game, because one of us was confident that virtually no B would choose (450,350) over (350,450). If (say) we ran a dictator session of 32 participants and found one who sacrificed, then the sacrifice rate in Barc9 and Berk25 collectively would be significant at $p \approx .08$; if 0 of 32 sacrificed, the rates would differ at $p \approx .02$.

Hence, Barc29 and Berk25 provide some weak evidence for positive reciprocity. Comparing Berk19 to Berk15, the dictator version of B choosing (600,600) vs. (200,700), the percentage choosing (600,600) on the other hand, does not change significantly, and hence shows no sign of positive reciprocity:

Table 2.3:	**Games Where B Chooses Between (200,700) and (600,600)**	**(200,700)**	**(600,600)**
Berk15 (22)	B chooses (200,700) vs. (600,600)	.27	.73

Berk19 (32)	A chooses (700,200) or lets B choose (200,700) vs. (600,600)	.22	.78

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

Table 2.4:	**Games Where B Chooses Between (0,800) and (400,400)**	**(0,800)**	**(400,400)**
Berk26 (32)	B chooses (0,800) vs. (400,400)	.78	.22
Berk14 (22)	A chooses (800,0) or lets B choose (0,800) vs. (400,400)	.45	.55
Berk18 (32)	A chooses (0,800) or lets B choose (0,800) vs. (400,400)	.44	.56

Once more, these results do not support widespread positive reciprocity. Especially given the size of the stakes involved, the results from Berk14 might seem to show some positive reciprocity when compared to the dictator game Berk26, since 55% choose (400,400) over (0,800), whereas only 22% choose (400,400) in Berk26, significant at $p \approx .01$.[26] But the results from Berk18 call this interpretation into question. Berk18 certainly seems anomalous; we would have 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 $p \approx .01$. The only sense we can make of it—not much—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.

Table 2.5 shows all the games in which B is choosing between (400,400) and (750,375), and provides the starkest presentation of our two main findings about reciprocity:

Table 2.5:	**Games With the Choice Between (400,400) and (750,375)**	**(400,400)**	**(750,375)**
Barc10 (24)	C chooses (400,400,x) vs. (750,375,x)	.46	.54
Barc2 (48)	B chooses (400,400) vs. (750,375)	.52	.48
Berk17 (32)	B chooses (400,400) vs. (750,375)	.50	.50
Barc3 (42)	A chooses (725,0) or lets B choose (400,400) vs. (750,375)	.62	.38
Barc4 (42)	A chooses (800,0) or lets B choose (400,400) vs. (750,375)	.62	.38
Berk21 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,375)	.61	.39
Barc1 (44)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.93	.07
Berk13 (22)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	.82	.18

[26] In the spirit (but not the letter) of the model we develop below, one explanation for why Berk14 might generate more positive reciprocity than in other games is that in all other games besides Berk29 the obligation to sacrifice is ambiguous in the sense that some parameter values for quasi-maximin preferences do not demand sacrifice, whereas in Berk14 not sacrificing here is unambiguously unfair.

These games seen together reflect our general findings about two of the three types of reciprocity that our results help illuminate. The two findings revolved around the fact that a very large percentage of B's here are willing to sacrifice to pursue the quasi-maximin allocation when they feel neutrally towards A's. There is clearly no evidence of positive reciprocity in comparing the first three to the middle three games—B is in fact *less* likely to sacrifice in pursuit of the quasi-maximin outcome following kind behavior by A than in the dictator context (50%). The difference between Barc3, Barc4, and Berk21 collectively and Barc10, Barc2, and Berk17 collectively is significant in a two-tailed test at $p \approx .08$.

But comparing Barc1 and Berk13 to Barc10, Barc2, and Berk17, we see the illustration of the most consistent form of reciprocity that we do find. We call it *concern withdrawal*: B is likely to withdraw his willingness to sacrifice to give the quasi-maximin allocation to A if A has behaved selfishly. Comparing within subject pools, the percentage of B's that sacrifice to help A following a selfish action drops from 48% to 7% (from Barc2 to Barc1) and from 50% to 18% (from Berk17 to Berk13). These are both significant at $p < .01$.

The results above establish the weakness or non-existence of positive reciprocity and the prevalence of concern withdrawal. To investigate "strong" negative reciprocity, where a player sacrifices money to hurt another player, we turn to the final class of response games:

Table 1.6:	**Two-Person Response Games—B's Sacrifice Hurts A**	**Out**	**Enter**	**Left**	**Right**
Barc11 (35)	A chooses (375,1000) or lets B choose (400,400) vs. (350,350)	.54	.46	.89	.11
Berk22 (36)	A chooses (375,1000) or lets B choose (400,400) vs. (250,350)	.39	.61	.97	.03
Berk27 (32)	A chooses (500,500) or lets B choose (800,200) vs. (0,0)	.41	.59	.91	.09
Berk31 (26)	A chooses (750,750) or lets B choose (800,200) vs. (0,0)	.73	.27	.88	.12
Berk30 (26)	A chooses (400,1200) or lets B choose (400,200) vs. (0,0)	.77	.23	.88	.12

The most striking fact about the results in the games in Table 1.6 is that there is relatively little punishment by B. We simply do not find frequent willingness to sacrifice money to punish an unfair player. As we shall show, the level of strong negative reciprocity in our data *is* unquestionably statistically significant. But the fact that the extent of negative reciprocity was so much lower than both our expectations and previous results in the literature worries us most about our data, and invites agnosticism by skeptical readers about the conclusiveness of our data

in establishing the superiority of reciprocity explanations for punishments over distributional explanations.

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. Note that in Barc11, Berk22, and Berk30, the relevant notion of "unfairness" for interpreting A's move is by a quasi-maximin criterion, not difference aversion, as in Falk and Fischbacher (1998). We thought that B would find it inappropriate for A to sacrifice so much social surplus for a little extra money by attempting to get (400,400) rather than (375,1000)—or (worse) by trying to get (400,200) rather than (400,1200).[27]

To see why we suspect that the calibrational success of difference-averse models is an artifact of the confounds in the menu of games on which the models are based, consider first all the games where Pareto-damaging sacrifice is consistent with difference aversion: Berk23, Berk27, Berk31, and Berk30. In these games, 7.5% (9 of 120) of B's choose the Pareto-damaging outcome. Now consider the two games where Pareto-damaging sacrifice is *inconsistent* with difference aversion: Barc11 and Berk22. In these games, 7.0% (5 of 71) of the time.[28]

How does the model by Falk and Fischbacher (1998) combining reciprocity and difference aversion fare? B's are no more likely to punish when doing so reduces the disparity in outcomes as when it leaves inequality unchanged: 9/84 (11%) punish for a decrease in difference, 4/35 (11%) for no change despite getting significantly more bang for their retaliatory buck. But only 1/36 (3%) punish when it increases the difference, lending more credence to the difference-aversion model of retaliation. On the other hand, B's are nearly as likely to punish

---

[27] Our results in Barc11 and Berk22 confuse us. In Berk11, 4 of 35 participants chose (350,350) over (400,400) following an entry decision by A to forego (375,1000), whereas presumably 0/35 would choose (350,350) in the reciprocity-free context. This is clearly "retaliation" without difference aversion. But there is virtually no punishment in Berk22. We do not read too much into these results (the difference is significant at $p \approx .15$, two-tailed test), given the small numbers involved and given that the comparison is across subject pools. But if comparisons like this replicate, it would be evidence that either difference aversion or quasi-maximin preferences temper willingness to retaliate.

[28] While punishment in these two games only costs 50 cents/pesetas compared to the 200 in the comparison groups, the payoff from punishing is much lower too—in Barc11 and Berk22, B can punish A at a 1:1 or 3:1 ratio of harm to cost, rather than the 4:1 ratio involved in the (800,200) vs. (0,0) case.

