Begun: Wednesday, December 26, 2007
Presentation Draft version: Monday, May 11, 2009
Comments & suggestions are welcome

PARADOX LOST: Explaining and Modeling Individual Behavior in Social Dilemmas[a]

Joe Oppenheimer *(Corresponding author)*
(301) 405 - 4113; e-mail: joppenheimer@gvpt.umd.edu
&
Stephen Wendel
Both of: Department of Government and Politics
University of Maryland
College Park, MD 20742
swendel@umd.edu
&
Norman Frohlich, Professor Emeritus
I.H. Asper School of Business
University of Manitoba
Winnipeg, Manitoba R3T 5V4 and
Department of Social and Preventive Medicine
University of Montreal
(514) 904 7915; e-mail: frohlic@ms.umanitoba.ca

Key Words: rationality preference self-interest public choice prisoner dilemma game

Abstract: Despite a large body of experimental data demonstrating consistent group outcomes in social dilemmas, a close look at individual behavior at the micro level reveals a dirty little secret. From round to round, individual behavior appears to be almost random. Using a combination of formal deduction and agent based simulations, we argue that any theory of individual choice that accounts for the observed behavior of real people must 1) be premised on probabilistic choice, 2) have context dependent preferences, and 3) replace self-interested preferences with preferences that are, somehow, other-regarding.

---

[a] This paper builds on strands of argument developed in two recent papers: the first by Norman Frohlich and Joe A. Oppenheimer (2006). "Skating on Thin Ice: Cracks in the Public Choice Foundation." Journal of Theoretical Politics, 18(3): 235-266; and the second, by Stephen Wendel and Joe Oppenheimer (2007) "An Agent Based Analysis of Context-Dependent Preferences in Voluntary Contribution Games with Agent-Based Modeling." The latter is available at http://www.bsos.umd.edu/gvpt/oppenheimer/. Earlier drafts were presented at the 2008 Conference on Social Dilemmas at Florida State University, Tallahassee, FLA: February 22-23, 2008; the Conference on Conflict and Complexity. Conflict Research Society and Conflict Analysis Research Centre: University of Kent, Canterbury, UK, Sept. 2-3 2008. 2008; and will be presented at the Conference on Rationality, Behavior and Experiments, State University — Higher School of Economics, Moscow, June 4-6, 2009 and the ESA in Washington, DC , June 25-29, 2009..

Paradox Lost:
Explaining and Modeling Seemingly Random Behavior in Social Dilemmas

Norman Frohlich, Joe Oppenheimer and Stephen Wendel.

Data have been collected from experiments on the theory of collective action for more than a quarter of a century. Applications of the basic theory have cast a good deal of light on a number of collective action problems, spawning a considerable and well known literature. Our attention here is focused on the underlying premise that guides collective action theory: that one can explain group (political) behavior using the workhorse behavioral assumptions of micro-economics, rationality and self-interest. Despite various successes of collective action theory, its group-level analyses are undergirded by data that radically contradicts its theory of individual behavior.

This problem is not new: but it is much bigger than is commonly appreciated in the field. In the applications to non-market behavior, and especially in applications of interest to political scientists, rational choice theory has not predicted accurately at the individual level. Elsewhere, psychologists have chipped away at the simple rationality postulates with experimental evidence of preference reversals and more, using what is loosely known as 'prospect theory' (Kahneman and Tversky, 1979). The fragility of preference stability in the face of framing and other contextual effects is well known. Studies showing precisely this are now legion (see Quattrone and Tversky, 1988 for a fine if somewhat dated, review of the findings here). Preferences have been found to be probabilistic (see Regenwetter et al., 2008), and subject to deep manipulation by framing effects. Other experimentalists have identified limits to the assumption of self-interest. So while the macro program of the theory has appeared to succeed, the micro foundations of this apparent success appears to be problematic.

Frohlich and Oppenheimer (2006) describe how individual contributions to the public good are often inconsistent over time, appearing to fluctuate almost randomly or chaotically across levels over sequential rounds in experiments. Although they conjecture that individuals have complex context-dependent preferences, they neither tested their conjectures, nor did they develop a testable specification of the theory. Wendel and Oppenheimer (2007) showed that one could use agent based simulations to develop plausible specifications for a theory of complex context-dependent preferences that could generate aspects of the

patterns of contributions observed in the experiments. They suggested how analyses of situations using such simulations could help investigate theories that one could then test in further laboratory settings. They show that a substantial improvement in the fit with individual data is possible. This paper builds on these previous two efforts.

Using a combination of formal deduction and agent based simulations, we argue that any theory of individual choice that accounts for the observed behavior of real people must 1) be premised on probabilistic choice, 2) have context dependent preferences, or responses, and 3) replace self-interested preferences with preferences that are, somehow, other-regarding. Further, we establish a feasible model to explain behavior that preserves rational choice with preference theory, which can be expressed in terms of a differentiable, continuous utility function. We conclude that evaluations of political institutions must be reformulated to take into account the new premises. Conclusions of models of social choice and voting based on the traditional theories need to be amended, particularly where the conditions of traditional VCM experiments hold: collective action situations in a novel context with little opportunity for learning or deduction about individual level intentions.[1] But we do not preclude the integration of these models with a learning model that is built on probabilistic choice and other regarding behavior.

We begin by outlining the problem. In Section 2, we describe the general conjectures developed to explain the observations. In Section 3, we develop the model and discuss the results of the simulations. Finally, we discuss the implications of the findings and the procedures for the use of preference theoretic models in political science. In the appendix we extend the 2 person proofs and analysis to 3 and n person cases.

