Modeling Altruism and Spitefulness in Experiments[1]

David K. Levine[2]

October 26, 1995

revised October 17, 1997

Abstract: We examine a simple theory of altruism in which players payoffs are linear in their own monetary income and their opponent's. The weight on opponent's income is private information and varies in the population, depending, moreover, on what the opponent's coefficient is believed to be. Using results of ultimatum experiments and the final round of a centipede experiment, we are able to pin down relatively accurately what the distribution of altruism (and spite) in the population is. This distribution is then used with a reasonable degree of success to explain the results of the earlier rounds of centipede and the results of some public goods contribution games. In addition we show that in a market game where the theory of selfish players does quite well, the theory of altruism makes exactly the same predictions as the theory of selfish players.

[1] This work was supported in part by the UCLA Academic Senate and by NSF Grant SBR-93-20695. Discussions with Drew Fudenberg, Tom Palfrey, Robert Rosenthal, John Van Huyck and comments by participants at the University of Chicago Theory Workshop, the UCLA Theory Workshop, the Harvard/MIT joint theory workshop and the Hong Kong University of Science and Technology Theory Workshop are also gratefully acknowledged. I would also like to thank Ramon Marimon and two anonymous referees for their guidance.

[2] Department of Economics, UCLA.

1. Introduction

Standard theory applied to the study of experiments generally examines a refinement of Nash equilibrium such as subgame perfection, and assumes that participants are selfish in the sense that they care only about their own monetary income.[3] Some (but not all) experiments cannot be explained by this theory. Two robust sets of experiments of this sort are those on ultimatum bargaining and on public goods contribution games.

In ultimatum bargaining, the first player proposes a division of a fixed amount of money that may be accepted or rejected by the second player. According to the theory, any demand that leaves the second player with anything should be accepted, and consequently the proposer should either demand the entire amount or at least the greatest amount less than the entire amount. In fact proposers do not demand nearly this amount, generally demanding between 50-60% of the total, and ungenerous demands that are significantly less than the entire amount are frequently rejected.

In public goods contribution games, players may make a costly donation to a common pool that provides a social benefit greater than the contribution. Because of the free rider problem, it is typically a dominant strategy not to contribute anything. Never the less, with as many as 10 or more players, some players contribute to the common pool.

One explanation of these phenomena is that the equilibrium concept is wrong, and this has been explored by a variety of authors such as Binmore and Samuelson [1995]. However, such explanations are difficult to accept since in public goods games dominant strategies are involved, while in ultimatum bargaining it is puzzling that many demands leaving the second player with 30% or more of the total amount are rejected. The

[3] There is much informal discussion of fairness, but little in the way of formal modeling applied to experiments.

rejection of such an offer is not a failure of equilibrium theory, but a failure of the assumption of selfish players.

An alternative line of explanation is that players are not really selfish. One such explanation that is frequently discussed, particularly by the experimenters themselves, is that some notion of fairness plays a role in individual decision making. Rabin [1992] has proposed a formal model of what this might mean. The model presented here is similar in spirit to Rabin's model, but does not depart from the ordinary assumptions of game theory. Moreover, Rabin examined only qualitative predictions of his model. The goal of this paper is to examine the quantitative implications of the theory: to what extent can a *simple* model of players who are not selfish can explain the data from a variety of experiments, including both ultimatum and public goods games. A similar effort can be found in Andreoni and Miller [1996]. That effort differs from this one in focusing only on public goods contribution games, and on altruism, but not spitefulness. In addition, they allow players to have non-linear preferences over contributions.[4] They find, as do we, a remarkable degree of consistency in the attitude of players towards one another in different games.

The basic theory is that players care not only about their own monetary payoffs, but also about their opponent's monetary payoffs. The simplest such model is of the type described by many authors, (see Ledyard [1995] for example) in which utility is a linear function of both the player's own monetary payoff, and his opponent's. It is convenient to normalize the coefficient on the own monetary payoff to one. The question is then, what is the coefficient on the other player's payoffs? Public goods contributions games suggest that this coefficient should be positive; ultimatum bargaining suggest that it

[4] They assume that utility is defined over contributions in the particular game at hand, and not over total wealth. This, combined with non-linearity, can lead to some of the same paradoxes that occur when non-Von Neumann Morgenstern preferences are considered in decision theory. For example, it is possible to construct a series of problems involving contributions in such a way that the optimal solutions to the individual problems are sub-optimal in the joint problem of solving the problems simultaneously.

should be negative (so that offers will be rejected). We adopt the point of view that the coefficient differs between different individuals in the population, with some individuals having positive coefficients and some negative coefficients, and with each individual's coefficient being private information. The theory gains strength from the hypothesis that this distribution is fixed across games of different types, so that the same distribution of coefficients should be used to explain, for example, both ultimatum bargaining and public goods contribution games.

In fact this simple model is inadequate to explain even the results of ultimatum bargaining. From the rejection rates in the experiment we can calculate how many players moving second must be spiteful and how spiteful they must be. However, the players moving first must be drawn from the same population, so we can use the distribution of spitefulness calculated from the rejection rates to compute what demands should be made by the players moving first. In fact the demands that should be made, according to the theory, are substantially larger than those observed in the data.

As a result of this, we are led to a more complicated model of altruism. Introspection suggests that players care not only about other players' utility, but also that their attitudes towards other players depend on how they feel they are being treated. One way to model this is to use the psychological game approach of Rabin [1983] and Geanakoplos, Pearce and Stachetti [1989]. However, these models are complicated, and depart quite radically from ordinary choice theory. We will discuss the Rabin model in greater detail below. As an alternative we propose a simpler model with the same flavor: player's weights on opponents monetary payoffs depends both on their own coefficient of altruism (or spite), and on what they believe their opponents coefficient to be. In particular, a more positive weight is placed on the money received by an opponent who is believed to be more altruistic, and a more negative weight on one that is believed to be more spiteful. Notice that the game played is now a signaling game, since players actions will potentially reveal how altruistic they are, and their opponents care about this.

It is useful to think what consequence this has in ultimatum bargaining. First, larger demands are a signal of greater spite. Hence the degree of altruism needed to explain the rejections is less than it would be in the absence of the signaling effect: even a relatively altruistic player may behave meanly to a player believed to be spiteful. In addition to demands being lower because players are less spiteful, demands are lower because an altruistic player will realize that acceptance of a large demand is a signal of an altruistic opponent, and consequently will be less desirous of making such a demand.

Using this model, we examine ultimatum bargaining experimental results due to Roth et al [1991]. We are able to fit the data quite well using the model, and pin down most of the distribution of altruism, except that the data contains little information about how altruistic the altruistic players are.

The departure of preferences from selfish preferences is substantial, so we next examine whether the model is consistent with other experiments in which the selfish theory together with subgame perfection has worked well. One example of such an experiment is the market experiment also reported in Roth et al. Here the selfish theory predicts the competitive equilibrium, and this is in fact what is observed. However, the structure of the game is such that there is little opportunity for transferring utility to or from other players. As a result we show that regardless of how altruism is distributed in the population, there exist equilibria in which the coefficient of altruism does not matter, and that consequently these equilibria are the same as the equilibria of the selfish model.

We then turn to another well-known experiment inconsistent with selfishness and subgame perfection, the variation on grab-a-dollar studied by McKelvey and Palfrey [1992]. In this experiment a player may either grab or pass. If he passes the money is doubled and his opponent gets the move, except in the final round. The grabber gets 80% of the money, his opponent 20%. This is structured so that with selfish players the unique Nash equilibrium is to grab immediately. In fact only 8% of the population does so. However, as in ultimatum bargaining, there is also a simple failure of decision theory. A substantial fraction of players choose to give money away in the final round. So many do so that it is optimal for a selfish player to stay in as long as possible in hopes of getting lucky and meeting an altruistic opponent in the final round. The distribution of altruism from ultimatum is applied to study this game. The play in the final round pins down the extent to which the altruistic part of the population is altruistic, a parameter that could not be identified from the ultimatum experiment. This gives a complete distribution of altruism, and we find that play in the earlier rounds is generally consistent with this distribution. This is a genuine test of the theory in the sense that there are no free parameters available to explain play in the early rounds.