A's who forego outcomes that are disadvantageous to A as they are to punish A's who forego equal splits.[29]

While it remains to be seen whether difference aversion has significant explanatory power in explaining which punishments are implemented by angered parties, our results clearly reinforce those of Blount (1995) and others that show that reciprocity plays a role in rejections in the ultimatum game. This can be seen by comparing all the games in which B is choosing between (800,200) and (0,0):

Table 2.6:	**Games Where B Chooses Between (800,200) and (0,0)**	**(800,200)**	**(0,0)**
Berk23 (36)	B chooses (800,200) vs. (0,0)	1.00	.00
Berk27 (32)	A chooses (500,500) or lets B choose (800,200) vs. (0,0)	.91	.09
Berk31 (26)	A chooses (750,750) or lets B choose (800,200) vs. (0,0)	.88	.12

0% (0 of 36) chose the (0,0) outcome outside the context of retaliation, while 6/58 chose (0,0) in the two treatments where retaliation is a motive. The difference between Berk23 and each of the other two games is significant separately at $p < .03$.

While we have emphasized B's behavior in reaching our strongest conclusions, obviously A's behavior may also be motivated by social preferences. The strongest—and most tenuous—way to interpret A's choices is to assume that A's correctly anticipated the empirically observed responses by B's and hence that A's made a binary choice between that expected payoff and the payoff from the outside option. Tables 3.1 and 3.2 present these choices between expected payoffs in all the two-player response games. Table 3.1 lists all of the cases where A's sacrifice increases B's payoff.

Table 3.1:	**A's Sacrifice Helps B**				**Maximize**	**Sacrifice**
Barc5 (36)	A chooses	(634,400)	or	(550,550)	.61	.39
Barc7 (36)	A chooses	(750,0)	or	(729,400)	.47	.53
Berk28 (32)	A chooses	(108,125)	or	(100,1000)	.50	.50
Barc3 (42)	A chooses	(725,0)	or	(533,390)	.74	.26
Barc4 (42)	A chooses	(800,0)	or	(533,390)	.83	.17

[29] In contrast to B's behavior, the behavior by A in this last set of games can be seen as quite supportive of either difference aversion or extreme maximin preferences. Over 50% of A's enter in Barc11 and Berk22, where they gain very little and hurt B's by a lot. Far fewer—23%—A's enter in Berk30, where A can't possibly gain from doing so. But we consider 23% is a substantial number, providing support for either difference aversion or competitiveness given that A's must surely have anticipated lowering their expected payoff from doing so.

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

By and large, we would interpret A behavior as being significantly more supportive of difference aversion than B behavior. In Berk28, Barc11, and Berk22, for instance, the best A could hope for by entering is a gain of 25, while costing B's anywhere from 600 to 875. Yet 52% (54/103) of A's enter. This could, of course, reflect extreme maximin preferences and optimistic (and, in each case here, justified) belief that most B's will not punish A's for entering. But it seems more likely that a significant amount of entry in these games reflects either difference aversion or competitive preferences.

Table 3.2 lists all of the cases where A's sacrifice decreases B's payoff.

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

The behavior by A's in our experiments help shed light on the much-emphasized observation that in ultimatum games proposer behavior is not discernibly inconsistent with narrow self interest. This is because proposers have an incentive to make generous offers out of fear of having their offers rejected by responders. It is not clear what would be the generalization of this fact to situations besides the ultimatum game, but the hypothesis that first-mover behavior is likely to be approximately compatible with self interest is (as with many hypotheses) not sustainable when analyzing games besides the ultimatum game. In our data, there appears to be deliberate attempts by A's to sacrifice money. Combining Tables 3.1 and 3.2, we find that 30% of A's take the action that, given actual B behavior, involved an expected sacrifice. While this could, of course, be an artifact of misprediction by A's, note that of A's whose sacrifice helps B,

35% sacrificed, whereas only 15% sacrificed to hurt B's. This difference (179/517 vs. 22/144) is significant at $p \approx .00$. Even more directly, note that in the eight games in which A's decision to enter could only lose her money but could help B, 33% (92/276) sacrificed. In the two cases where entry by A could not help either player, 19% (10/52) entered.[30]

A final facet of our data worth presenting is the relationship of each person's behavior in each game across the two roles. 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 their motivations behind their own choices.

As a clarifying example, consider Barc5, where A can select either (550,550) or let B choose between (400,400) and (750,400). Is the probability that B chooses (400,400) higher if that person chooses (550,550) when in the A role? In fact, 9 of the 14 people who chose (550,550) when A chose (400,400) while B, while only 3 of the 22 people who entered as A's chose (400,400) as B's. The difference between these proportions is significant at $p \approx .00$.

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

Table 4: For each type of behavior as A, did the person help A as B?		**if Out**	**if Enter**	**p-value**
Table 4.1: Helping A Doesn't Affect B's Payoff				
Barc5 (36)	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	5/14	19/22	.00
Barc7 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	15/17	19/19	.12
Berk28 (32)	A chooses (100,1000) or lets B choose (75,125) vs. (125,125)	10/16	11/16	.73
Berk32 (26)	A chooses (450,900) or lets B choose (200,400) vs. (400,400)	16/22	1/4	.06
Table 4.2: Helping A is Costly to B				
Barc3 (42)	A chooses (725,0) or lets B choose (400,400) vs. (750,375)	10/31	6/11	.19
Barc4 (42)	A chooses (800,0) or lets B choose (400,400) vs. (750,375)	11/35	5/7	.05
Berk21 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,375)	3/17	11/19	.01
Barc6 (36)	A chooses (750,100) or lets B choose (300,600) vs. (700,500)	8/33	1/3	.73
Barc9 (36)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	2/25	0/11	.24
Berk25 (32)	A chooses (450,0) or lets B choose (350,450) vs. (450,350)	3/20	3/12	.48
Berk19 (32)	A chooses (700,200) or lets B choose (200,700) vs. (600,600)	13/18	12/14	.36
Berk14 (22)	A chooses (800,0) or lets B choose (0,800) vs. (400,400)	6/15	6/7	.05
Barc1 (44)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	1/42	2/2	.00
Berk13 (22)	A chooses (550,550) or lets B choose (400,400) vs. (750,375)	1/19	3/3	.00
Berk18 (32)	A chooses (0,800) or lets B choose (0,800) vs. (400,400)	0/0	14/32	

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

Table 4.3: Helping A is Beneficial to B

Barc11 (35)	A chooses (375,1000) or lets B choose (350,350) vs. (400,400)	15/19	16/16	.05
Berk22 (36)	A chooses (375,1000) or lets B choose (250,350) vs. (400,400)	13/14	22/22	.20
Berk27 (32)	A chooses (500,500) or lets B choose (0,0) vs. (800,200)	11/13	18/19	.34
Berk31 (26)	A chooses (750,750) or lets B choose (0,0) vs. (800,200)	16/19	7/7	.26
Berk30 (26)	A chooses (400,1200) or lets B choose (0,0) vs. (400,200)	19/20	4/6	.06

Consider the five games in Table 4.3, where A's entry hurt B, and B could sacrifice to hurt A. Of the participants who themselves entered, only 4% (3/70) "punished"; of those who chose Out, 13% (11/85) punished. This is significant at the $p \approx .03$ level. At first blush this would seem to lend significant support to the reciprocity model. But there is also a difference-aversion explanation in games Berk22, Berk27, and Berk31 for why the same subjects who would enter would 'punish'. Berk30, in fact, is more consistent with difference aversion than with retaliation. But Barc11 is certainly more compatible with retaliation than with difference aversion. Overall, of all the participants who entered as A's, 71% (180/252) took an action as a B that helped A; of all the A's that did not enter, only 44% (178/409) took actions as B's to help A. The difference in rates is significant at $p \approx .00$.[31]

In addition to these general patterns, we also obtain some specific insights from the role-reversal data. Role-reversal can be useful, for instance, for disentangling those motivated by difference aversion from those who are competitive. The two people who chose (400,400) over (750,400) in Barc7 when B's chose Out [(750,0)] when A's and so seem more competitive than difference-averse.[32] In Berk32, three of the four who entered as A's rather than choosing (450,900) chose (200,400) over (400,400) as B's, implying competitiveness as the motivation rather than either difference aversion or retaliation. The play in Berk30 is more consistent with difference aversion, but there are few observations of Pareto-damaging behavior by either A's or B's.