The Problem

The problems with the standard theory of collective action can be depicted quite simply. At the bottom of the theoretical pyramid are a set of heavy duty working premises: Individuals are assumed to be *rational* and

[1] In many contexts, such as markets, competitive elections, military tactics, etc. in which, the traditional model performs admirably.

self-interested with *non-probabilistic preferences* and have *no explicit response to the behavior of others*. Each of these core assumptions has been subjected to significant experimental and theoretical critique. Work to refine the assumption of rationality, defined as having and exercising the ability to choose the best alternative, has extended from bounded rationality (Simon 1947, 1995; Lindblom and Cohen, 1979; Zeckhauser and Shaefer 1968) to broader work on whether *consistent* maximizing, bounded or not, might not be a part of human nature after all (see the good summaries of the findings in Rabin, 1998; Quattrone and Tversky, 1988; Grete and Plott, 1979; Simon, 1986; Tversky and Kahneman, 1986; as well as Shafir and Tversky, 1994). A second premise of most rational choice models, self-interest, appears to be relatively robust in the market contexts to which the 'standard theory' had originally been applied. But as rational choice theory was applied to areas ever more distant from markets, the assumption of self-interest proved more problematic. (See Frohlich and Oppenheimer, 1984, 2000; Fehr and Schmidt, 1999; Cain, 1988; Cox et al., 2001; Ahn et al., 2003; and a variety of dictator experiments. The literature is reviewed in Roth, Alvin, E. 1995.)

These critiques provide some insight into a well studied macro-failure of collective action theory. When Olson (1965) boldly described the collective action problem in rational choice terms, he provided an explanation for why groups of rational self-interested individuals often failed to act to achieve an attainable group benefit. The empirical tests of Olson's insights, first in the field (see for example the discussion in Baumol and Oates, 1979), and subsequently in the laboratories (Ledyard, 1995, is a fine starting point) failed to bear out this drastic, and pessimistic corner result of non-contribution to group goals. In 'voluntary contribution mechanism' experiments (VCM's), which test contributions to public goods in repeated plays, the main reported results found from the experiments can be characterized as

1. significantly higher group contributions than the theory predicts;
2. a decline in group contributions over time (roughly from 40% to 20%)
3. a level that never approaches the corner solution of zero contributions (Ledyard, 1995).

Looking at summaries of group behavior over time gives one hope that the theory is partially supported by the data. Often the analysis stops at that point: research question asked, test designed, completed and reported. But all of the experimentalists have micro data: data about the choices of individual subjects in their

experiments. And at that level, the level the theory is built upon, the theory's performance is abysmal, but this is basically unreported (see Frohlich and Oppenheimer, 2006). Of course, there is much more noise in the data at the individual choice level than at the aggregate group level. And, it is natural that not all of this noise is acknowledged. How that noise may reflect complex individual behavior; and how that behavior might relate to the foundational premises discussed above is the subject of this paper.

With these general musings in mind we begin rethinking the premises of our arguments. We start by examining some data.

Some Preliminary Evidence: Micro-Data and Collective Action

Consider a typical VCM experiment conducted by the authors (e.g., Frohlich and Oppenheimer, 1996; Frohlich and Oppenheimer, 2003), with the parameters and results given in Table 1.[2] The size of the group is 5, and each individual is endowed with a budget of 10 and must decide the proportion to be allocated to a group good that yields a return of 40% to each individual. The decisions are repeated (with the same group) 7 or 15 times (depending upon the particular experiment).[3] The standard theoretical prediction from the theory of collective action, is that individuals will choose their dominant strategies: that of *giving nothing* (top row in the table).

{Table 1 About Here}

If individuals behaved according to standard assumptions of rationality and pure self-interest, their motivation to contribute can be formally described as in the following equation:

$$U_i(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_{\sim i,t} \qquad \textit{Self interested utility function } \textbf{(1)}$$

In this series of experiments, each individual (i) has an endowment of 10, and is trying to maximize the function $U_i(X_{i,t})$ by determining $X_{i,t}$, the level of the contributions in the current, t^{th} round. $Y_{\sim i,t}$ is the average

[2] T.K. Ahn has helped pool together the data from many of the VCM experiments and very kindly made this available to us. We have examined all of this data. All of our observations that we report below regarding our data hold in the larger set.

[3] The subjects don't know when the experiment will end.

contribution level of the other four players.[4] The utility maximizing level of this function is straightforward: $X_{i,t}$ equals 0 for all players and all rounds.

Non-cooperative game theory indicates that other equilibria can be sustained (i.e., the folk theorem), but at least when the experiment is designed so that individual contributions aren't identifiable, the non-cooperative Nash outcome of all playing their dominant strategy is most compelling on theoretical grounds. Indeed, it is the standard prediction for models of n person PD games. And as noted above, a standard experimental result is getting higher levels of contributions, which decay somewhat over time. Without some further 'bridge principles' (Hempel, 1965) the theory is in trouble (Green and Shapiro, 1994).

Existing Micro-level Hypotheses for Aggregate Responses

Varying hypotheses have been put forward to explain individual-level contributions to public goods. In an earlier work (Frohlich and Oppenheimer, 2006) we analyzed this literature and considered the logical implications of alternative theories. Here, a brief summary should suffice to provide context. Some say there are two or more types of individuals: some give and some don't. Is it that, for example, 70% say, "Not me! I'll be no one's fool" and 20% say, "Yes, count me in!" And maybe some individuals (e.g. 10%) slowly learn and hence move from the count me in to the not me posture? Perhaps there are just these two basic types and the learners: the rational egoists - equilibrium playing self-interested responses and the unconditional cooperators. Such a conjecture can be tested: both of these sorts would have a response line over the rounds that would be horizontal or constant. And the learners should gradually shift down over the course of the experiment. Others say that those who don't cooperate, and those who shift to non-cooperation, leave a 'meaningless' tip, of say 15% so as to not feel selfish. This too can, and has been, tested. In this case, the learning would leave a curve asymptotic to 15% rather than 0%. Some of the conjectured types of contributors are illustrated in Figure 1.

{Figure 1 About Here}

[4] Individuals were informed of the average contributions of the others at the end of each round of decision making.