Finally, we turn to a public goods experiment by Isaac and Walker [1988]. Here the model we use imperfectly represents the way in which the experiment was conducted, and the results of the experiments are not perfectly consistent with the distribution of altruism calculated in the other experiments. Never the less the amount of altruism found in the public goods experiments is reasonably consistent with the amount and degree of altruism calculated from the other experiments.

In our concluding section we discuss in more detail the extent to which the simple theory succeeds in explaining experiments, and speculate about how it might apply in other experiments. One important issue we do not address is the issue of why players should be altruistic or spiteful. It is natural to speculate about evolutionary explanations for preferences of this type, and perhaps future research will identify evolutionary forces that lead to the types of preferences modeled here.

2. Altruistic Preferences

We will be considering n person games with players $i = 1, \ldots, n$. At terminal nodes of the extensive form, player i receives a *direct utility* of u_i. Player i also has a coefficient of altruism $-1 < a_i < 1$ and receives an *adjusted utility* of

$$v_i = u_i + \sum_{j \neq i} \frac{a_i + \lambda a_j}{1+\lambda} u_j ,$$

where $0 \leq \lambda \leq 1$. The objective of players is to maximize their adjusted utility. The adjusted utility reflects the player's own utility, and their regard for their opponents. If $a_i > 0$ we refer to the player as *altruistic*, as such a player has a positive regard for his opponents. If $a_i = 0$ we refer to the player as *selfish*, the usual case. If $a_i < 0$ we refer to the player as *spiteful*. The assumption that $-1 < a_i < 1$ means that no player has a higher regard for his opponents (positive or negative) than for himself.

The coefficient λ reflects the fact that players may have a higher regard for altruistic opponents than spiteful ones. When $\lambda = 0$ the model is one of pure altruism of the type discussed by Ledyard [1995] as an explanation of the results of public goods contribution games. When $\lambda > 0$ the model can be regarded as incorporating an element of fairness, not in the sense that players have a particular target they consider "fair," but in the sense that they are willing to be more altruistic to an opponent who is more altruistic towards them. One of our major conclusions is that $\lambda = 0$ is not consistent with data from the ultimatum game.

Obviously the coefficient a_i is not independent of the units in which utility is measured, and utility must be measured in the "same" interpersonally comparable units for all players. The linearity of payoffs in opponents' utility may be taken as a convenient approximation. It has the important implication that players respect each other's preferences over outcomes and gambles. In general players regard for one another may depend on who the opponent is, but in the types of experiments we will be considering, players interact with each other anonymously, so regarding all opponents in a symmetric manner seems not only sensible, but necessary.

Prior to the start of play, players are drawn independently from a population with a distribution of altruism coefficients represented by a common cumulative distribution function $F(a_i)$. Each player i's altruism coefficient a_i is privately known, while the distribution F is common knowledge. Consequently, we model a particular game as a Bayesian game, augmented by the private information about types. It is of some importance in our analysis that players may reveal information about their altruism coefficient through their play. This can both act as a signal of how they are likely to play in the future, and may change opponents' attitudes towards them (when $\lambda > 0$).

In studying experiments, we will identify the participants' utility with their monetary income from the experiment. Since the sums of money involved are trivial, it is hard to believe that curvature in the utility function can play much of a role in explaining behavior in the experiments. It is important to note, however, that the money that is not received by the participants reverts to the experimenter, and there is no reason for the subjects to feel differently about the experimenter than the other subjects. However, it does not seem sensible to identify the utility of the experimenter with the amount of money that reverts to him. Instead we will assume that the marginal utility of the experimenter for money that is not disbursed to the subjects is zero, so that in effect, from the subjects point of view, the money is thrown away, and the altruism coefficient a_i does not matter. Notice that it is possible to design experiments to control more carefully for the effect of money that is not received by the subjects. Rather than having the money revert to the experimenter, one subject can be chosen to be the residual claimant with all money not disbursed to the subjects being given to the residual claimant, who does not otherwise participate in the experiment. In this case, the utility of money not going to the participants other than the residual claimant can be identified also with money income, and the residual claimant should be viewed by the other subjects as having the population mean value of a_i. According to the theory, this should have an effect on the outcome of the experiment.

Our basic notion of equilibrium is that of sequential equilibrium: each player optimizes given preferences and beliefs that are derived from a limit of strictly mixed perturbations from equilibrium play and Bayes law. As a technical aside, note that all the games considered here are relatively simple, and sequentiality in these games coincides with the simpler notion of a perfect Bayes Nash equilibrium. As all the distributions over types *F* we will consider have finite support, for any given monetary payoffs in a game, our theory of preferences induces an ordinary Bayesian game. This game can be analyzed by the ordinary tools of game theory: there are no new theorems or results about games of this type. The general theorems about sequential equilibria from Kreps and Wilson [1982b] apply directly. In particular, introducing altruistic preferences need not eliminate the multiplicity of equilibria. However, we should emphasize that the number of Nash equilibria is reduced by means of sensible refinements. The use of refinements has fallen into disrepute in the study of experiments, because these refinements (including sequentiality and subgame perfection) do quite poorly in describing actual play. However, that is not the case in this theory: once preferences for altruism and spite are taken into account refinements do relatively well. Indeed, all the equilibria we explicitly discuss are not only sequential, but satisfy the obvious monotonicity requirement on beliefs in a signaling game: beliefs are that the type most likely to deviate is the type for whom it is least disadvantageous.

We will explore the theory by means of quantitative examples drawn from the experimental literature. It is well known that there is considerable learning taking place in the early rounds of experiments. Since our model is a theory of equilibrium, we focus on experiments in which players get several opportunities to play, and focus on the outcome in the final rounds after the players have had time learn an equilibrium. Our goal is to explain why, even after the system appears to have stabilized, play does not resemble an equilibrium with traditional preferences. Our ideal experimental design is one in which players are matched with different opponents every period, so that we may legitimately ignore repeated game and reputational effects between rounds. With the exception of the public goods experiments of Isaac and Walker [1988], all the experiments reported here follow that design. The results of Isaac and Walker [1988] are included because, despite the possibility that repeated effects might have mattered, the experimental design was well suited in other respects for examining the theory described here.

3. Other Models of Altruism and Fairness

Before studying specific experimental data using the model of altruism outlined in the previous section, it is useful to put the model in a broader perspective. The model described can be viewed as a particular parameterization of a class of models in which preferences depend on payoffs to an individual player and to his rivals, as well as depending on his own type and the type of his rivals. As we indicated, because of the stakes involved in the experimental setting, we have chosen a parameterization that is linear in monetary payoffs. This specification would obviously be unsuitable in a setting where the stakes were large.

Within the linear framework, we can consider a broader class of models in which adjusted utility is given by

$$v_i = u_i + \sum_{j \neq i} \beta_{ij} u_j \ ,$$

and the coefficients β_{ij} are determined from players types or other details about the game. For example, Ledyard [1995] proposes a specification in which $\beta_{ij} = \gamma_i (u_j^f - u_j)$, where u_j^f is a "fair amount" which he does not define. A more detailed specification can be found in Rabin [1993] who assumes that a player is interested in what is fair for himself, rather than what is fair for the other player. His specification is $\beta_{ij} = \gamma_i (u_i - u_i^f)$ where the "fair amount" is a fixed weighted average of the maximum and minimum Pareto efficient payoff given player *i*'s own choice of strategy, and the coefficient γ_i itself is endogenous in a somewhat complicated way. Both of these theories suppose that players care about whether their opponents play "fairly" and run into the problem that there is no obvious notion of fairness that applies to all games. The strategy of the specification we have chosen is to suppose that players do not care about whether their opponents play "fairly," but rather whether their opponents are nice people. This has the advantage that there is no need to answer the question of what is "fair."

4. Ultimatum

We begin with the study by Roth et al [1991] on ultimatum bargaining in four countries. The extensive form of this game is shown below.