We conclude the presentation of our results with some summary statistics that attempt to tie our game-by-game analyses together into some coherent patterns. We begin by comparing the explanatory power of various distributional preferences (competitive, difference-averse, and

[31] Note that Table 4 has 19 comparisons in all. If the behavior were random, we should expect to see half of the two-tailed p-values above .5 and half below .5. Instead, we find that the p-value is below .5 17 times and above it only 2. Randomness is rejected by the binomial test at $p \approx .00$.

quasi-maximin) in our data. In Section 2, we briefly discussed representing the simple linear forms of competitive, difference-averse, and quasi-maximin preferences together in one formula, with the appropriate restrictions for the parameters. We can analyze the data from the two-person games for consistency with the three types of distributional preferences we discussed, as well as for consistency with narrow self interest. Because we are not considering reciprocity motivations, it is most appropriate to make comparisons using only the seven two-player dictator games presented in Table 1.2. On the other hand, as some of the distributational models have been designed to predict behavior in all settings, even where reciprocity might play a role, we also examine the consistency with each of the four distributional preferences of B's behavior in all 27 two-person games. Table 5.1 presents statistics on both sets of games.[33]

	Total # Observations	Competitive	Difference Aversion	Quasi-Maximin	Narrow Self interest
B's behavior in the seven two-person dictator games	232	140 (60%)	175 (70%)	224 (97%)	158 (68%)
B's behavior in all twenty-seven two-person games	903	579 (64%)	685 (74%)	836 (93%)	690 (76%)

Table 5.1 – Consistency of B Choices with Different Distributional Models

Table 5.1 shows how many observations are consistent with any value of ρ and σ permitted by the restrictions for each type of social preferences.[34] For either set of games, the

[32] Of course, if they expected the other player to act as they did and choose (400,400), playing Out may simply be a choice for more money.

[33] Our determination of which choices are consistent with which models, upon which we base the following statistics, is shown in Appendix C. 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 generic games. In the many games in which B's own payoffs are identical, 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.

[34] In calculating consistency, we deemed the choice by B in Games Barc8 and Barc6 of (700,500) over (300,600) as consistent with difference aversion, even though it is probably not consistent with plausible parameter values of difference aversion. (As noted earlier, for instance, it is inconsistent with the combination of α and β parameters that

proportion of observations explained by quasi-maximin preferences is significantly higher ($p \approx .00$) than the proportions explained by the other three types of preferences.[35] 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 social motivations, our data include results from games where these preferences make different predictions. We cannot define a "fair" test of the different distributional preferences because we do not know the most appropriate array of games to study, but it is our sense that, even without invoking the additional explanatory power of reciprocity, quasi-maximin preferences offer a more promising means of predicting responder behavior than does difference aversion.

Interpreting the consistency of A behavior with different preferences is more problematic, since A's perceived distributional consequences of his choice can depend on his beliefs about what B will do. One approach is to make no assumptions about what A believes B will do—and say that A's choice is consistent with a distributional preference if his choice is consistent with that distributional preference given any belief about what B might do. Under this liberal interpretation of consistency, of the 671 choices made by A, all 671 are consistent with difference aversion, 661 are consistent with quasi-maximin preferences, 636 are consistent with narrow self interest, and 579 are consistent with competitiveness. Few choices by A are entirely inconsistent with any of the models, but clearly difference aversion and quasi-maximin do very well, narrow self interest does a little worse, and competitiveness does relatively poorly.

A second approach to inferring A's preferences is to assume that they correctly predict the distribution of B's behavior. To get a rough estimate of the implications of this approach, we can for each of the games assume that A's believed they were making the choice between the payoff from exiting, and the average payoff from entering, as entered in Tables 3.1 and 3.2. By this count, of the 671 choices by A, 649 are consistent with quasi-maximin preferences, 603 are consistent with difference aversion, 488 consistent with competitiveness, and 466 are consistent with narrow self interest. While this seems to indicate the superiority of quasi-maximin

---

Fehr and Schmidt's (1999) model allows.) There were 21 participants who chose (700,500); had we designated them as inconsistent with difference aversion, the entries for difference aversion would have changed to 163 (70%) and 664 (74%).

[35] Since we often have multiple (up to four) observations for each individual, treating each of the observations as independent overstates the statistical significance. However, even if we divide the number of independent observations by four, the differences are still statistically significant at $p < .01$ in both cases.

preferences, we urge caution in thus interpreting the numbers because (opposite to the case of B behavior) there are more observations where intuitively implausible parameter values are needed to reconcile choices with quasi-maximin preferences than with difference aversion.[36] Overall, it is our impression that that the behavior of A's is more consistent with difference aversion than is B behavior in these games.

Table 5.2 tallies up the consistency of all choices in two-player games by adding A's choices to B's choices in the second row of Table 5.1—and measuring consistency using each of the two methods discussed above:[37]

	Total # Observations	Competitive	Difference Aversion	Quasi-Maximin	Narrow Self interest
Consistency of choice, without assumptions about A's beliefs.	1574	1158 (74%)	1356 (86%)	1497 (95%)	1326 (84%)
Consistency of choice, assuming A's correctly predict B behavior.	1574	1067 (68%)	1288 (82%)	1485 (94%)	1156 (73%)

Table 5.2 – Consistency of A and B Choices with Different Distributional Models

Finally, and as a preface to our formal model incorporating reciprocity, we consider the determinants of both kind and unkind behavior by Player B, as a very rough test of different models. In Table 6.1 we consider the frequency with which B takes an opportunity to engage in

[36] Note that the percentage of A's behavior consistent with narrow self interest given responses by B is 466 out of 671 = 69%, which is modestly less than the percentage (76%) of B's that behave consistently with narrow self interest, and virtually identical to the percentage (68%) of B's that behave consistently with narrow self interest in dictator games. Hence, while we ourselves get the impression that A's behave more selfishly than B's in these games, our data also suggest that the common observation that proposer behavior in ultimatum games is more consistent with self interest than proposers in classical dictator games is somewhat misleading. The ultimatum game is not well suited for identifying the motives of proposer behavior, since self-interest and fairness are often hard to distinguish. In our array of games, we can distinguish whether first movers behave more selfishly than responders or dictators, and we find no such manifest pattern.

[37] As the number of participants in each game varied, our percentages could be correspondingly distorted. Thus, we also checked these percentages by assigning an equal weight to each game (and eliminating duplicate games). We

Pareto-damaging behavior—lowering A's payoff when doing so either decreases B's payoff or leaves B's payoff the same—as a function of various factors, and in Table 6.2 we consider the frequency with which B takes an opportunity to sacrifice to help A as a function of a various factors. Both Tables can help us see the determinants of unkind and kind behavior by B—with our usual caveats that there is no principled sense in which the set of games we have is a random sample of possible games, that we are comparing across subject pools, and across games with different degrees of self interest at stake, etc.[38]

Table 6.1 parses the determinants of Pareto damage in several different ways:

Class of Games	Games in that Class	Chances	Taken	Percent
All games allowing Pareto-damage	5, 7, 11, 22, 23, 27, 28, 29, 30, 31, 32	357	59	17%
Punishing decreases inequality	5, 7, 23, 27, 29, 30, 31	228	34	15%
Doesn't decrease inequality	11, 22, 28, 32	129	25	19%
A has helped B	7	36	2	6%
A has had no play	23, 29	62	8	13%
A has hurt B	5, 11, 22, 27, 28, 30, 31, 32	259	49	19%
When A has refused a deficit	11, 22, 28, 30, 32	155	28	18%
Deficit demanded by QMM	30, 32	52	12	23%
QMM allows refusal of deficit	11, 22, 28	103	16	16%
When A's choice to hurt:				
Lowers the maximin payoff	5, 27, 30, 31, 32	156	33	21%
Raises the maximin payoff	11, 22, 28	103	16	16%

Table 6.1: Determinants of Pareto-Damaging Behavior by B

Table 6.1 shows that B's caused Pareto damage in 17% of the opportunities they had to do so. The first category is perhaps the most important for calling into question difference

find that the percentages changed very little—with this approach, the first row reads 73%, 87%, 94%, 84% and the second row reads 67%, 82%, 94%, 73%.