Other individual-level behavioral patterns have been conjectured to generate these aggregate results. There may be conditional cooperators as conjectured by Rabin (1993) and Cain (1998) - individuals who are learning, or trying to get something going by signaling, or trying to contribute when it could make a difference (and their lines of behavior might look quite different in this graphical space). Perhaps some may even be trying to react to others' non-cooperativeness: attempting a punishment by withholding contributions, although in this simple VCM design there is no way of directly targeting punishment at a particular individual (see Ostrom et al. 1991 and 1992).

The graph in Figure 2 displays some typical results for five groups (aggregated) over the 15 rounds of repetitions, along with a prediction of learning behavior. The example from an actual data set (blue line) does not conform to the pure self-interest prediction (which would be a straight line along the X axis) but does come quite close to the conjectured learning (represented by the red line) and is typical of results in experiments of this type. So what does this mean for the various conjectures?

{Figure 2 About Here}

Micro-level Hypotheses for Individual Responses

While the micro-level hypotheses could have produced observed aggregate behavior reasonably well, they flatly fail to conform to observed individual level behavior. A cursory examination of the individual level data shows that *no one* is an unconditional cooperator (at any level, including the notion of a 'steady tipper'). Very few (fewer than 10%) are purely self interested and they are the only ones who appear to have unconditional behavior. Figure 3 presents sample individual level data. The problem is that behavior of virtually everyone is erratic. We can note that there are 1) no other simple non-conditional players (that always give a set amount) other than those that give 0; 2) looking at the 3 representative graphs in Figure 3 the patterns of behavior do not appear to be easily decipherable into a simple aphorism, such as either reflecting 'learning' or conditionalized behavioral responses to the behavior of others. Indeed, we could have put up thousands more such graphs: all showing jagged responses, but these illustrate the problem: the individual behavior appears to be random or chaotic.

{Figure 3 About Here}

If most everyone is behaving erratically, what is driving their behavior? Many conjectures are possible. Inconsistency can stem from erratic learning, complex interactive cuing among the subjects in the experiments, general instability of preferences, a lack of care in the decision making, or some other violations of rationality and/or self-interest. Perhaps Charles Plott's observation is appropriate "things are just *too* complicated at the level of the individual."[5] We clearly have behavior that is varying in a seemingly erratic, if not random fashion: perhaps it is responsive to the ever changing decision environment facing the individual.

How can we explain such fluctuations and inconsistency? Do we have to give up the rational choice assumptions? Can we explain it as a reflection of the responses to the behavior of others in the group? And if so, what is that response function like? In previous research (Frohlich and Oppenheimer 1996; 2006), we considered these larger issues of rational choice theory and individual level behavior, and performed statistical analyses on the experimental data against existing theories of individual-level behavior. We looked at the effects of conditionalized behavior such as signaling and negative reactions to others' low contribution rates statistically, finding that self-reported concerns for both *equity* and *group benefit* are strong predictors of (higher) contribution levels in these VCM experiments. Those two elements of other regarding behavior explained more than 40% of the variance in contributions on their own (Frohlich and Oppenheimer, 1996). Those findings are supported by authors such as Gunnthorsdottir, 2001, who also find differences in personality, or values, between free-riders and cooperators. We concluded that line of argument with a discussion of how such context dependent behavior provided a promising route of analysis at the individual level. However, we developed no specification of what such a model would entail. Here, we apply those considerations to a set of specific models discovered through a process of simulation-enabled theory development (Wendel and Oppenheimer, 2007).

---

[5] In personal conversation in with Professor Plott, at an NSF funded conference (at the University of Indiana, in January, 2003). The NSF Conference, featuring attendees including John Ledyard, Elinor Ostrom, and John Walker, was organized to analyze the unexplained individual-level behavioral patterns underlying this paper.

The Conjectures

How to generate erratic responses, when the basic environment (as in 'The Experiment') is a constant, is the central question at hand. Clearly, if the response is 'rational' (i.e. reflective of preferences held by the individual) then something must be triggering the changes. The obvious suspect is the behavior of others.[6] In other words, we begin by assuming a reactive utility function that incorporates the behavior of others in one's evaluation of one's own behavior. One problem with this sort of answer is: how to specify the relationship between choice and preferences. To assist us we combine traditional mathematical deduction with agent based modeling (initially explored in Wendel and Oppenheimer, 2007) to grapple with the complex responsive, iterative, behavior.

Other-regarding Behavior

We begin by making more specific the decision environment faced by the individual. Equation 1 reflected the experimental setting as viewed through the traditional lens of rational self-interested behavior. It did not encompass possible group dynamics. In the laboratory, after the first round, each individual is faced with a decision environment that contains actual people and each has to respond to the decisions of the other individuals. Specifically, in our experiments, as in most other VCM experiments, each subject knew the aggregate behavior of the others in the group after each round, prior to making a decision for their contribution in the subsequent round. Presumably this continually changing situation (information about the aggregate behavior of others and of one's own historical behavior) leads to changing responses. Of course, this wouldn't happen if the individual were motivated by self-interest only, for then she would just never contribute, regardless of the behavior of others. Similarly, an individual who contributes, regardless of the behavior of others, will not be affected. But, as we indicated above, there are no unconditional givers and there are mighty few purely self-interested types out there.

---

[6] To keep the story manageable, and as a first approximation we will presume the individual always keeps in mind only a one-round history.