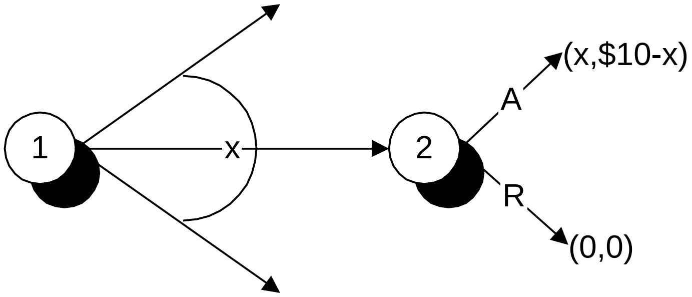


Figure 1

Player 1 demands an amount *x* between 0 and $10.[5] Player 2 may either accept or reject. If he accepts, player 1 gets the demanded amount, and player 2 gets the remainder of $10. If he rejects neither player receives anything.

In the usual selfish case where the altruism coefficient is $a_i = 0$ for all players, player 2 should accept any demand less than $10. Subgame perfection then requires that player 1 demand at least $9.95.[6] The actual results of the experiments were quite different. Table 1 below pools results[7] of the final (of 10) periods of play in the 5 experiments reported in Roth et al [1991], with payoffs normalized to $10. It should be emphasized that in order to eliminate as much as possible the effect of learning, only data

[5] This is the base case. In other countries than the United States payoffs were calibrated to match $10 US in local currency. In one treatment in the US the payoff was $30 rather than $10. Since the magnitude of the payoffs do not matter for the theory here, we normalized all the payoffs to $10.

[6] Players were constrained to demands stated in units equivalent to $0.05.

[7] All four countries are pooled. There are differences between the results of the different treatments, but even with the pooled data, the evidence on altruism is very limited, as we shall see. While the differences are statistically significant, they are not large in absolute terms (in the US demands were generally $5, in Israel generally $6). Since the theory says the treatments should not make a difference we elected to pool the data.

from the last of 10 rounds is used. The majority of players demanded no more than $6 and many demands of less than $10 are rejected.

Demand	Observations	Acceptance
$9.00	1	100%
$8.25	1	100%
$8.00	4	50%
$7.50	5	80%
$7.00	10	80%
$6.75	5	20%
$6.50	6	83%
$6.25	5	80%
$6.00	30	83%
$5.75	9	100%
$5.50	17	71%
$5.25	5	100%
$5.00	31	100%
$4.75	1	100%

Table 1

The altruistic model has implications for this game, independent of the distribution *F*.

Proposition 1: Regardless of *F*, in no sequential equilibrium will any demand be made for less than $5.00, and any demand of $5.00 or less will be accepted.

Proof: First observe that $10-x+\beta x>0$ if $-1<\beta$ and $x\leq 5$, so indeed any demand of $5.00 or less will be accepted. But $x+\beta(10-x)$ is increasing in *x* for $\beta>-1$. Since a demand of less than \$5.00 can be increased without reducing the probability it will be accepted, it cannot be optimal to make such a demand.

■

In fact in the data only one demand of less than \$5.00 was ever made, and it was for \$4.75 and was accepted, so the data are consistent with Proposition 1.

To simplify the remainder of the analysis, it will be convenient to pool the results. The demands are grouped into three categories: demands in the range \$4.75-\$5.25 are treated as \$5.00 demands; demands in the range \$5.50 to \$6.50 are treated as \$6.00 demands, and demands of \$6.75 and higher are treated as \$7.00 demands. For simplicity, we will only allow players to make demands in even dollars. The pooled data is summarized in Table 2; the column labeled "Adjusted Acceptance" is explained below.

Demand	Observations	Frequency of Observations	Accepted Demands	Probability of Acceptance	Adjusted Acceptance
\$5.00	37	28%	37	1.00	1.00
\$6.00	67	52%	55	0.82	0.80
\$7.00	26	20%	17	0.65	0.65

Table 2

We will assume that the distribution F of altruism coefficients places weight on three points $\overline{a} > a_0 > \underline{a}$. We refer to these as the altruistic, normal and spiteful types of players. Since there are three demands made in equilibrium, and more altruistic types will prefer to make lower demands, we will look for an equilibrium in which the altruistic types demands \$5.00, the normal type \$6.00 and the spiteful type \$7.00. Consequently the probabilities of the three types must be 0.28, 0.52 and 0.20 respectively, as this is the frequency of demands in the sample. The \$5.00 demand is clearly accepted by all three types. The \$6.00 demand is accepted by 82% of the population. However, because of sampling error, it is impossible to reject the hypothesis that the actual acceptance rate is 0.80 at less than a 28% level of confidence. Since 80% of the population corresponds to the spiteful types rejecting the demand, we will assume that this is in fact the actual acceptance rate (the column "Adjusted Acceptance" in Table 2). In other words, we assume that the $6.00 demand is rejected by the spiteful types and accepted by the normal and altruistic types. The $7.00 demand is accepted by 65% of the population, corresponding to all the altruistic types (28%) and a fraction 71% ($0.71\times 0.52\approx 0.37$) of the normal types. This implies that the normal types must be indifferent between accepting and rejecting the $7.00 demand.

Based on these considerations, we can characterize the parameters consistent with equilibrium by means of 6 incentive constraints given in Lemma A in the Appendix. Our next set of results examines more closely which parameter values are consistent with the data.

Proposition 3: There is no sequential equilibrium with $\lambda = 0$.

Proof: Since the normal type must be indifferent to accepting or rejecting the $7.00 demand, we have

$$3+\frac{a_0+\lambda\underline{a}}{1+\lambda}7=0$$

(also equation (6) in Lemma A). Setting $\lambda = 0$, we see that $a_0 = -3/7$; that is, the normal type must be relatively spiteful. But we may calculate in this case that the utility the normal type gets from making the $6.00 demand is $3.43, while the utility from making the $7.00 demand is $3.71. This contradicts the fact that the normal type must prefer the $6.00 demand.

■

This is actually a corollary of the next proposition, but we give a separate proof, because of the importance of the result. What this proposition says is that a model of pure altruism is not consistent with the data from the ultimatum experiments. The problem is that the acceptance of demands is such that players must be relatively spiteful. But spiteful players would not make the modest demands observed in ultimatum. We have experimented with several other specifications of the distribution of the coefficient of altruism in the model of pure altruism, and none can explain this feature of the data.

Proposition 4: In sequential equilibrium $-.301 \le a_0 \le -.095$, $-1 < \underline{a} < -0.73$, $0.584 \ge \lambda \ge 0.222$.

Proof: From manipulating the incentive constraints characterizing and equilibrium; see the Appendix for details.

■

There are a variety of parameter values for which there are sequential equilibria consistent with the data. Each column of Table 3 gives a set of parameter values for which there exists an equilibrium of the type described that is consistent with the data.

.

$\bar{a}$	0.10	0.30	0.40	0.90	0.90	0.90	0.90
a_0	-0.22	-0.22	-0.22	-0.22	-0.27	-0.26	-0.20
$\underline{a}$	-0.90	-0.90	-0.90	-0.90	-0.87	-0.90	-0.90
λ	0.45	0.45	0.45	0.45	0.36	0.35	0.49

Table 3

As can be seen a wide range of values of $\bar{a}$ is consistent with the data. Experimentation indicates somewhat less flexibility in the remaining parameters than is indicated in Proposition 3. In particular, it appears to be difficult to get $\underline{a}$ larger than -0.87 (versus the known lower bound of -0.73). Values of λ are difficult to find lower than 0.35 (against the known lower bound of 0.22). Values of λ are difficult to get higher than 0.49, as against the known upper bound of 0.58. We were also unable to discover equilibria with values of a_0 below -0.2, although the known lower bound is only $-.301$. The parameters $\lambda = 0.45$, $\underline{a} = -0.9$ and $a_0 = -0.22$ lie in the midrange of parameter values consistent with proposition 3, and with the range of parameters experimentation shows is feasible. Moreover, from the table above, these parameters are consistent with a wide variety of different values of $\overline{a}$. In the remainder of the paper, in evaluating other experiments, we will, somewhat arbitrarily, choose to work with these parameter values.