[38] And the same caveat to this caveat—that we feel that our chaotic constellation of games is clearly more of a cross section of conditions than the standard menu of games studied and presented in support of existing theories.

aversion as an explanatory variable in Pareto-damaging behavior. Namely, in our sample, B's are *less* likely to cause Pareto damage when doing so decreases inequality than when it doesn't decrease it. We don't believe this would be the pattern more generally: We suspect people are more likely to engage in Pareto-damaging behavior when it reduces inequality than when it increases inequality. But we also suspect the role for inequality reduction in punishment behavior has been exaggerated, and our results highlight the overwhelming confound—even in previous research that disentangles reciprocity from distributional preferences—between inequality reduction and Pareto-damaging behavior.

The difference shown in Table 6.1 in the percentage of time Pareto-damaging behavior is taken as a function of when A has helped, not affected, or hurt B is indicative of the reciprocity component in reciprocal behavior. When A hurts B, B is more likely to hurt A than otherwise. The difference in 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 $p \approx .02$.[39]

The bottom half of Table 6.1 considers the determinants of negative reciprocity when A has hurt B by entering. Different theories of reciprocity make different predictions about when a player is bothered by another's harmful behavior. The models in Rabin (1993), Dufwenberg and Kirschteiger (1998), and Levine (1998) each say that B will be bothered by any A decision that lowers B's payoff, without any emphasis on whether the harm is justified or not. By contrast, Falk and Fischbacher (1998) deem that B will not be bothered if A is avoiding a deficit. In our model below, we assume that B is not bothered when A's behavior is in pursuit of the quasi-maximin outcome.

We consider our results on these matters weak and inconclusive. We do not find support for the Falk and Fischbacher variant in our data. The percentage who punish when A has avoided a deficit is not significantly lower—18% rather than 19%—when A has avoided a deficit than in all cases where A's behavior has hurt B. As we discussed earlier, we hypothesized that it is only when deficit-avoidance by A's is consistent with quasi-maximin preferences that B's are likely to treat it charitably; in games where A is avoiding a deficit where quasi-maximin

---

[39] The difference in B behavior comparing when A has helped B and when A has no play is significant at $p \approx .12$; the difference in B behavior comparing when A has no play and when A hurts B is significant at $p \approx .13$.

preferences say that she should be willing to accept the deficit, B is likely to resent the harm A's entry has done. As Table 6.1 shows, breaking down the deficit-avoidance into these two cases seems to support this hypothesis, yielding a difference in punishment rates of 23% vs. 16%. While this difference is not significant at conventional levels ($p \approx .13$), it is nevertheless suggestive.

More generally, the role of quasi-maximin preferences in determining when B will punish A is not strongly demonstrated by our data. In the last two lines of Table 6.1, we provide some evidence on this point. Here we compare the five games where A's harm to B also lowered the minimum payoff—in each case, B's payoff following entry is lower than the minimum payoff that either A or B would have gotten had A exited—to the three games where A's entry would raise the minimum payoff if B does not punish A. B punishes A 21% of the time when the minimum payoff is lowered, and only 16% of the time when the minimum payoff is not lowered by entry. This is significant at the $p \approx .13$ level, and lends some further support to the role of quasi-maximin preferences in negative reciprocity.

While our model follows Falk and Fischbacher's in spirit by positing forgiveness by B for some harmful behavior by A, we assume that B will be angered by A's unwillingness to pursue the quasi-maximin outcome, even if it involves A coming out behind. This distinction is manifestly operative in games Berk30 and Berk32, where A's refusal to come out behind involves Pareto-damaging behavior of her own. In hindsight, our games were not ideal for this purpose, since in all five games B might be plausibly be angered by A's decision to enter.[40] But we can observe that the 18% punishment rate is not significantly lower when A has hurt B by refusing a deficit than the 20% for other cases where A has hurt B (not shown on the Table 6.1, A's harm to B didn't involve deficit-avoidance in Games 5, 27, and 31, where B punished 21 out of 104 opportunities), while the violation of quasi-maximin standards increases B's punishment rate from 16% to 21%.[41]

---

[40] Note that the difference between the last two lines is misleading, since if we compared cases according to whether or not B's choice to cause Pareto damage required sacrifice, we'd find that B's punishment rate when punishment is free increases insignificantly (34% to 35%) when moving from Game Berk28 to Berk32, and not very significantly (7% to 12%) when moving (across subject pools) from Games Barc11 and Barc22 to Berk30.

[41] But this support is tenuous, and we feel the summary statistics reported in Table 6.1 are somewhat misleading. Notice, for instance, that while entry in Barc5 lowers the minimum payoff, and hence is inconsistent with maximin preferences, if B does not punish it can raise total payoffs, and hence is consistent with disinterested *quasi*-maximin

All said, we find support for the hypothesis that B's propensity to harm A is determined in part by reciprocity, and that this reciprocity may be based on A's refusal to pursue the quasi-maximin outcome. Though clearly violation of the quasi-maximin criterion is not the main determinant in our data of Pareto-damaging behavior, we nonetheless build this particular aspect of preferences into our stylized model.

Table 6.2 presents the determinants of B's willingness to sacrifice to help A:

Class of Games	Games in that Class	Chances	Taken	Percent
All games where a sacrifice by B helps A	1, 2, 3, 4, 6, 8, 9, 13, 14, 15, 17, 18, 19, 21, 25, 26	546	199	36%
Helping decreases inequality	6, 8, 14, 15, 18, 19, 26	212	99	47%
Helping increases inequality	1, 2, 3, 4, 9, 13, 17, 21, 25	334	100	30%
A helped B	3, 4, 6, 9, 14, 19, 21, 25	278	100	36%
A had no play	2, 8, 15, 17, 26	170	74	44%
A hurt B in violation of QMM	1, 13	66	7	11%

Table 6.2: Determinants of B's Willingness to Sacrifice for A

B sacrifices to help A 36% of the time when he has the opportunity to do so. There is a significant relationship ($p \approx .00$) between helping behavior and whether such helping increases or decreases inequality, consistent with the predictions of both difference aversion and quasi-

---

preferences. If Barc5 were instead grouped with the Barc11, Berk22, and Berk28 as the set of games where A's harmful entry is conceivably consistent with quasi-maximin preferences, and contrasted with Berk27, Berk30, Berk31, and Berk32 as a group of games where A's entry is unambiguously in violation of a disinterested quasi-maximin criterion, then the punishment rates would switch to 16% for the first group and 21% for the second group—leading to the conclusion that B is less likely to punish clear violations of the distinterested quasi-maximin criterion than less clear violations. We are inclined to deem entry by A in Barc5 as a clear violation of quasi-maximin norms for two reasons. First, unlike Barc11, Berk22, and Berk28, where entry by A accords to the quasi-maximin criterion given actual B responses, given actual B responses—which A could reasonably predict—A's entry in Barc5 violates the disinterested quasi-maximin criterion because it in fact lowers total social surplus given the high punishment rate.[41] Second, as demonstrated in games Barc10 and Barc12, and by inference elsewhere, the small increase in total surplus by entry in Barc5—from 1100 to 1150—is unlikely to be deemed a socially acceptable factor in outweighing the lowering of the minimum payoff from 550 to 400. Despite these interpretations, it is clear that our evidence about the determinants of Pareto-damaging retaliation—including any role of quasi-maximin preferences—is only tentative.

maximin preferences. The fact that 30% of inequality-increasing opportunities to sacrifice are taken, however, indicates much stronger support for quasi-maximin preferences than for difference aversion, reflecting the results presented in Table 5.1. The data in Table 6.2 also support the view that positive reciprocity plays little role in helping behavior, and that negative reciprocity does play a role. The table crystallizes the fact that our data show that a nice prior choice by A is *less* likely to yield nice treatment by B than is no choice by A at all—reducing helping behavior from 44% to 36%. By contrast, when A has hurt B, helping behavior reduces to 11%.[42] While involving only two games and 66 observations, this last comparison forms part of the basis for our incorporation of "concern withdrawal" as the primary form of reciprocity in our model of the next section. Hence, we see that violation of quasi-maximin norms plays a stronger role in determining when a person sacrifices to help another player than it plays in determining when a player sacrifices to harm another.