INTERACTION OF SELF WITH OTHERS: To make sense of and to model these patterns of reaction, we need a few more variables. After round *t -1*, individual *i* makes a decision about what to contribute ($X_{i,t}$) given the average donations of others ($Y_{\sim i,t-1}$) (see footnote 4), and herself ($X_{i,t-1}$). Individuals compare their effort with that of others in the group. In this comparison they can discover that their contribution is greater, equal to, or smaller than the average contribution of others. Finding themselves giving more than others something like a sense of justice or fairness (as evidenced in Frohlich and Oppenheimer, 1996) may well be expected to lead to a feeling of frustration, or anger, that others aren't sharing the burden proportionately. Similarly, finding herself relatively less generous than their fellows can lead the individual to feel guilty regarding her smaller contribution. There are many conjectures already in the literature regarding how one responds to the behavior of others. These can be expressed generically without being precise about their content. Label the current round as *t*, and the previous as *t-1*. Then we conjecture that the individual's current payoff calculations are a function, *f*, of both her previous behavior and the previous behaviors of others and that these behaviors impact on the player's current contribution level.

$$U_i(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_t + f(X_{i,t-1}, Y_{t-1})* X_{i,t} \qquad \textit{other-regarding utility function } \textbf{(2a)}$$

What are reasonable properties of this function *f*? Most experimentalists (see Fehr and Schmidt, 1999; Janssen and Ahn, 2006; Cox et al., 2001; and Charness-Rabin, 2002 for a few examples) agree that people want to be treated equitably. They get more upset (i.e. they care more) about being wrongly taken advantage of (i.e. when $X_{i,t-1} > Y_{t-1}$ than they care about being wrongly advantaged (when $X_{i,t-1} < Y_{t-1}$). In the VCM decision situations, this would mean that we want *f* to take on values that reflect the difference in the giving patterns of the group and the individual or $X_{t-1} - Y_{t-1}$. The effect of *f* should be larger when X > Y than when Y > X. Further, we would want the response to monotonically increase as the difference between X and Y increase. Although there are many possible ways of getting these properties, below, we accomplish them by letting *f* be a function ($Y^s - X^s$) when Y > X, and by ($X^{s*} - Y^{s*}$) when X > Y, presuming that $s^* \geq s$, and specifying that $s > 1$.

ALTRUISM: But why would other-regarding behavior only exhibit itself in response to the relative giving of self and others? If one relates to the group and its behavior, might not the individual care about the quality of

the group outcome for all the participants, rather than just for herself? That is, a second form of other-regarding behavior, we here call altruism should be considered.[7] This was also evidenced in Frohlich and Oppenheimer, 1996. Since any contribution by the individual helps the outcome for the group, we can capture altruism with an additional function that reflects an additional utility going to *i* for contributions which increase the others welfare: $g(X_i)$.[8] Here *g* would be increasing in X_i. Adding this term for altruism to the equation for other-regarding behavior gives us:

$U_i(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_t + g(X_{i,t}) + f(X_{i,t-1}, Y_{t-1})*X_{i,t}$ *Altruistic & other-regarding utility function* **(2b)**

Although these changes could lead to less constant behavior, none of this is likely to lead to the typically erratic, jagged type of responses displayed in Figures 3. In our current abstraction, everyone has an incentive to move toward the average response in the first round, although to different degrees. Hence one would expect a relatively quick dampening of the erratic behavior. We are not likely to see the sorts of continual flailing about in response that are reflected in the graphs of the behaviors of actual experimental subjects.

Probabilistic Response

The notion that one has a specific value for different outcomes is a basic presumption of traditional utility theory. Traditionally, theorists developed a non-probabilistic theory of choice: if *i* chooses *a* over *b*, it is because aP_ib. Such a 'certain' tie of choice to preference leads to knife edge results. When *b* becomes

---

[7] Altruism is defined as placing a (positive) value on the welfare of others, even when there is no direct feed back to you of that welfare. There is a big literature now on the nature and form of altruism. It is a topic Frohlich and Oppenheimer have worked on for years, beginning with Frohlich 1974; followed up by Frohlich, et al. 1984, 2001, and 2004. Others have worked even earlier on these problems. See, for example Valavanis, 1958.

[8] Expressing altruism this way avoids a number of nuances. For example, if one only cares about the quality of the group outcome, rather than one's own part in it, then the donations of others can be a perfect substitute for one's own contribution. Caring about one's own giving is often attributed to a 'warm glow effect.' For purposes of simplicity, we don't bother with such distinctions here.

preferred to *a*, behavior shifts completely. The preferences of individuals are presumed to be stable, and unique.

But modern theories of memory and recall support a probabilistic theory of choice. These theories attribute storage of an experience to memory to the value that the experience has. One stores, and then can remember things precisely, because they have some value (Edelman, 1992). Most items are remembered in multiple contexts, and hence with differing values. And as one recalls an item, one brings up the attached memory and values. As Damasio (1999, 163-64) puts it:

> . . (T)he memory of . . . (an) object has been stored in a dispositional form. Dispositions are records which are dormant and implicit rather than active and explicit, as images are. Those dispositional memories of an object that was once actually perceived include not only records of the sensory aspects of the object, such as the color, shape, or sound, but also records of the motor adjustments that necessarily accompanied the gathering of the sensory signals; moreover the memories also contain records of the obligate emotional reaction to the object. As a consequence, when we recall an object . . . we recall not just sensory characteristics of an actual object but the past **reactions** (emphasis added) of the organism to that object.

So the valuations have a distribution, and are context dependent in complex ways. Such complex valuations as are embedded in these compound memories involve non unique elements, and what is recalled by any trigger is likely to be probabilistic in details and valuation. And we argue, that to avoid clearly unsubstantiated predictions of individual behavior such as smooth learning, early convergence, or purely oscillating response functions in these voluntary contribution mechanism experiments, it appears that we need probabilistic response, and hence probabilistic choice, functions. It may not be that god plays dice but we do. Equation 2c illustrates our belief that the response function relating to a sense of justice or unfairness is the element most likely to be subject to probabilistic fluctuation. After all, it is inherently reactive to the relative behavior of others. That behavior is subject to the greatest ambiguity. We presume that the underlying valuations of the monetary returns and of basic altruism are not subject to probabilistic fluctuation.

$$(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_t + g(X_{i,t}) + \rho\,(X_{i,t-1}, Y_{t-1})*f(X_{i,t-1}, Y_{t-1})*X_{i,t} \quad \textit{Altruistic \& other-regarding utility} \quad \textbf{(2c)}$$

Non-linear Valuations

If we are to identify optimizing behavior on the part of individuals that is responsive to the factors in question, and not simply a corner solution of contributing all or nothing, we must introduce some non-linearities into the utility function. One traditional way of doing this in public goods situations is to introduce a squared term in the valuation function of the public good.[9] This is not plausible in an experimental context in which the public good is actual monetary payoffs over a relatively small range. However, Kahneman and Tversky (1979) have shown that losses, even small ones from a status quo, are felt more strongly than could be represented by a linear utility function. Therefore, to model contributing behavior that is not restricted to corner solutions, we introduce a quadratic loss term in the contributions.