While we have found a set of parameter values that is consistent with both sequential equilibrium and with the data. One question we have not yet answered is how much predictive power the theory has. In particular, for our favored parameter values, are there other sequential equilibria than the one found in the data? In particular can there be a pooling equilibrium in which all players play the same way? The question is affirmative:

Proposition 5: If $\lambda = 0.45$, $\underline{a} = -0.9$, $a_0 = -0.22$, $\overline{a} = 0.29$[8] and the corresponding probabilities of the spiteful, normal and altruistic groups are $0.20, 0.52, 0.28$ then there are two pooling equilibrium outcomes that are sequential: one in which all demands are \$7.00 and one in which all demands are \$8.00. In both cases, the sequential equilibrium offers are accepted by normal and altruistic types, and rejected by the spiteful types.

Proof: By computation; see the Appendix

■

The predictive power of the theory is about what we would expect from a signaling model. As usual, it is difficult to rule out pooling at different levels, and likely there are several separating equilibria as well as the one observed. On the other hand, it is by no means true that anything is an equilibrium, and indeed, we are able to rule out pooling equilibria at \$5.00, \$6.00, \$9.00 and \$10.00 which was not *a priori* obvious.

[8] The value of $\overline{a}$ is the estimate from the Centipede experiment discussed below, and is in the range consistent with the separating equilibrium observed in ultimatum.

Notice also that while these pooling equilibria are inconsistent with the data, they are considerably close to the data than the equilibrium without altruism in which only very large demands are made, and all are accepted.

5. Competitive Auction

For any value of λ if the distribution F of altruism coefficient a is degenerate placing all weight on $a = 0$ the model is the traditional model of all selfish players. Thus the extent to which the distribution F distributes weight away from the origin measure the extent to which the model is different than the selfish players model. To explain the ultimatum experiments, the departure from the selfish players model is quite large. For example, at least 20% of the population as a group has a mean coefficient of -0.73 or worse; even the middle group of 52% of the population seems to have a substantial degree of spite. In other words, we are proposing a substantial departure from the model of selfish agents. This, however, poses a potential problem: in many experiments, especially in market games, double oral auctions, and so forth, the model of selfish agents explains the data well. If the model proposed here is useful, then it must continue to explain the results of games already explained by the selfish player model.

From an intuitive perspective, the experiments in which the selfish player model has worked well are experiments with a high degree of competitiveness. In a relatively competitive environment, player can have a significant effect on their own utility, but it is difficult for them to transfer utility to or from other players. Consequently, we might expect that spite or altruism would play very little role in such environments. To explore this issue, we turn now to another experiment conducted by Roth et al under very similar conditions to the ultimatum experiment reported above. We argue that regardless of the distribution of altruism and spite in the population, we would expect to see the competitive equilibrium occur in this experiment (as was indeed observed).

In the market game experiment nine identical buyers submit an offer to a single seller to buy an indivisible object worth nothing to the seller and $10.00 to the buyer. If the seller accepts he earns the highest price offered, and a buyer selected from the winning bids by lottery earns the difference between the object's value and the bid. Each player participates in 10 different market rounds with a changing population of buyers. This game has two subgame perfect equilibrium outcomes: either the price is $10.00, or everyone bids $9.95. In fact by round 7 the price rose to $9.95 or $10.00 in every experiment, and typically this occurred much earlier.

Altruistic equilibria may be partially characterized by:

Proposition 6: In any sequential equilibrium all offers of $5.00 or better are accepted. There exist sequential equilibria in which buyer offers are independent of how altruistic they are and the seller always sells. If other buyer offers are independent of how altruistic they are and the seller always sells, then buyer preferences are independent of how altruistic the buyer is. Consequently the set of sequential equilibria in which buyer offers are independent of how altruistic they are is independent of the distribution of altruism in the population.

Proof: Let β be the coefficient of altruism adjusted for the opponent's altruism. If the seller accepts an offer of x he gets $x+\beta(10-x)$; if he rejects he gets 0. Hence he accepts if $x+\beta(10-x)\geq 0$. Since $\beta>-1$ this is true provided that $x\geq \$5.00$, so all offers of $5.00 or better are accepted, just as in the case of bargaining.

Turning to the buyers, if there are multiple offers at $10.00 then no buyer can have any effect on their own utility, since the seller always gets $10.00 and the buyers $0.00 regardless of how any individual buyer deviates. More generally, suppose that buyer offers are independent of how altruistic they are, and the seller always buys. The key observation is that by bidding low a buyer does not prevent the transaction from taking place, he merely fails to get a valuable item for himself. In particular, if the buyer fails to buy, but the transaction takes place anyway, this yields a net benefit to the rest of the population of 10. In other words, an offer *x* accepted with probability *p* gives utility

$$p((10-x)+\beta x)+(1-p)\beta 10=10\beta+(1-\beta)p(10-x)$$

which regardless of β are the same preferences as $p(10-x)$. Since preferences are independent of altruism, players are willing to use strategies that are independent of how altruistic they are, so every equilibrium without altruism is an equilibrium with altruism.

■

6. Centipede

So far we have merely fit parameters to observations; when the model with $\lambda=0$ did not fit the data, we simply introduced a new parameter to explain the results. To actually test the theory, we must hold fixed the parameters we found from ultimatum, and use them to explain the results of a different experiment. One famous experiment that is not explained well by selfish players is grab-a-dollar, and we will next examine such an experiment.

The specific experiment is a variation on grab-a-dollar that McKelvey and Palfrey [1992] call the centipede game. The extensive form, together with the actual conditional probabilities of moves computed from the 29 experiments over the last 5 of 10 rounds of play, is reported in Figure 2. Each round, players played against an opponent they had not previously played against so repeated game effects should not be an issue here.

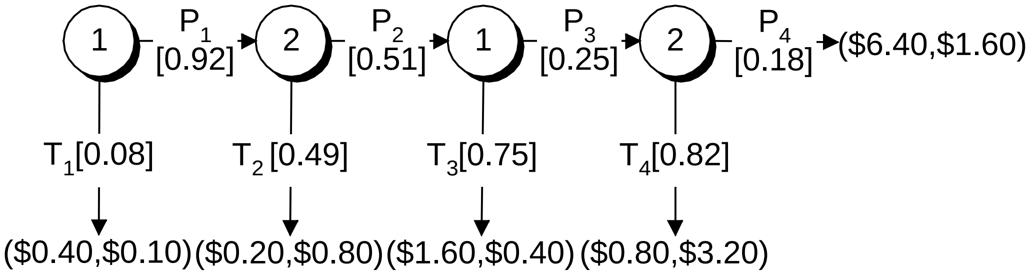


Figure 2

Without altruism, these results do not make much sense: 18% of player 2's that reach the final move choose to throw away money, for example. Moreover, with normal preferences, the unique Nash equilibrium is for all player 1's to drop out immediately. Note however, that altruistic players may wish to give money away at the end, since the gain to the other player is much greater than the loss, and that this will give other players reason to stay in the game. Notice also that there is a kind of reputation effect of the type discussed by Fudenberg and Levine [1989], in the sense that by staying in a player signals he is an altruistic type, and as a result is more likely to receive kind treatment from his opponent.

We will model this game using the same model of three types we used to analyze ultimatum. We will assume $\lambda = 0.45$, $\underline{a} = -0.9$ and $a_0 = -0.22$, which are parameters that have been narrowed down by the data on ultimatum. The probabilities of the spiteful, normal and altruistic groups are $0.20, 0.52, 0.28$ respectively. Notice that virtually no player 1's drop out in the first move, so that the distribution of types the second time player 1 moves should be essentially the prior distribution. Moreover, in this second move by player 1, 25% of the players choose to continue, which, within the margin of sampling error, is quite close to the 28% of player 1's that are altruistic. So we will assume that in player 1's final move, all the altruistic types pass, and all the other types take, and we will analyze the following modified data

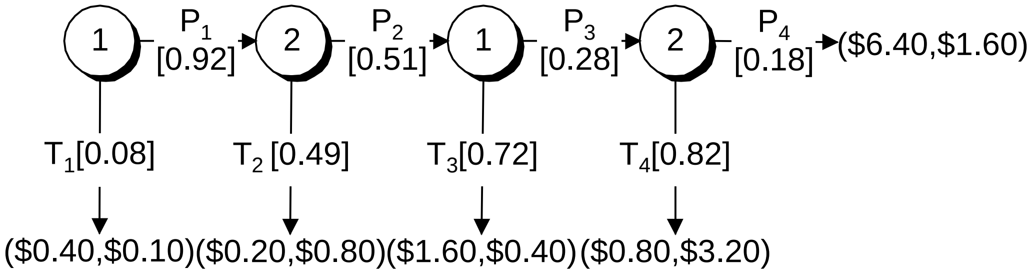


Figure 3

We examine the play of player 2's at the final node first. Since spiteful and normal types will drop out before altruists, and fewer players pass than the 28% of the population that are altruists, we conclude that the altruistic types must be indifferent between passing and taking. Since all player 1's are known to player 2 to be altruists at this point, this means that

$$3.20+\frac{\overline{a}+\lambda\overline{a}}{1+\lambda}0.80=1.60+\frac{\overline{a}+\lambda\overline{a}}{1+\lambda}6.40 .$$

From this we may calculate $\overline{a}=2/7\approx 0.29$. This is one of the wide range of values consistent with the ultimatum data. Notice that this does not yet represent a test of the theory; we are merely calibrating the final parameter that was not pinned down by the ultimatum experiment. However, now that the entire distribution of altruism is pinned down, we can test the theory be asking whether the decisions of players in earlier rounds are consistent with the theory.