To keep our paper brief, we conclude our analysis of the data here, and move on to develop our model.

5. A Model

In this section we develop a model meant to capture two of the important features of social preferences—quasi-maximin motivations and intentions-based reciprocity—identified above. Many subjects clearly don't have these preferences, and we feel that the model that ultimately comes out of this literature will clearly need to incorporate a more complicated and more heterogeneous conception of social motivations than is embedded in our model. While we believe that variants of our model *will* help make well-calibrated interpretations of experimental

[42] Notice that we are excluding Berk18 from our tally of games where A's entry hurts B, even though the only possible effect of A's decision to forego (0,800) and give B the choice of (400,400) and (0,800) is to lower B's payoff. We exclude it from this category because A's choice to enter still leaves B with the chance to implement A's exit outcome, and, as denoted in Table 6.2, A's choice here is clearly compatible with quasi-maximin preferences. Including Berk18 would raise the proportion of B's sacrificing to help after A has hurt B from 11% to 26%.

evidence, our model here is much more barebones, and omits many realism-increasing factors that will be crucial to tightly fitting a model to experimental data.[43]

Our model captures the assumptions that each player is motivated by both self interest and a desire to give each other player a fair share according to a quasi-maximin criterion, but loses this desire when such a player is pursuing her self interest rather than the quasi-maximin allocation. We also include the possibility that a player may go further in response to unjustified self-interested behavior by another, and sacrifice to punish her. We proceed in steps that reflect three components of our model. We first posit a person's "disinterested social ideal"; we then specify the weight the person puts on this social ideal relative to her self interest; we finally determine the "reciprocity" component of preferences by specifying for which beliefs the person will sacrifice to pursue her social preferences, for which beliefs she will withdraw her willingness to sacrifice to pursue social preferences, and perhaps punish misbehaving players.

We denote by $W(\pi_1,\pi_2,\ldots,\pi_N)$ a disinterested social-welfare function. The quasi-maximin criterion is:

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

where $\delta \in (0,1)$ is a parameter measuring the degree of concern for helping the worst-off person versus maximizing the total social surplus.[44] Setting $\delta = 1$ corresponds to the pure maximin criterion; setting $\delta = 0$ corresponds to total-surplus maximization.[45]

---

[43] One important step for interpreting experimental data is to develop a non-equilibrium solution concept. Our model also does not incorporate any sophisticated notion of sequential rationality, as have some recent reciprocity models, such as Dufwenberg and Kirchsteiger (1998) and Falk and Fischbacher (1998). We do not do so, partly to keep our model simple, and partly because some of the better predictions made by these models are obtained in our model as well without sequential refinements, by assuming that players are motivated to help others even in the absence of sacrifice by others. Moreover, we suspect that much of the intuition in these models—and the evidence invoked in favor of these intuitions—derive from heterogenous and non-equilibrium play in experiments, rather than from a notion of how players should behave at points in a game that really are "off the equilibrium path". If it is unrealistic to assume that the second mover in a sequential prisoner's dilemma will play a strategy of unconditional cooperation no matter what a first mover does, it is probably not because unconditional cooperation is not a best response to certainty that the first mover will cooperate—in which case we would never observe the second mover's behavior following non-cooperation. It is rather probably because in reality there is a positive probability, due either to heterogenous preferences or disequilibrium, that a first mover will defect—and that the second mover will 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.

Second, we designate a weight that players put on self interest versus social interest. Consider Player *i*'s "reciprocity-free" preferences:

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

where $\gamma \in [0,1]$ measures how much Person *i* cares about pursuing the social ideal vs. pursuing his self interest. Combined with the quasi-maximin social preferences, the function V_i translates into:

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

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

To put these preferences in the context of games, let A_i be Player i's pure strategies, S_i be Player i's mixed strategies, and $S_{-i} \equiv \times_{j\neq i} S_j$ be the set of strategies for all players besides Player i. The material payoffs are determined by actions taken, where $\pi_i(a_1,\ldots,a_N)$ represents Player i's payoffs given actions $(a_1,\ldots,a_N)$.[46]

---

[44] It would surely be more realistic 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. Complicating the model thusly is likely to be important in some applications.

[45] For simplicity we assume in our formal model that the players have identical preferences. Clearly, δ and several of the following parameters of the model might be player-specific, and any serious attempt to calibrate our model to experimental data would have to allow for such variation in parameter values.

[46] An important question in modeling distributional preferences is how players treat probabilistic outcomes, generated for instance by mixed strategies. How, for instance, does a person with pure selfless maximin preferences feel about a 50/50 chance of (8,2) and (4,10), as opposed to the payoffs (5,5) for sure. Is the lottery perceived as a social payoff of 6, since the expected payoffs to each player is 6, or is it perceived as payoffs of 3, since the expected value of the lowest payoff to the players is .5(2) + .5(4) = 3? If the first, then a person will prefer the lottery to (5,5), but if the second, she prefers (5,5). Our formal definition assumes it is a payoff of 3—utilities for the players are the expected value of the quasi-maximin payoffs, rather than the quasi-maximin taken over expected material payoffs. We believe that this "expected-distribution" rather than "distribution-of-expectations" approach is implicit in all distributional models.

While our full model will incorporate reciprocity, we first define an equilibrium notion based just on the quasi-maximin preferences formalized by the V_i functions, by defining Nash equilibrium in the game where players' payoffs are transformed into the quasi-maximin payoffs rather than the original material payoffs.

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

Because $\pi_1,\ldots\pi_n$ are continuous in the players' actions, the functions $V_i(\bullet)$ are well-defined and continuous in the players' actions. Hence, a QME always exists.[47]

QME is useful for two reasons. First, in both reciprocity-free environments—where players are unlikely to be motivated by reciprocity—and in "simple-model environments"—where researchers want the most tractable model possible—QME can provide more explanatory power than other distributional models. Second, it turns out to have a special status in our reciprocity model: With an important restriction placed on the parameters of our model, every QME will be an equilibrium in our full reciprocity model.

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

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

---

[47] As with other distributional models, one could readily define a range of solution concepts with respect to quasi-maximin utility functions. Both refinements of Nash equilibrium (such as subgame-perfect Nash equilibrium) and less restrictive concepts (such as rationalizability) can be applied directly to the transformed games.

$$(1-\delta)\cdot(\pi_i + \Sigma_{m\neq i} \max[1-k\rho_m,0]\pi_m) - f\,\Sigma_{m\neq i}\,\rho_m\cdot\pi_m],$$

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

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

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

where π is the profile of material payoffs.[48] The material payoffs are a function of players' actions, and hence strategies; we suppress this fact in our notation.

We let $g_i(s,\rho)$ be some upper hemi-continuous and convex-valued correspondence from (s,ρ) into the set [0,1] such that, for values (s,ρ) where $\{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$ is non-empty, $g_i(s,\rho) \approx \{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$. In the model, $g_i(s,\rho)$ will serve as a measure of how appropriately other players feel that Player *i* is behaving when they determine how to reciprocate. It can be interpreted as the degree to which Player *i* is pursuing the social good (that is, pursuing the disinterested quasi-maximin criterion) by choosing s_i in response to s_{-i}, given that she has disposition ρ_{-i} towards the other players. Except for a technical fix to assure that $g_i(s,\rho)$ is upper

[48] "Typically", $S_i^*(s_{-i},\rho_{-i},g)$ will be a singleton set.

hemi-continuous and convex-valued, this interpretation holds when there *exists* some degree of concern for the social good that, combined with self interest, can explain Player *i*'s choice. But some strategies may not be consistent with any such weighting—as, for instance, when a person chooses a Pareto-inefficient allocation even when the others have no demerits. In such cases, our model does not pin down a particular functional form, and hence in some cases can be unrestrictive.[49] We make this assumption partly for technical convenience and because it doesn't matter much.[50] But we don't restrict $g_i(s,\rho)$ when $\{g \mid s_i \in S_i^*(s_{-i},\rho_{-i},g)\}$ is empty also because we don't feel we know the right psychology for how people interpret seemingly unmotivated Pareto-damaging behavior (do they think the person is being a jerk, accidentally slipped, feels money is bad?) or behavior that seems motivated by different norms of fairness than expected.