$(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_t + g(X_{i,t}) + \rho(X_{i,t-1}, Y_{t-1})*f(X_{i,t-1}, Y_{t-1}) *X_{i,t} - h(X^2)$ *Altruistic & other-regarding utility function* **(2d)**

The Model

We now put flesh on these presumptions. Begin with the model minus any probabilistic term. Optimization takes place after it has been determined whether how one's contribution compares with the average of others in the previous round. Using the parameters from 'The Experiment' it is captured in equation 3.

To discount the response to guilt (when Y>X), relative to anger (when X>Y), we use a parameter, b <1 in the utility function *f* . To have the utility function vary so as not to produce corner solutions, we add a final term $(-(X_{i,t})^2/2)$ to represent the intensity of the loss associated with contributing X.

Then, the value to *i* of contributing X is different when *i* contributes more than the average than it is when *i* contributes less. In the latter case, when others gave more than the individual, the equation becomes:

$U_i(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_t + a_i(X_{i,t}) + b[(Y_{t-1})^s - (X_{i,t-1})^s]*X_{i,t} - (X_{i,t})^2/2$ *The value of giving when Y>X* **(3)**

i will choose an amount to give that will maximize her payoff given the valuation of $U_i(X_{i,t})$. That is, we can identify this maximizing contribution by simple differential calculus. Using the first derivative, to solve for an optimal X, we get:

[9] For a typical recent example, see Arifovic and Ledyard, 2008.

$$X_{i,t}^{opt} = -.6 + b[(Y_{t-1})^s - (X_{i,t-1})^s] + a_i \qquad \textit{The optimal value to give when } Y>X \quad \textbf{(4)}$$

And when others have given less than *i*, this yields

$$U_i(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_t + a_i(X_{i,t}) + [(X_{i,t-1})^{s^*} - (Y_{t-1})^{s^*}]X_{i,t} - (X_{i,t})^2/2 \qquad \textit{The value of giving when } X>Y \quad \textbf{(5)}$$

The optimal X then is determined by:

$$X_{i,t}^{opt} = -.6 + (X_{i,t-1})^{s^*} - (Y_{t-1})^{s^*} + a_i \qquad \textit{The optimal value to give when } X>Y \quad \textbf{(6)}$$

Equations 4 and 6 provide our benchmark model for non-probabilistic choice, which, because of mathematical and conceptual simplicity, we will analyze along side the more sophisticated probabilistic choice model.

To deal with probabilistic choice, we need to discuss the shape of function ρ that determines the probability. We believe that the perturbations are coming about from the reaction to the difference between one's own and others' behavior. That means, we need to map the probability of response changes (ρ) to the changing differences in behavior. Obviously, we want ρ to be monotonically related to the differences in behavior: the bigger the gap, the more certain the response. In other words, we can write the probabilistic response function as:

$$U_i(X_{i,t}) = 10 - .6X_{i,t} + 1.6Y_t + a_i(X_{i,t}) + \rho(X_{i,t-1}^{s^*} - Y_{t-1}^{s^*})*X_{i,t} - (X_{i,t})^2/2 \quad \textit{The probabilistic value of giving when } X>Y \quad \textbf{(7)}$$

But even then, we need to posit a shape for the probabilistic function. Besides having a positive slope as a function of X-Y (monotinicity), what else might we consider? We have made the judgment that our first best guess is that the response function is likely to be a sigmoid function (i.e. S shaped: as in Figure 4). Such functions are rife in many 'natural' events, and since we believe we are dealing with a 'natural response' chose this form as a reasonable starting point. The functional form of a sigmoid curve responding to the difference between X and Y is:

$$\rho = -h/(1+e^{[(-|X-Y|)/k]}) \qquad \textit{A Sigmoid response function} \quad \textbf{(8)}$$

{Figure 4 About Here: Sigmoid Curve}

In the graph, h is -1 and k is one. We want the value of ρ to be zero when there is no difference between the individual and the average group behavior, so we need to 'center' the function around zero which is done

by adding ½ to the function. A larger h would give a bigger 'height', and a larger k would give a bigger 'stretch.' For example, if h=-2, the function goes from -2 to 0. And with k=1, rather than 3, as in the diagram, the curve would acquire a range of -5 to + 5 to retain its current form. We have set k = 3 so that the response is close to a constant after the edge of the feasible range is hit. In other words we have to set the valuations of some of these variables to insure that the curve generates responses in keeping with the nature of the decision problem the individuals confronted. With this function, it should be noted, there is no difference between the anger and the guilt in the inequality aversion portion of the utility function. That asymmetry would here be taken into account by the 'b' term in equations 3 and 4.[10] Setting 'b' so that anger is about 6 times stronger than guilt as a motivator, generates the results that we show below but sensitivity tests were run that show that over a wide range of values, the specific value of 'b' makes little difference to the results. On the other hand, the model is now close to intractable analytically. To handle the complications that have been added requires an increase in calculating power: one achievable by agent based modeling.