We consider player 1's final decision to pass or take. Since 51% of the player 2's previously passed, including all the altruistic player 2's, this means that $0.28/0.51=0.55$ of the player 2's are altruists and the remaining 0.45 are normal types. If player 1 takes, he then places a weight on his opponent's utility of

$$a_T\equiv\frac{a_0+\lambda(0.55\times\overline{a}+0.45\times a_0)}{1+\lambda}=-0.13 .$$

His utility if he takes is $1.60+a_T 0.40=1.55$. On the other hand, if he passes, he has a 0.18 chance of an altruistic opponent and getting $6.40 for himself and $1.60 for the opponent, resulting in a utility of $6.31. He also faces a 0.82 chance of an opponent who is $0.45/0.82 = 0.55$ likely to be normal and 0.45 likely to be altruistic. This yields a utility of \$0.33. Averaging over his opponent passing and taking in the final round, yield the expected utility to passing of \$1.40, less than the utility of taking. In other words, the normal type should take. This implies the spiteful type should take, and it is apparent from the fact that the normal type is nearly indifferent that the altruistic type should pass. This is as the data suggest.

Continuing on in this way, we can at each node compute the utility from taking and passing as shown in Table 4:

Node	Type	Take Utility	Pass Utility	Difference
1's last move	a_0	\$1.55	\$1.40	\$0.14
2's first move	a_0	\$0.76	\$0.85	-\$0.09
1's first move	$\underline{a}$	\$0.33	\$0.49	-\$0.16

Table 4

From the table we see that the spiteful type 1 player is never the less willing to pass in the first period. The only inconsistency is that the normal type of player 2 in his first move should be indifferent between passing and taking, and in fact prefers to pass. Notice however, that we have made no effort to calibrate any of the parameters to the exact indifference of this type, and despite this, the preference for passing is very slight: a mere \$0.09 advantage. In fact the data seem strikingly consistent with the model and the estimates of altruism and spitefulness from the ultimatum game and the final period of this game.

One way to understand how well this model explains the data is to compare it to the standard non-altruistic model. In Fudenberg and Levine [1996] we argue that an appropriate metric for measuring departures from the theory is the expected loss of players. We just calculated these losses for the altruism model. The results are summarized below in Table 5. Here the column marked loss replicates the final column of Table 4. The column marked population is the fraction of players (of both types) who suffer the loss. The second two columns report the same data for selfish players. The basis of the calculation it is optimal for a selfish player to pass in every round but the final one. In the penultimate round, player 1 can gets an expected money earning of $1.80 by passing; in his first move, player 2 gets an expected money earning of $1.18 by passing, while in the first move, it is worth $1.02 for player 1 to pass. We do not include the losses of selfish player 1's that give money away in the final round, since in the altruism model we used this data to fit a free parameter.

	Altruism Model		Standard Model	
Node	Loss	Population	Loss	Population
1's last move	$0.00		$0.20	0.17
2's first move	$0.09	0.17	$0.38	0.23
1's first move	$0.16	0.04	$0.62	0.04

Table 5

We can summarize the results of this table by computing an overall expectation: the deviation from the theory in the altruism model is an expected loss per player per game of about 1.5¢. By way of contrast, the deviation of the data from the standard model of selfish players is an expected loss per player per game of about 14.5¢, nearly an order of magnitude higher.

7. Public Goods Contribution Game

It is well known that there is a great deal of altruism in public goods contribution games. Our examination of ultimatum bargaining and centipede suggests a relatively spiteful population with few (28%) altruists. Can this be reconciled with the large amount of altruism found in public goods contribution games? To answer this question, we examine a public goods contribution game studied by Isaac and Walker [1988].

The game is a simultaneous move *n* person game, in which each individual must decide whether or not to contribute a number of tokens to a common pool, or consume them privately. If m_i is the number of tokens contributed (we may normalize so that the total number of available tokens per player is 1), the direct utility is given by

$$u_i = -m_i + \gamma \sum_{j=1}^{n} m_j \ .$$

There were four different treatments (pairs of values of γ, n), and each treatment was repeated 6 times.

As in the case of ultimatum and centipede we assume $\lambda = 0.45$, $\underline{a} = -0.9$, $a_0 = -0.22$ and $\bar{a} = 0.29$. The corresponding probabilities of the spiteful, normal and altruistic groups are $0.20, 0.52, 0.28$ respectively. We may calculate the mean population altruism equal to $\hat{a} = -0.21$. An individual contemplating a contribution to the public goods game assuming his opponents are drawn randomly with a population with degree of mean altruism

$$v_i = -m_i + \gamma\left(m_i + \hat{m}_{-i}\right) + \frac{a_i + \lambda \hat{a}}{1+\lambda}(n-1)\left(-\hat{m}_{-i} + \gamma(m_i + (n-1)\hat{m}_{-i}\right)$$

where $\hat{m}_{-i}$ is the mean contribution by players other than player *i*. Differentiating this with respect to the own contribution m_i we see that the player will wish to contribute if and only if

$$-1 + \gamma + \frac{a_i + \lambda \hat{a}}{1+\lambda}(n-1)\gamma \geq 0 \, .$$

From this we may compute the unique cut-off value $a*$ such that a player with a higher degree of altruism contributes, and player with a lower degree of altruism does not contribute. This is given by

$$a* = \frac{(1-\gamma)(1+\lambda)}{(n-1)\gamma} - \lambda \hat{a} \, .$$

Using $\lambda = 0.45$ and $\hat{a} = -0.27$, we can compute the different cut-off values corresponding to the different treatments.

In the actual experiment four treatments were used with different numbers of players and different values for the marginal per capital return γ. Following Isaac and Walker [1988] we will consider the final round of play only; each treatment was repeated three times. The different treatments, the data from the experiments, and the cutoff values a are all reproduced in Table 6. The column labeled $\overline{m}$ reports the fraction of the population that would have had to contribute if all contributions were either zero or the maximum allowable.

γ	n	$\overline{m}$	a^*
0.3	4	0.00	1.13
0.3	10	0.07	0.38
0.75	4	0.29	0.17
0.75	10	0.24	0.06

Table 6

We should begin by noting that the experimental design was not ideal in the sense that it does not reflect the simple one shot model we use to explain it. In fact players played repeatedly against the same opponents, and we can not be sure what information was revealed about their types prior to the final round reported above.

We should begin by observing that most players that contributed tokens contributed less than the maximum allowable. This is inconsistent with our theory. Because payoffs are linear in own and opponents monetary payoff, except in the zero probability event of exact indifference, the contribution should either be the minimum or maximum allowable.[9] Although not part of the theory as exposited here, this failure is

[9] It is natural to speculate that this problem could be remedied by the type of non-linear preferences considered by Andreoni and Miller [1996]. However, while they are equally successful in predicting the

mitigated by the fact that both theoretically (see Fudenberg and Levine [1995]) and empirically (see McKelvey and Palfrey [1995]) there is reason to believe players near indifference randomize; the partial contribution can be the result of such randomization. Because the theory cannot explain individual contributions, we take our objective to explain the aggregate contributions made by the population. This is measured by $\overline{m}$, the fraction of the population contributing all of their tokens required to match the aggregate contribution level.