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

---

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

[50] This unrestrictiveness would be more problematic if we were to use it to predict non-equilibrium outcomes, or outcomes for heterogeneous preferences.

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

1) $s_i \in$ Argmax $U_i(s,\rho)$, and

2) $\rho_i = \text{Max}[\gamma^* - g_i, 0]$.

A strategy profile is a RFE if every player is maximizing her expected utility given other players' strategies and given some demerit profile that is itself consistent with the profile of strategies.[51] The implications of RFE depend, of course, on the specific parameter values assumed, and hence it is unrestrictive insofar as there are many degrees of freedom in interpreting behavior as consistent with RFE. It is too restrictive to be directly applied to experimental evidence, on the other hand, because it does not allow for other social preferences, heterogeneity in players' preferences, or non-equilibrium play. Nonetheless, to cursorily illustrate the intuition for how the model reflects our interpretation of experimental behavior, we return to the three games with which we began the paper, and analyze which of the outcomes in each of these games can, for plausible parameter values, be the outcome in a pure-strategy RFE.

In Game 1, where B unilaterally chooses between (400,400) and (750,375), both choices are consistent with plausible values of the parameters of δ and γ. Clearly from our data, where 50% of B's choose each, there is heterogeneity. Moreover, from the fact that only 54% of C's choose (750,375), we can surmise that most of those choosing (400,400) are doing so not because they are selfish (have a very low γ), but rather because they believe strongly in maximizing the minimum payoff rather than the total payoff (have a high δ). It is useful for our purposes, however, to concentrate on preferences in which B would choose (750,375) in Game 1, and see how the reciprocity component of RFE influences that decision.

In Game 2, either outcome in which A exits choosing (800,0) is a RFE for any values of the parameters. Because B is not influencing the outcome at all here, our model allows A to

---

[51] 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). Psychological games are where players' utilities depend not just on the material outcomes, but also on players' higher-order beliefs. Incorporating beliefs directly into utility functions allows us to assume that players care about the motivations of others, which depends not just on what a player thinks other players are doing, but what she thinks other players believe are the consequences of their actions. *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.

assign any demerit level to B, and hence, for a sufficiently strong degree of concern withdrawal (high *d* and *k*) or negative reciprocity (high *f*), (800,0) will be a RFE outcome no matter A's γ and δ. Of greater interest is RFE's predictions about "entry" equilibria, in which A gives B a choice. If A is sufficiently unselfish, and the players put very strong weight on maximin, then (400,400) can be a RFE, since A prefers it to (800,0) and B prefers it to (750,375)—and B's choice thusly is forgiven by A. More plausibly, however, (400,400) won't be a RFE even if A is selfless: Either B would deviate to the socially better (750,375), or, if not, A would deviate because of concern withdrawal for B's refusal to do so. The outcome (750,375), on the other hand, is very likely to be a RFE, since B is likely to be so motivated, and even a relatively small degree of selflessness—and any relative weight of maximin and surplus—would make A prefer (750,375) over (800,0) given she will feel no hostility to B.

In Game 3, by contrast, it is likely the exit payoffs (550,550) are the only payoffs consistent with RFE. Although (400,400) is a possible negative-reciprocity equilibrium if players have a strong taste for negative reciprocity and an odd constellation of parameter values, it is more likely that A would deviate to (550,550) over (400,400).[52] Moreover, (750,375) is unlikely to be a RFE, for much the reason we discussed when presenting our results. Even if B would choose (750,375) over (400,400) when feeling positively or neutrally towards A, unless the players put a huge weight on surplus over the maximin criterion—so that B feels that the gain in surplus of 25 in going from (550,550) to (750,375) is justifiable despite the 175 loss in minimum payoff—she is likely to withdraw her concern for A or feel hostile towards A, and thus choose (400,400) over (750,375).

A final example with which we illustrate RFE is the prisoner's dilemma:

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	400,400	0,500
	Defect	500,0	100,100

[52] In order for (400,400) to a RFE, A would have to resent B's choice of (400,400) over (750,375), and have strong enough retaliatory preference, *f*, that she is willing to punish B at a one-for-one cost.

PRISONER'S DILEMMA

The prisoner's dilemma illustrates a couple of issues. First, the cooperative outcome can be a RFE despite our assumption of no positive reciprocity for the simple reason that it is likely to be a QME. Second, the reason the mutual-defection outcome is likely to be a RFE is *also* because it is a QME. In particular, it is likely that players put sufficient weight on increasing the minimum payoff relative to the surplus that A prefers (100,100) over (0,500) and B prefers (100,100) over (500,0) even if they are not particularly selfish. If not, there is a good chance that this outcome would not be a RFE. It might be a concern-withdrawal equilibrium if, say, each player self-servingly holds the other player to a higher standard than herself (i.e., if γ^* is higher than γ.) But it may not be a RFE because each player might forgive the other player for being selfish since she would recognize that the other player is justified given she herself is being selfish.

Beyond specific examples, we note two more general results of interest. First, every game of the form we are studying has a reciprocal-fairness equilibrium:

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

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

Existence clearly enhances the applicability of the solution concept. A second feature also enhances the applicability of the model despite potential complications due to incorporating reciprocity. Above we noted that quasi-maximin equilibria would play a prominent role in our model. Because of the reciprocity component in preferences, which becomes operative when $\rho_k > 0$ for some k, reciprocal-fairness equilibria might not correspond to quasi-maximin equilibria. Outcomes such as non-cooperation in the prisoners' dilemma can be "concern-withdrawal

equilibria". Indeed, if players hold each other to very high standards of selflessness—if γ^* is very high—it may be that such negative outcomes are the only RFE. But if all players' intrinsic desire, γ, to pursue the social good rather than self interest is at least as great as the standard, γ^*, to which people hold each other, then all quasi-maximin equilibria will be reciprocal-fairness equilibria:

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

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

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

6. Summary and Conclusion

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

Our tentative and rough estimate is that, when reciprocity is not an issue, about 70% of people are motivated by quasi-maximin preferences, 20% by difference aversion, and 10% by competitive preferences.[53] We also believe that quasi-maximin preferences are more "robust" than is difference aversion, in the sense that the 70% motivated by quasi-maximin preferences are less likely to forego pursuit of those preferences in response to other goals—such as self-interest and reciprocity—than are the 20% motivated by difference aversion likely to forego their pursuit of difference aversion. We believe that reciprocity in the form of concern withdrawal is likely to be an important facet of behavior. And though the relative lack of Pareto-damaging behavior in our data makes us more tentative on this point, we feel that it is clear that most Pareto-damaging behavior is more likely caused by reciprocity than by difference aversion (or competitive preferences).

On the other hand, we do find non-trivial amounts of difference aversion (or competitive preferences) in some circumstances, indicating that this motivation may influence behavior. Indeed, since running the experiments reported in this paper, we have gathered some survey evidence from students (in Barcelona) that lends more credence to such models. These data are varied and confusing, but one game provides the strongest evidence we have yet observed for difference-aversion: Given a choice between payoffs of (2000,400) and (400,400), 62% of B's stated they would choose (400,400), and when given a choice between (2000,400) and (375,375), 28% stated that they would sacrifice 25 pesetas for equality. These results suggest that either our results in this paper are not robust, or perhaps that difference aversion is more of a factor when disparities in payoffs are larger.[54] In any event, we are not fully confident that difference aversion is so rare that it is unable to explain significant amounts of the data it purports to

---

[53] 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 it makes relative payoff salient. In any event, we see laboratory experiments as only one mode, probably the best starting point, but certainly not the best finishing place, for investigating social preferences.

[54] There is debate within experimental economics on the necessity of linking choices to actual monetary payoffs. We do not believe that the hypotheticality of games *per se* renders results invalid, and do not have an intuition as to why participants would lie about their preferences for (2000,400) vs. (400,400), but still feel that, in the realm of social preferences, hypothetical results results must be viewed with caution.

explain, but we are skeptical, and are quite confident that it would be wise for researchers to avoid anchoring on these models as the natural base model of distributional preferences.