Results

Our analysis proceeds in two parts. First, we consider the logical properties of the non-probabilistic utility function, under the analytically tractable circumstance of two competing players. Second, we delve into the full system of five players, under the deterministic and probabilistic model variants, with a set of simulation models. Finally, the simulation models allow us to explore alternative specifications of the model for rival hypotheses (Wendel and Oppenheimer, 2007). Both the formal and the simulation analyses, follow directly from the VCM lab experiments.

Formal Analysis:

[10] Of course, a slightly different psychological assumption would substitute a 'b' in the top or bottom portion of the sigmoid function so as to distinguish the probabilistic response to the behaviors, rather than the valuation itself. The literature on inequality aversion, however, suggests that it is the valuation itself that changes (Fehr and Schmidt, 1999).

First, we examine the non-stochastic, individual optimal contribution function from equations 4 and 6. We demonstrate that the entire family of related contribution functions, covering both our own model and numerous others conjectured in the literature, will necessarily settle down either into periodic oscillation or to a convergence to a constant level of aggregate giving. Since neither of those conforms to the observed behaviors, we reject non-stochastic individual behavior as an assumption.

To establish this result of either periodic oscillation or convergence, we assume that the experimental subjects knew neither the identity of the other participants nor their individual behavior. They knew only the aggregate behavior of the others in the group after each round.

To show this, start with the optimal contribution functions, Equation 4 and 6, provide further generalization into a wider range of VCM games, and reiterate the minimal assumptions made in the text. First, the linear rate of return on any contribution, rather than being .4, is specified as k. As above, the individual's current payoff calculations are a function, *f*, of the previous behavior of both her and others and are potentially affected by altruism: a_i. $Y_{i,t}$ is still the average of the contributions of others at time t. In other words with the utility function for any giving of $X_{i,t}$ with non-stochastic behavior and other-regarding preferences would be a function of altruism, the difference between the subject's and the average of others behaviors, the rate of return, and the contribution made. With these generalizations on the optimal contribution function, we have:

$$X_{i,t}{}^{opt} = (k\text{-}1) + f\,[(Y_{i,t-1})^s - (X_{i,t-1})^s] + a_i \qquad \textit{The optimal value } (Y>X \textit{ or } X\geq Y) \textbf{ (9)}$$

Consider the beginning of the interaction. All the participants have no behavior of the others to respond to. We assume that in such situations, they make contributions as if there were no differences between their behavior and the behavior of others: (letting $d_{i,t} = (Y_{i,t-1})^s - (X_{i,t-1})^s$) as if $d_i = 0$. Call this contribution level of the individual, $(k\text{-}1) + f\,[0] + a_i$, that individual's *neutral contribution level*. We can note this as N_i. Define function r, the *reaction function*, as $f\,[d_{i,t}]\text{-}f[0]$. Rewriting equation 9 using this notation we have:

$$X_{i,t}{}^{opt} = N_i[k, a_i] + r[d_{i,t}] \qquad \textit{Optimal contribution function, in terms of neutral contribution level } \textbf{(10)}$$

We will want to begin each player with this *neutral contribution level*. Of course, not all individuals will have the same N_i. So there will normally be a gap between neutral contribution levels of any participant *i* and the

average of the others participating. Call this gap $G = N_i - N_o$, where N_o is the average neutral contribution level of others.

Now, as above, we assume both that $\partial X_{i,t}^{opt} / \partial a_i > 0$ and that $\partial X_{i,t}^{opt} / \partial d_{i,t} < 0$: that is, one's response to doing 'differently than others' is to change the level of X upward if one gave less than others and downward if one gave more.

Assume that the individual, *i*, begins each round with "N_i" and then responds with a discounted value of the difference he found between his behavior and the average of others that he has observed in the previous round. We will discuss the situation for two, and then 3, and then n person VCM games are given in the Appendix. Since the participants in the experiments only knew the aggregate behavior of the others in the group after each round, the logic of the 2 person case will be somewhat useful in understanding the other cases.

TWO PERSON CASE: We can show that the general solution of optimizing behavior under these simple assumptions, for a two-person game, will always either converge or oscillate. This will be so regardless of each player's response to differences in contributions (as long as $\partial X_{i,t}^{opt} / \partial d_{i,t} < 0$), of each player's a_i, and regardless of each player's initial contribution levels. First note, with two players, *i* and *j*, $G_i = - G_j$. The following table illustrates what happens if the response by the individual to the gap is linear, say the players respond to a gap with some proportional constant r (0>r>1):

{Table 2 about here}

More generally, using G as the gap as experienced by *i* (that is $G = N_i - N_j$) in the first round, we can call the difference in the behavior between the two players as in any round t^* ($t^* > 1$) d_{t^*}. Then the behavior of individual *i* in the next round, t+1, is simply $N_i - r^* d_{t^*}$. Similarly, *j* will respond with $N_j + r^* d_{t^*}$.[11] This then lets us express d_{t^*} as follows:

$$d_{t^*} = G(1 + \sum_{t=2}^{t^*} (-2r)^{t-1}).$$

[11] Note one peculiarity of these assumptions is that the sum of the donations remains the same.

It follows then that what happens to d (the difference) will depend upon what is happening with $(-2r)^{t-1}$ as t increases. Recall that r is presumed to be >0 and <1. Hence there are three cases of concern:

1. When r is precisely .5, then $(-2r)^{t-2}$ will oscillate between +1 and -1 depending upon whether t - 2 is positive or negative. Hence the sum of these strings over intervals of t will oscillate between 0 and -1 and the difference, d, will also oscillate between 0 and G.
2. When r is less than half, then $(-2r)^{t-2}$ will getting geometrically smaller as *t* gets larger. So we will be subtracting from G a sum that is growing quite quickly, but by less and less. The value of the difference, d, will hence converge to some value between the two values 0 and G. The closer r is to ½ the closer the convergence value will be to ½ G. When r is smaller, that convergence value goes up toward G and the convergence is quicker.
3. When r is more than one half, then $(-2r)^{t-2}$ will grow as *t* gets larger. And the sum of these will also grow, so that the difference, d, will be oscillating increasingly, until we reach the maximums possible (in this case 0 and E).