In the first treatment the theory predicts (since $a_i \leq 1$ for all players) there should be no contributions, and indeed there are none. In the second treatment, our theory has 28% of the population altruistic with an average coefficient of 0.28; here we see 7% of the population with a coefficient of at least 0.38, consistent with the degree of altruism from the previous experiments.

The third treatment yields 29% of the population with a coefficient of at least 0.17, also generally consistent with 28% of the population having an average altruism coefficient of 0.28.

The final treatment is also generally consistent with the theory: 24% of the population has an altruism coefficient of at least 0.06. Notice however, that there were fewer contributions in the fourth treatment than the third treatment, despite the fact that the theory predicts the opposite. This is consistent, however, with the possibility that the anomalous results of the third treatment are due to sampling error: since each experiment was repeated 3 times, there are only 12 observations.

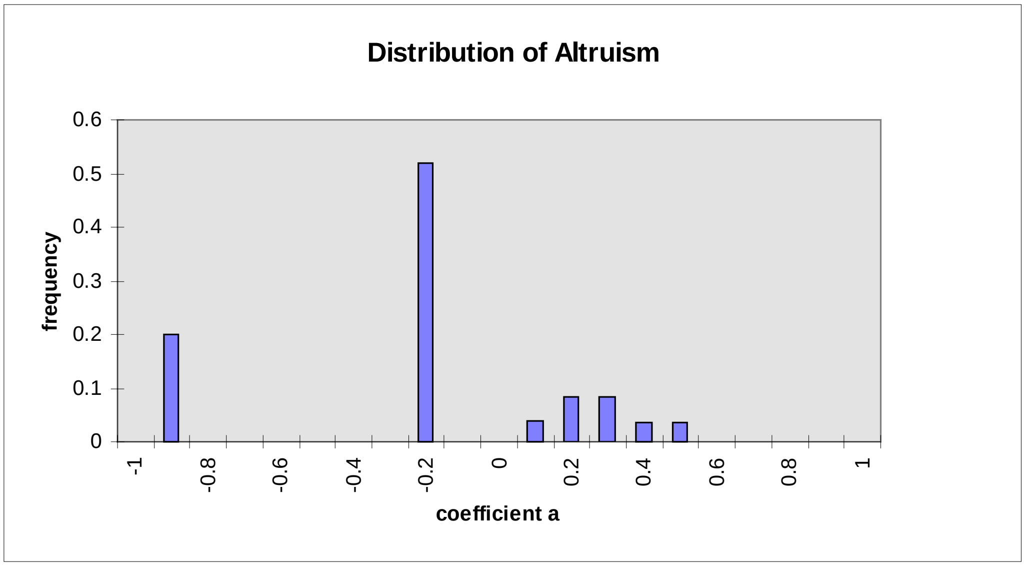
If we assume that the large fraction of contributions in the third treatment is due to sampling error, then we should conclude that the actual fraction of the population that would contribute should be 0.24; the fraction of altruists 0.28 less the 0.04 of the population who are altruists with coefficients below 0.17. If we assume three types of

aggregate outcome of these experiments, they also have little success in predicting the number of individuals who contribute nothing.

altruists with coefficients 0.06+*x*, 0.17+*x* and 0.38+*x* and frequencies 0.04, 0.17 and 0.07 having, according to our previous conclusions, a mean of $\bar{a} = 0.29$, then *x*=0.073, and the three altruistic groups have coefficients of 0.133, 0.243 and 0.453.

8. Conclusion

The theory fares relatively well in the experiments examined here, particularly in comparison to the selfish theory, which makes wildly wrong predictions except in the market game. We can summarize the distribution of altruism coefficients that works relatively well:



The one really surprising feature of this distribution is the large mass of individuals with relatively negative coefficients; that is, the group of 20% of the population with mean coefficient -0.9.

We should point out the significance of this theory for games in which mixed strategy equilibria are observed. With selfish players, at the Nash equilibrium, a player can transfer money to or from his opponent at no cost (since he is indifferent). The deviation from Nash equilibrium depends on whether at the Nash equilibrium the marginal (indifferent) player is spiteful or altruistic. In the case in which the spiteful player is marginal he will wish to transfer money away from his opponent. To prevent him from doing so, the opponent's strategy will have to be adjusted to keep him indifferent. In a symmetric game, this means that (with the usual perverse comparative statics of mixed strategies) equilibrium payoffs will be higher than would be the case with purely selfish players. It is known that non-zero sum mixed equilibria differ systematically from Nash equilibria (see, for example, McKelvey and Palfrey [1995]): it remains to be seen whether this can be explained by the population distribution of altruism reported here.

There are several experiments that cannot be explained by this theory: one such is the dictator experiment in which one player decides whether or not to contribute money to an opponent. In these experiments contributions are positive, although with linear utility and $a_i < 1$ the theory predicts no contributions will ever be made.[10] It possible to dismiss this result along the lines of the partial contributions in the public goods experiment: players who are very altruistic are near indifferent and are randomizing. In addition, we am unaware of any dictator experiments conducted in the same way as the experiments here where players played repeatedly and had a chance to gain experience with the game.

A less radical departure from the predictions of the theory, but an important one can be found in recent work by Van Huyck, Battalio and Rankin [1996]. This is a public goods experiment similar to Isaac and Walker's, except that players had the opportunity to be spiteful as well as altruistic. In fact, despite the fact that the estimates here are that many players are quite spiteful, very little spitefulness is observed in Van Huyck, Battalio and Rankin [1996]. There also appears to be less altruism than in Isaac and Walker's experiment, which suggests that the experimental conditions may have been different in an important way, but hardly begins to explain the lack of spitefulness. There are two significant ways in which the Van Huyck, Battalio and Rankin experiment differs from

[10] The results of dictator are evidence in favor of the type of non-linearity favored by Andreoni and Miller [1996], and indeed, their experiments focus on dictator.

the experiments reported here in which spitefulness appeared to play an important role (ultimatum and centipede). The game is a one-shot game, so players could not react to "spiteful" play by opponents, and the game is a four-player game rather than a two-player game.

The fact that there was less spitefulness with four players raises an issue about the theory: we have assumed that the altruism/spite coefficient is independent of how many opponents there are. This is sensible in the case of altruism, but one explanation of spite is that it is really "competitiveness," that is, the desire to outdo opponents. In this case, it is not the total utility of opponents that matter, but some measure of their average or maximum utility. However, while there is obviously much scope for the systematic study of how spite might vary with the number of opponents, even the simple adjustment of deflating our estimated spite coefficients by the number of opponents does not reduce predicted spiteful play to the low level observed in Van Huyck, Battalio and Rankin. This is evidence in favor of the hypothesis that the extensive form of the game matters: that is, that retaliation for past poor performance is not due only to signaling of types.

In discussing multi-player public goods experiments, we should mention also the work of Palfrey and Prisbrey [1997] on altruism and the warm glow effect. They argue that the value of contributions to other players are not so important as the cost of the donation; that there is a "warm glow" effect that players wish to incur a particular cost of contribution, regardless of the benefit. If the cost is lower, there will more contribution, but if the benefit is higher there will be no increase in contribution. They study a 4-person public goods contribution game in which players must decide whether or not to contribute a single token.[11] Each period each player would randomly draw a value ξ_i for his token, uniformly distributed on 1 to 20. If the token were kept, the value of the token

[11] There was also a treatment in which players could contribute up to 9 tokens.

would be paid; if the token were contributed, a fixed amount γ would be paid to each player. The monetary payoffs, are given by

$$u_i = \xi_i - \xi_i m_i + \gamma \sum_{j=1}^{n} m_j \ .$$

Each player played 20 rounds with a fixed value of γ. They did this four times with different values of γ. Each round players were shuffled so as to minimize repeated game effects.[12] We consider only results from the second 10 rounds with each value of γ, so that players are relatively "experienced" and we can observe equilibrium behavior rather than learning behavior. Data from their experiment is reported in Table 7 below.[13]

	$\gamma = 3$		$\gamma = 15$	
$\xi_i - \gamma$	Gain ratio	$\overline{m}$	Gain ratio	$\overline{m}$
5	1.8	0.00	9.0	0.60
3-4	2.7	0.18	13.1	0.67
1-2	6.8	0.27	33.7	0.79
0	∞	0.88	∞	0.86

Table 7

Because there are relatively few observations in each cell, the data is pooled as indicated in the table.