There are many specific ways in which we suspect our model is incomplete or incorrect, and we don't have strong evidence or intuition for some aspects of the model. We recognize that there is substantial heterogeneity among the participants, but make no accommodation for this in the model. Extensions permitting heterogeneous preferences and informational assumptions will raise many other issues not at play in our simple model. For instance, if it is common knowledge that players share a different norm of fairness, there would be two different directions in which to extend the model. We could assume that Person 1 is not angered by Person 2's behavior that helps Player 2 and hurts Player 1 in a way that violates Person 1's preferred norm of fairness, so long as Person 1 is convinced that Person 2 was adhering to a genuine norm of fairness she would hold even if it did not benefit her.[55] Or we could asume that a person is angered whenever others violate his own norm.

We are also not entirely convinced that positive reciprocity is so rare that it is appropriate to exclude it from a model of social preferences, even though re-reading evidence of positive reciprocity in the experimental literature indicates that much of the evidence for positive reciprocity may be misidentified fairness or concern-withdrawal outcomes.[56] We observe scattered positive reciprocity in our games, and feel that it is possible that this motivation is actually strong for a minority of the population or under certain (as yet unidentified) circumstances. For example, one hypothesis is that people are willing to sacrifice to achieve the fair outcome, and willing to sacrifice for the sake of positive reciprocity—but *not* willing to sacrifice more to achieve both. This would in turn suggest that we should see reciprocity in contexts where sacrificing is neither required by fairness, nor manifestly in contradiction to fair treatment of oneself. Some of our data suggest this interpretation, in fact: In games where A

---

[55] That is, people are not angered by the behavior of others so long as those others don't seem to be behaving too self-servingly. Positing the determinants of anger does not fully address the question of how players react behaviorally. If Person 1 is not angered by Person 2's behavior, then it is likely he won't retaliate. But it is less clear whether or not he withdraws willingness to sacrifice for a norm of fairness that Player 2 evidently does not believe in.

[56] This said, there is at least one reason that evidence for positive reciprocity in psychological research and common intuition is not being found in laboratory economics. In most experiments, money is used, which makes fairness norms and social comparison manifest, and does not allow as much ambiguity as to whether or not one is supposed to help another.

could choose between (450,0) and giving B the choice between (450,350) and (350,450), small but non-negligible numbers of B's chose (450,350). This choice is clearly not mandated by any notion of distributional fairness. Hence, if very few B's would do so in the absence of a choice by A, this should probably be attributed to positive reciprocity.

Further, our data seem to indicate a dependence on the behavior of others that does not lend itself to any sort of natural reciprocity interpretation. We see some evidence of a *complicity effect*: The mere fact of another player being involved in a decision seems to change a player's behavior, generally in the direction of making him more selfish. Does a person act more favorably when she knows that the other person has had no opportunity for a decision, so that the full responsibility for a final allocation rests with the decider? There is some evidence which suggests that impulses towards pro-social behavior are diminished when an agent does not feel the full responsibility for an outcome.[57]

We can think of many additional games to run that could help resolve some of the outstanding issues. In the sessions for this paper we did not, for instance, run a simple dictator version of a choice between (375,375) and (750,400), which would have helped identify the bounds on difference aversion. While we cited evidence in Charness and Grosskopf (1999) that suggests few B's would choose (375,375), we do not have direct evidence in our context. Also, one interpretation of our data we have emphasized is the centrality of A's violation of fairness norms in B's concern withdrawal or retaliation. It would be better to have more evidence than we do that B is not—more straightforwardly—apt to punish *any* behavior by A that harms her. Consider, for instance, what B's strategy would be in a game where A chooses between (200,800) or giving B the choice between (500,500) and (0,525). How would B react? Our model predicts pretty clearly that virtually all B's would choose (500,500), as they would if given the same choice in a dictator format. The more straightforward inclination to punish another when that other has caused you harm would predict that B would choose (0,525).

Some other issues concern behavior in games with more than two players. For example, our emphasis throughout has been on a distributional model that pays attention to only the average payoff and the minimum payoff. This is likely to be an unrealistic simplification, as

---

[57] See Charness (1996b) for a discussion of *responsibility alleviation*, and a review of papers with evidence related to the phenomena.

players are likely to care about other features of the distribution; while the lowest payoff may be particularly salient, the second lowest payoff (and so on) may also be germane.

Our model does not predict stronger resentment towards somebody hurting you rather than hurting someone else. Consider a game where A chooses between the outcome (800,100,800) and giving C the choice between (850,0,200) and (0,850,100). Our model predicts just as much propensity by C to punish A if A's choice were between (800,800,100) decision and giving C the same choice between (850,0,200) and (0,850,100), since A's desire to get (850,0,200) is clearly just as egregious in the second case. Yet the victim in this case is B rather than C; entry helps C, but is manifestly unfair. We don't know how C would behave, but suspect that she might not retaliate, and more generally we suspect that some of the determinants of retaliation may be more self-centered than in our model.

We have in our model assumed, as in all models we are familiar with, that a player is more likely to punish others the less it hurts her the punisher to do so. We believe this is correct, but can think of at least one reason why it is not obvious. Compare, say, the game where A chooses between (600,600,400) and giving C a choice between (750,375,375) and (350,350,400) to the game where A chooses between (600,600,400) and giving C the choice between (750,375,375) and (350,350,350). A problem with punishing A, however, is that it also punishes B. "Punishment" in the first game is beneficial to C, and hence C may worry that A and B may interpret this as money-grabbing at their expense, whereas in the second game there is, absent difference aversion, only one natural interpretation—that C is attempting to punish A. If C cares about how she appears to others, then she may punish in the second game but not the first.

There are many other games that would provide useful insights into social preferences. Our view is that the range of games the literature has studied has been much too narrow; we hope to encourage researchers to employ alternative experimental games to test hypotheses. An important reason to study one particular class of games is that they are more economically realistic or relevant. The ultimatum game and the prisoner's dilemma are parsimonious representations of important phenomena of bargaining and public-goods situations, and hence it may be argued that it is most important to develop models that do well in explaining behavior in those contexts. But they are not the only representations of these phenomena. Indeed, we believe again that their adequacy as stand-ins depends on the assumption of narrow self interest;

features of these games that would not matter were people are not narrowly self-interested matter when they are not.[58]

While the range of games typically studied has been too narrow, the particular games studied have typically been too complicated to lend themselves to easy interpretation. One benefit of the sort of simple games we run is that it is easier to discern what subjects believe are the consequences of their actions. But even in our simple games—and inherently in any games with enough strategic structure to make reciprocity motives operative—we could not reach sharp conclusions about the motivations of first movers because we could not be sure how they thought the responders would play. Hence, we feel one avenue for research would be to pay more attention in experimental design to ways to more directly discern participants' beliefs about the intentions or likely behavior of others in their group or session. 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.

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

REFERENCES

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

Andreoni, J., P. Brown, and L. Vesterlund, "What Makes an Allocation Fair? Some Experimental Evidence," 1999, mimeo.

Bar-Hillel, M. and M. Yaari, "Judgments of Distributive Justice," in Psychological Perspectives on Justice: Theory and Applications, Barbara A. Mellers and Jonathan Barron, eds., 1993, 55-84.

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

Bolton, G. and A. Ockenfels (1999), "ERC: A Theory of Equity, Reciprocity, and Competition," forthcoming, American Economic Review.

Bolton, G. and A. Ockenfels (1998), "Strategy and Equity: An ERC-analysis of the Güth-van Damme game," Journal of Mathematical Psychology **42** Jun-Sep: 215-226.

Bolton, G., J. Brandts, and E. Katok, "How Strategy Sensitive are Contributions? A Test of Six Hypotheses in a Two-Person Dilemma Game," 1996 (forthcoming, Economic Theory)

Bolton, G., J. Brandts, and A. Ockenfels, "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game," mimeo, 1998 (forthcoming, Experimental Economics).