Similarly, we can express the behavior of individual i in any round t* (t* >2) as:

$$X_{i,t}{}^{opt} = N_i - Gr(1 + \sum_{t=3}^{t^*} (-2r)^{t-2}).$$

Here again, it is the performance of $(-2r)^{t-2}$ that will determine what is going on. The same properties hold, although now, to show what the values will be as a function of N, G, and r, it is obviously slightly modified. Thus, we have the same story. The contributions will oscillate N and a lower number, when r = .5. Overall:

1. If r > ½, the two players will oscillate with increasingly extreme contributions, until oscillating between maximum and minimum levels allowable given their assets. G and r will both determine the speed of reaching those limits.
2. If r < ½, then $(-2r)^{t-2}$ will getting geometrically smaller as t gets larger. So we will be subtracting a sum from N that is growing quite quickly, but by less and less. The value of the difference will hence converge to some value between the two neutral contribution levels. More specifically, the two

contributions of each of the players will each steadily converge back toward a value of between their neutral contribution levels ($N_{i,j}$). G determines the size of the oscillations. 'r' determines speed of the convergence: the closer r is to .5 the longer it takes for the convergence to occur.

3. Finally, if r = ½, the two players will continually oscillate between their neutral contribution levels and a level that is N - ½ G. In other words, the oscillations will be between N_i and the average of N_i and N_j. Hence, if the other person has a higher N, the gap is negative, and the oscillation will be between N_i and ½ the way toward N_j.

While started with optimal contribution functions 4 and 6, it should be noted that any function of the form given in equation 10 will provide similar results.

These results clearly indicate that a general family of equations, which follows directly from our own prior research and that of numerous other researchers in the field, is unable of generating the individual-level behavior observed in the labs, at least in the 2-player case. Moving beyond two players analytically is much more challenging. In the Appendix, we provide a similar but more detailed proof covering the n-person case. The results are analogous: either oscillating or converging rapidly. In the next section, we pursue another route of analysis, directly simulating the 3+ person case in an agent-based model under a suite of reasonable parameter sets; as expected, oscillation or convergence ensure. We then move on to the probabilistic model, which is intractable for analytic solutions, but amenable to simulation methods.

***Simulations*:**

First, we programmed simulated agents to follow (exactly) the conjectured non-stochastic behavior provided in equations 3 and 5. The agents are embedded in an environment that mimics the experimental condition, in which a group of participants, i.e., agents, are each given their endowment at the beginning of each round, and then are allowed to contribute some or all of it to the production of a public good. Conceptually, the simulations consist of three parts: the VCM parameters (as used in the experimental studies and more specifically, starting with those displayed in Table 1), the motivational structure of the individuals participating in the iterated VCM game (as provided by equations 4-8), in which each player learns of the average contributions of the others from the previous round. The simulation software also includes a set of

tools to visualize and record data from the game, and to explore the sensitivity of the model to its parameters.[12] The simulation design provided the flexibility to simulate a wide range of alternative theories that might generate jagged contributions.[13] Here we only focus on the components used in the theoretical models given above. Other components used in sensitivity analyses will be discussed but not detailed here (but see Wendel and Oppenheimer, 2007).

In round 1, contribution levels are set by us. We stochastically select starting values to both reflect typical patterns of human subjects and adjust agents' altruism patterns to reflect this initial move. At the start of each subsequent round, each agent receives information about the average contribution of the other agents in the prior round, if any. The agents then maximize their utility (i.e. the equation being tested) over the range of potential contributions via a simple numerical approximation: one hundred evenly spaced contribution increments are evaluated in each utility function, and the contribution that generates the highest utility level is chosen. All results have been archived.

In the stochastic models, individual runs are inconclusive on their own. As it would be tedious to execute the software manually a sufficient number of times to generate meaningful results, we executed the simulation in batch mode across one hundred unique random number seeds. The results were logged to a file and analyzed in the statistical package R.[14]

Sample results from the deterministic utility function agents are given in Figures 5-7; each of these graphs display behavior by 5 (simulated) individuals, or agents, making decisions in 30 consecutive rounds under the

---

[12] The model was built upon the RePast Simphony agent-based modeling platform (Tatara et al. 2006), with subsequent elements programmed in Java. RePast provides helpful visualization and data logging tools.

[13] The previous paper (Wendel and Oppenheimer, 2007) gives the details of the alternatives that were considered as well as much of the logic for the rejection of many of the alternatives.

[14] This design is inspired by Catherine Dibble's concept of a "Computational Laboratory", which employs an agent-based model at the center of a larger process of sensitivity testing, optimization, and statistical analysis (2006).

conditions specified in Table 1. They maximize their utility function as specified in equations 3 and 5. Each agent is given a unique level of altruism that then generates a randomly set starting level of contributing. As expected from our analytical solutions, we can get either oscillations or rapid convergence The selection of graphs here is made to show typical patterns; indeed, all agents under these circumstances generate simple oscillation or convergence to an equilibrium level of contribution.

{Figures 5-7 About Here}

Without a probabilistic response, we thus get early convergence or continued periodic oscillations: simulated behavior that does not mirror the patterns of the data. No real live subjects exhibit pure oscillation and very few even approximate oscillation. Further, none of the subjects we have examined have exhibited behavior that converged to any value other than zero. And not very many do that, either. A final problem we have with the non-probabilistic models is a lack of general decline in average contributions over time. Although there are circumstances where this happens, it is not a 'typical' event in the simulations; it is typical for human subjects (compare Figure 2 with the current Figure 8). In Figure 8, we display overlay the aggregate contributions of 8 separate groups of simulated individuals over time. Note that half of them oscillate periodically, and the other 4 quickly converge to a non-zero sum. The lack of decline takes place even though there is a ratio of 6 fold in the anger versus guilt response.