The most significant feature of this data is that it does not bear out the Palfrey and Prisbrey conclusion of a "warm glow" effect. For a given net cost of contributing $\xi_i - \gamma$, far more contributions are made when $\gamma = 15$ than when $\gamma = 3$, indeed this is true whenever the cost of contribution is positive. This raises a methodological issue: their conclusion is based on a hypothesis test using maximum likelihood estimation in a fully

[12] A repeated game treatment was also considered with little consequence to the results.
[13] I am grateful to Tom Palfrey for providing me with the raw data.

specified model. To account for heterogeneity, they allow the warm glow effect to differ from player to player, but not the degree to which they are altruistic. As a result the coefficient representing how altruistic players are (the weight on other players' utility) is effectively averaged over the population. Since, as the table above shows, for many players the slope was zero this specification error leads them to substantially underestimate the extent to which a portion of the population was indeed altruistic.

While the maximum likelihood approach deals well with the problem of sampling error, it deals poorly with specification error. The approach we have taken here is to use method of moments estimation, and, recognizing that the model is misspecified, to give easily interpretable measures of the departures from the theory. This for example, is the approach we take in analyzing the centipede game, where we use the Fudenberg and Levine [1996] approach of reporting players' losses. Similarly, while the model is clearly misspecified in the public goods experiment, failing as it does to explain the fact that individual players do not contribute all or nothing, we can ask the question of whether it does a good job of predicting aggregate outcomes. This idea that the model may work well to explain features of the data we are interested in, while poorly describing some of the details we are less interested in, is extremely difficult to capture in a maximum likelihood approach.

Finally, we turn to other implications of the theory that could be tested in future experiments. For example, there is a set of implications of the theory for one-player games that has not been examined experimentally: The theory predicts that spiteful play should take place even in a one player setting. In other words, if a single player is given the option to deprive an opponent of money at a sufficiently modest cost to himself, then we should observe 20% or more of players availing themselves of this option. Moreover, in two move games such as the best-shot game discussed by Prasnikar and Roth [1992], in which all first players are observed to play the same way, the second mover should exhibit the same degree of altruism (or spite) when posed with the simple second period decision problem with no first player move.[14] In a similar vein, it is possible to confront players with the same choices faced by the second player in ultimatum bargaining, but also as a simple decision problem and no first period move. The theory predicts that this change will effect the play of the second players due to the absence of signaling by the first player, but it makes very specific and easy to calculate predictions about the frequency of acceptance and rejections. Since this is a two-player experiment, it provides a more direct test than Van Huyck, Battalio and Rankin that spite involves explicit retaliation and not merely signaling. Finally, a referee made the interesting suggestion that it would be possible to have players engage in dictator experiments, with their behavior in the experiment announced to their opponent in a subsequent ultimatum game. According to the theory, relatively ungenerous players in dictator should get poor offers in ultimatum.

[14] In the full information best-shot reported in Prasnikar and Roth [1992] there is slightly less altruism and slightly less spite exhibited by the second player than in the experiments reported here. However, only eight second players were involved in the experiment, so sampling error is a major problem.

Appendix

Lemma A: A sequential equilibrium matching the data will be given by parameters $1 > \overline{a} > a_0 > \underline{a} > -1, 0 \leq \lambda \leq 1$ such that

$$(1) \qquad (6+\frac{a_0+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8-(5+\frac{a_0+\lambda(.28\overline{a}+.52a_0+.20\underline{a})}{1+\lambda}5) \geq 0$$

$$(2) \qquad (6+\frac{\overline{a}+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8-(5+\frac{\overline{a}+\lambda(.28\overline{a}+.52a_0+.20\underline{a})}{1+\lambda}5) \leq 0$$

$$(3) \qquad 4+\frac{\underline{a}+\lambda a_0}{1+\lambda}6 \leq 0$$

$$(4) \qquad (7+\frac{\underline{a}+\lambda(.43\overline{a}+.57a_0)}{1+\lambda}3).65-(6+\frac{\underline{a}+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8 \geq 0$$

$$(5) \qquad (7+\frac{a_0+\lambda(.43\overline{a}+.57a_0)}{1+\lambda}3).65-(6+\frac{a_0+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8 \leq 0$$

$$(6) \qquad 3+\frac{a_0+\lambda\underline{a}}{1+\lambda}7 = 0$$

$$(7) \qquad (7+\frac{\underline{a}+\lambda(.43\overline{a}+.57a_0)}{1+\lambda}3).65 \geq 2.80$$

Proof: We first consider the $5.00 demand. Since all types will accept this demand, the adjusted utility received by a player demanding this amount is

$$5+\frac{a+\lambda(.28\overline{a}+.52a_0+.20\underline{a})}{1+\lambda}5$$

In addition, if the spiteful type accepts, all types will accept the demand. Since the demand is known to be made by the altruistic type, for spiteful type to accept we must have

$$5+\frac{\underline{a}+\lambda\overline{a}}{1+\lambda}5 \geq 0$$

However, this inequality is always satisfied for $\underline{a}, \overline{a} > -1$.

We turn next to the \$6.00 demand. Since only the altruistic and normal types accept this demand, the adjusted utility received by a player demanding this amount is

$$(6+\frac{a+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8$$

For the normal type this must yield more utility than the \$5.00 demand (and therefore it does also for the spiteful type)

(1) $$(6+\frac{a_0+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8-(5+\frac{a_0+\lambda(.28\overline{a}+.52a_0+.20\underline{a})}{1+\lambda}5)\geq 0$$

while for the altruistic type it must yield less utility

(2) $$(6+\frac{\overline{a}+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8-(5+\frac{\overline{a}+\lambda(.28\overline{a}+.52a_0+.20\underline{a})}{1+\lambda}5)\leq 0$$

Moreover, the spiteful type must reject, and the normal type accept (in which case the altruistic type will also accept) the \$6.00 demand. Since the demand is known to be made by the normal type for the spiteful type to reject, we must have

(3) $$4+\frac{\underline{a}+\lambda a_0}{1+\lambda}6\leq 0$$

while for the normal type to accept, we must have

$$4+\frac{a_0+\lambda a_0}{1+\lambda}6=4+6a_0\geq 0$$

Next we have the \$7.00 demand. Since the altruistic and 71% of the normal types accept this demand, the adjusted utility received by a player demanding this amount is

$$(7+\frac{a+\lambda(.43\overline{a}+.57a_0)}{1+\lambda}3).65$$

The spiteful type must prefer this to the \$6.00 demand so that

(4) $$(7+\frac{\underline{a}+\lambda(.43\overline{a}+.57a_0)}{1+\lambda}3).65-(6+\frac{\underline{a}+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8\geq 0$$

while the normal type must prefer the \$6.00 demand (implying that the altruistic type does as well)

(5) $$(7+\frac{a_0+\lambda(.43\overline{a}+.57a_0)}{1+\lambda}3).65-(6+\frac{a_0+\lambda(.35\overline{a}+.65a_0)}{1+\lambda}4).8\leq 0$$

The normal type must be indifferent between accepting or rejecting the $7.00 demand (in which case the spiteful player rejects, and the altruistic player accepts). Since the demand is known to be made by the spiteful player, this forces

(6) $$3+\frac{a_0+\lambda\underline{a}}{1+\lambda}7=0$$

Notice that this implies that a weighted average of $a_0, \underline{a}$ is equal to -3/7; since $a_0 > \underline{a}$ this implies that $a_0 > -3/7 > -2/3$, which implies the inequality above, that the normal player accepts the $6.00 demand.