Brandts, J. and G. Charness, "Hot vs. Cold: Sequential Responses and Preference Stability in Simple Experimental Games," 1998 (forthcoming, Experimental Economics).

Brandts, J. and G. Charness, "Retribution in a Cheap-Talk Game," 1999, mimeo.

Brandts, J. and C. Solà, "Reference Points and Negative Reciprocity in Simple Sequential Games," 1998, mimeo.

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

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

Charness, G., "Responsibility and Effort in an Experimental Labor Market," 1996b, (forthcoming, Journal of Economic Behavior and Organization).

Charness, G. and B. Grosskopf, "Relative Payoffs and Happiness: An Experimental Study," 1999, mimeo.

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

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

Fehr, E., E. Kirchler, A. Weichbold, and S. Gächter, "When Social Forces Overpower Competition: Gift Exchange in Experimental Labor Markets," Journal of Labor Economics, 1998, **16**, 324-351

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

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989), "Psychological Games," Games and Economic Behavior, 1989, **1**, 60-79.

Gibbons, R. and L. Van Boven, "Contingent Social Utility in the Prisoners' Dilemma," 1999, mimeo.

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

Güth, W., S. Huck, and W. Müller, "The Relevance of Equal Splits: On a Behavioral Discontinuity in Ultimatum Games," 1999, mimeo.

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

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

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

Levine, D., "Modeling Altruism and Spitefulness in Experiments," Review of Economic Dynamics, 1998, **1**, 593-622.

Liebrand, W., "The Effect of Social Motives, Communication, and Group Size on Behaviour in an N-person Multi Stage Mixed Motive Game," European Journal of Social Psychology, 1984, **14**, 239-264.

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

McClintock, C. and W. Liebrand, "Role of Interdependency Structure, Individual Value Orientation, and Another's Strategy in Social Decision Making: A Transformational Analysis," Journal of Personality and Social Psychology, 1988, **55**, 396-409.

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

Offerman, T., J. Sonnemans, and A. Schram, "Value Orientations, Expectations and Voluntary Contributions in Public Goods," Economic Journal, 1996, **106**, 817-845.

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

Rabin, M., "Bargaining Structure, Fairness, and Efficiency," mimeo, U.C. Berkeley, January 1997.

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

Selten, R., "Die Strategiemethode zur Erforschung des Eingeschränkt Rationalen Verhaltens im Rahmen eines Oligopolexperiments," in Beiträge zur Experimentellen Wirtschaftsforschung, H. Sauermann, ed., 1967, 136-168.

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

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

APPENDIX A - SAMPLE INSTRUCTIONS

INSTRUCTIONS

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

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

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

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

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

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

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

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

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

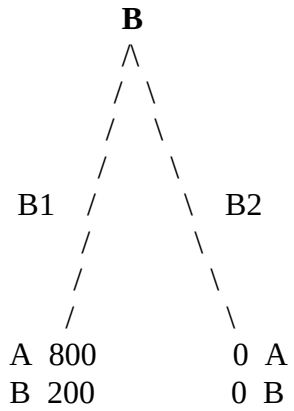
We will proceed to the decisions once the instructions are clear. Are there any questions?

PERIOD 1

GAME 3

In this period, you are **person A**.

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



DECISION

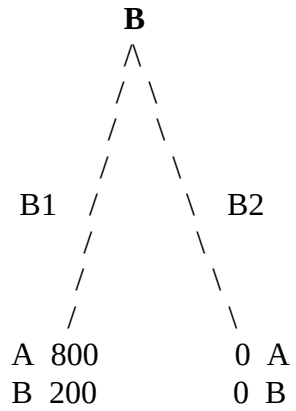
I understand I have no choice in this game _____

PERIOD 1

GAME 3

In this period, you are **person B**.

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



DECISION

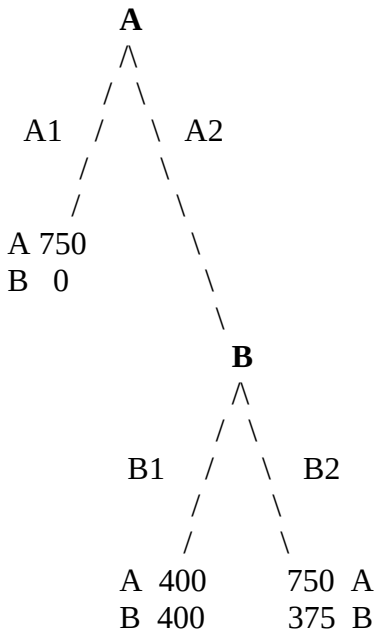
I choose: **B1** **B2**

PERIOD 1

GAME 1

In this period, you are **person A**.

You may choose A1 or A2. If you choose A1, you would receive 750 and player B would receive 0. If you choose A2, then player Bís choice of B1 or B2 would determine the outcome. If you choose A2 and player B chooses B1, you would each receive 400. If you choose A2 and player B chooses B2, you would receive 750 and he or she would receive 375. Player B will make a choice without being informed of your decision. **Player B knows that his or her choice only affects the outcome if you choose A2, so that he or she will choose B1 or B2 on the assumption that you have chosen A2 over A1.**



DECISION

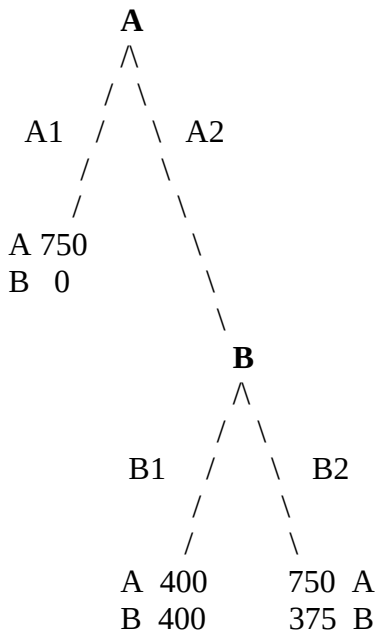
I choose: **A1** **A2**

PERIOD 1

GAME 1

In this period, you are **person B**.

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



DECISION

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

APPENDIX B: Results from All Games

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

Two-Person Dictator Games		**Left**	**Right**
Barc2 (48)	B chooses (400,400) vs. (750,375)	.52	.48
Berk17 (32)	B chooses (400,400) vs. (750,375)	.50	.50
Berk29 (26)	B chooses (400,400) vs. (750,400)	.31	.69
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

Two-Person Response Games - B's Payoffs Identical		**End**	**Enter**	**Left**	**Right**
Barc5 (46)	A chooses (550,550) or lets B choose (400,400) vs. (750,400)	.39	.61	.33	.67
Barc7 (36)	A chooses (750,0) or lets B choose (400,400) vs. (750,400)	.47	.53	.06	.94
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

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

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

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

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

In this Table, we allow A to have any beliefs about Bís response to Enter.

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

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

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

Game	"A" Exit		"A" Enter		"B" plays Left		"B" plays Right	
	N	*Prefs.*	*N*	*Prefs.*	*N*	*Prefs.*	*N*	*Prefs.*
1 A(550,550); B(400,400)-(750,375)	42	C,D,Q,$	2	C	41	C,D,Q,$	3	Q
2 B(400,400)-(750,375)	-		-		25	C,D,Q,$	23	Q
3 A(725,0); B(400,400)-(750,375)	31	C,D,Q,$	11	D,Q	26	C,D,Q,$	16	Q
4 A(800,0); B(400,400)-(750,375)	35	C,D,Q,$	7	D,Q	26	C,D,Q,$	16	Q
5 A(550,550); B(400,400)-(750,400)	18	D,Q	28	C,D,Q,$	15	C,D,$	31	Q,$
6 A(750,100); B(300,600)-(700,500)	33	C,D,Q,$	3	D,Q	27	C,D,Q,$	9	D,Q
7 A(750,0); B(400,400)-(750,400)	17	C,D,Q,$	19	D,Q	2	C,D,$	34	Q,$
8 A(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 = 488	D = 603	Q = 649	$ = 466
Total B choices = 903	C = 579	D = 685	Q = 836	$ = 690