From Proposition 1, we may ignore demands of less than $5.00. However, we need to consider demands of more than $8.00.[15] Since in our proposed equilibrium, only spiteful types will demand as much as $7.00, so we consider the most favorable case for equilibrium, that in which beliefs are that any demand greater than $7.00 is made by a spiteful type. Because the spiteful type has the most reason to make large demands, these beliefs are consistent with quite strong refinements such as the intuitive criterion (Cho and Kreps [1987])). Given these beliefs, it is clear that since normal types are indifferent between accepting and rejecting the $7.00 demand that only the altruistic type will accept larger demands. Since it is most favorable for making large demands, let us suppose that the altruistic type is sufficiently altruistic as to accept all demands. In this case if any type is to make a demand above $7.00 the spiteful type will wish to do so, and will wish to demand a full $10.00. The demand is accepted with probability 28% corresponding to the fraction of altruistic types, so the expected utility is $2.80. On the other hand, a demand of $7.00 is accepted with probability 65%, and gives a spiteful type a utility of

[15] There were actually 4.6% of the offers for $8.00 or more (almost all for $8.00), but we have elected to treat this as approximately zero.

(7) $(7+\frac{\underline{a}+\lambda(.43\overline{a}+.57a_0)}{1+\lambda}3).65\geq 2.80$

■

Proposition 4: In sequential equilibrium $-.301\leq a_0\leq -.095$, $-1<\underline{a}<-0.73$, $0.584\geq\lambda\geq 0.222$.

Proof: We begin by showing that the bounds $-.301\leq a_0\leq -.095$, $-1<\underline{a}<-2/3$, $0.584\geq\lambda\geq 0.222$ hold, then strengthen the bound on $\underline{a}$ as indicated below.

Note that the lower bound on λ follows from the bounds on a_0 and $\underline{a}$, and equation (6) which may be solved for λ as a function of the other two variables. The upper bound on λ follows from substituting (6) into (3) and observing that $\underline{a}\geq -1$.

Note that equation (6) says that a convex combination of $\underline{a},a_0$ is equal to -3/7. This implies immediately $\underline{a}\leq -3/7\leq a_0$. Solving (6) for λ and substituting into the condition (3) that the spiteful type reject the \$6 demand, we find

$$4+\frac{6\underline{a}(\underline{a}+3/7)-6a_0(a_0+3/7)}{\underline{a}-a_0}\leq 0$$

Through straightforward algebraic manipulation, it can be shown that it is possible to satisfy this equation together with $\underline{a}\leq -3/7\leq a_0$ only if $\underline{a}<-2/3$. Observing that if this condition can be satisfied at all, it can be satisfied when $\underline{a}=-1$ then yields the upper bound $a_0\leq -.095$.

Finally, substitute the solution of (6) for λ into (5), the condition that the normal type prefers to demand \$6 rather than \$7. Inspection of the resulting condition shows that if it can be satisfied at all, it can be satisfied when $\overline{a}=1$. Making use of the condition that $\underline{a}<-2/3$ yields the lower bound $-.301\leq a_0$.

Finally we can strengthen the bound on $\underline{a}$ by substituting (6) into (3) to find

$$\underline{a}\leq -\frac{2-(9/7)\lambda}{3(1-\lambda)}$$

Since the right-hand side is decreasing in λ, the largest value of $\underline{a}$ is obtained when λ takes on its smallest value of 0.22 yielding $\underline{a} \le -0.73$.

■

Proposition 5: If $\lambda = 0.45$, $\underline{a} = -0.9$, $a_0 = -0.22$, $\overline{a} = 0.29$[16] and the corresponding probabilities of the spiteful, normal and altruistic groups are $0.20, 0.52, 0.28$ then there are two pooling equilibrium outcomes that are sequential: one in which all demands are \$7.00 and one in which all demands are \$8.00. In both cases, the sequential equilibrium offers are accepted by normal and altruistic types, and rejected by the spiteful types.

Proof: Observe from our previous calculations that normal and altruistic players will accept a \$7.00 demand or less even from a spiteful type; both will strictly prefer to accept a \$6.00 demand. Moreover, a spiteful type, regardless of beliefs about his opponent, faced with acceptance of both normal and altruistic types, will always prefer to demand \$6.00 rather than get \$5.00 for sure. This means that any pooling equilibrium must involve all players demanding at least \$6.00. Next, the population average altruism coefficient is $\hat{a} = -0.21$, so that both the normal and spiteful type will reject pooled demands of \$9.00 and \$10.00, and the spiteful type will reject all demands of more than \$5.00. This enables us to rule out a pooling equilibrium at \$6.00, since the altruistic type will prefer \$5.00 for certain to \$6.00 with 80% chance. We can similarly rule out pooling equilibria of \$9.00 and \$10.00, since the altruistic type will prefer \$5.00 for certain to even \$10.00 received with only a 28% probability. So this further narrows the range of pooling equilibria to \$7.00 and \$8.00 demands, which are accepted by the normal and altruistic type, and rejected by the spiteful type. These are equilibrium with off-equilibrium path play in which all types (including the altruistic type) reject \$10.00 offers

[16] The value of $\overline{a}$ is the estimate from the Centipede experiment discussed below, and is in the range consistent with the separating equilibrium observed in ultimatum.

on the grounds that only spiteful types make them. In this case even the spiteful type prefers to have an $7.00 offer accepted 80% of the time by an altruistic or normal type, to having a $9.00 offer accepted 28% of the time by the altruistic type. Similarly, the altruistic type prefers $7.00 80% of the time against $5.00 for certain. Consequently both of these are equilibria. Notice that like the separating equilibrium, these pooling equilibria satisfy plausible refinements based on monotonicity: higher demands are thought to be made by less altruistic types.

■

References

Andreoni, J. and J. H. Miller [1996]: “Giving According to GARP: An Experimental Study of Rationality and Altruism,” University of Wisconsin, Madison.

Binmore, K. and L. Samuelson [1995]: “Evolutionary Drift and Equilibrium Selection,” University College London.

Cho, I. and D. Kreps [1987]: “Signaling Games and Stable Equilibria,” *Quarterly Journal of Economics*.

Fudenberg, D. and D. K. Levine [1989]: “Reputation and Equilibrium Selection in Games with a Patient Player,” *Econometrica*, 57 (July): 759-778.

Fudenberg, D. and D. K. Levine [1995]: “Consistency and Cautious Fictitious Play,” *Journal of Economic Dynamics and Control,* 19 : 1065-1090.

Fudenberg, D. and D. K. Levine [1997]: “Measuring Subject’s Losses in Experimental Games,” *Quarterly Journal of Economics*, 508-536.

Geanakoplos, J., D. Pearce and E. Stacchetti [1989]: “Psychological Games and Sequential Rationality,” *Games and Economic Behavior*, 1: 60-79.

Isaac, R. M. and J. M. Walker [1988]: “Group size effects in public goods provision: The voluntary contribution mechanism,” *Quarterly Journal of Economics*, 103: 179-200.

Kreps, D. and B. Wilson [1982b]: “Sequential Equilibrium,” *Econometrica*, 50: 863-94.

Kreps, D. and R. Wilson [1982a]: “Reputation and Imperfect Information,” *Journal of Economic Theory*, 50: 253-79.

Ledyard, J. [1995]: “Public Goods: A Survey of Experimental Research,” In *Handbook of Experimental Economics*, Ed. J. Kagel and A. Roth, (Princeton: Princeton University Press).

McKelvey, R. and T. Palfrey [1992]: “An experimental study of the centipede game,” *Econometrica*, 60: 803-836.

McKelvey, R. and T. Palfrey [1995]: “Quantal Response Equilibria for Normal Form Games,” *Games and Economic Behavior*, Forthcoming.

Milgrom, P. and J. Roberts [1982]: “Predation, Reputation and Entry Deterrence,” *Econometrica*, 50: 443-60.

Palfrey, T. R. and J. Prisbrey [1997]: “Anomalous Behavior in Public Goods Experiments: How Much and Why?,” *American Economic Review*, forthcoming.

Prasnikar, V. and A. Roth [1992]: “Considerations of fairness and strategy: experimental data from sequential games,” *Quarterly Journal of Economics*, 107: 865-888.

Rabin, M. [1993]: “Endogenous Preferences in Games,” *American Economic Review*, 83: 1281-1302.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara and S. Zamir [1991]: “Bargaining and market behavior in Jerusalem, Liubljana, Pittsburgh and Tokyo: an experimental study,” *American Economic Review*, 81: 1068-1095.

Van Huyck, J., R. Battalio and F. Rankin [1996]: “On the Evolution of Convention: Evidence from Coordination Games,” Texas A&M.