{Figure 8 About Here}

Adding a probabilistic response component provided in equation 8 above significantly changes the outcome. Instead of oscillation or rapid convergence, we find patterns that more adequately reflect both the almost chaotic responses which were displayed in the behavioral graphs earlier (see Figure 3) and also the decreasing total contributions by group members (Figure 2). These two features of both agent (Figure 10) and aggregate (Figure 9) behavior are shown in the graphs below. Both were selected for readability (when the lines are too similar and overlap it can be hard to trace behavior) and are typical in patterning. Figure 10 shows the 5 agents of one group, with all the properties of Equations 7 (the utility function) and 8 (the probabilistic response function). The very erratic behavior patterns of the real subjects now shows up as maximizing behavior with probabilistic response. And Figure 9 has the properties of decreasing donations

over time, not declining to zero just as in the laboratory experiments.[15] Earlier model explorations (Wendel and Oppenheimer 2007) included sensitivity analyses showing that these results are robust to wide variations in their parameters, including the initial contribution levels of agents, the weight and distribution of the other-regarding component of the utility function, and the injection of individuals motivated by alternate utility functions (e.g., purely self-interested individuals). Thus, while our analytical results thus indicate that deterministic preferences are logically insufficient to generate the observed pattern of behavior, the simulation experiments demonstrate that the theoretically appealing option of probabilistic, other-regarding preferences can, under the experimental conditions, generate the observed behavior.

{Figures 9 and 10 About Here}

Overview and Discussion

The results support a number of substantive conclusions that we, and others, have articulated for more than a decade concerning the lack of realism of the rational choice assumptions for the theorizing of public and social choice modeling of politically interesting phenomena.[16] It appears that the presumption that each individual has preferences separated from the welfare of others is wrong: it doesn't enable one to explain the facts. Second, the notion that one's behavior is deterministically set by one's preferences appears wrong. The results cast severe doubt on both of these traditional axioms of rational choice theory. Indeed, the basic results (not the particulars of their functional forms) of Fehr and Schmidt (1999) and their conclusions regarding the weight of anger and guilt seem strengthened by our analysis. We care substantially more that others do as much as we do, than we do about their doing more than we do.

Of course, the lack of fit of the most basic experimental data with the classic theory, as outlined in the first part of the paper, is a fundamental problem. Putting the 'excusniks' (e.g. Milton Friedman, 1953) aside,

---

[15] For each of the simulation models we ran, we did sensitivity testing over a range of parameters. Most of this is outlined, with detailed results in Wendel and Oppenheimer (2007). There we report results for groups of various compositions, various parameters values, etc.

[16] Again, the methodological implications are discussed more fully in Wendel and Oppenheimer (2007).

the severe lack of fit between reality and the quantifiable assumptions of a theory are always of concern in science. Realism must play a role somewhere in the path that scientific inquiry carves out in its intellectual journey (see Nola, 2004). Presumably the call of realism has been resisted because dealing with framing, probabilistic choice, and other-regarding preferences were thought to add too many complications to the calculations. It was impractical to reconstruct the theory to conform to reality.

But what has been established here is that relatively simple fixes to the foundations can go a long way to repairing the structure, without giving up the calculating capability of the models. Others have argued similarly (see, for example, Fehr and Schmidt, 1999; Rabin, 1993; Mackie, 2003; Ahn, et al. 2003; Frohlich, Oppenheimer and Kurki, 2004) in more limited fashions. We have added three interactive elements: a notion of altruism (or concern for the group outcome), a notion of fairness toward self and the group, and a probabilistic element tying behavior to values for fairness, evaluated in terms of relative effort. We have shown that incorporating the complexity of reality into the analysis can be done with little cost to calculability and yet can generate great improvement in the fit of the models to the non-market behavior at the individual level.

Few have argued that the normative prescriptions based on ill-fitting theories should be questioned and that there is a need to amend the foundational premises that yield the empirical anomalies (but see Mackie, 2003; and Green and Shapiro, 1994 who led the charge). However, the reader should note that the wrinkles we have added are likely to change some of the normative implications of some of the models based on the older, basic rationality and self-interest assumptions. Just as an example, consider the implications of probabilistic choice to spatial modeling of electoral competition. Coughlin (1984) discusses the implications of probabilistic voting for political candidates.[17] The major change in analysis is that electoral competition is

[17] It must be said, however, that Coughlin indicates that the voter's decision isn't probabilistic, only that the politician presumes it is because of incomplete information (told to Oppenheimer in repeated conversations). Be that as it may, the implications that follow from his analysis hold true for probabilistic choice by the individual voter.

far more likely to lead to a 'median voter' result, and the result has easily interpretable analogues to the normative theory of utilitarianism (see Mueller, 2003, chapter 12). Indeed, the electoral analysis results are far more optimistic than the traditional arguments that show the possibilities of cycles, and hence little normative interpret ability of electoral outcomes.[18]

We are certainly not asserting that the model we put forward and tested is the stopping point for development of an adequate behavioral theory. Rather, we view it as a starting point: one that shows the elements that must be incorporated in any further theoretical work that is to be grounded in reality. Others have already begun such work, and will certainly lead to a new foundation.[19]

As always in a productive moment for science, what is clear is not where we will end up, but the road that must be taken.

Bibliography

Ahn, T. K., Elinor Ostrom, and James M. Walker (2003) "Incorporating Motivational Heterogeneity into Game Theoretic Models of Collective Action" in Public Choice, V. 117, No. 3-4 (December.):

Arifovic, Jasmina & John Ledyard (2008). A Behavioral Model for Mechanism Design: Individual Evolutionary Learning. Mimeo, working paper (Caltech).

Arifovicy, Jasmina & John Ledyard (2009). Individual Evolutionary Learning, Other-regarding Preferences, and the Voluntary Contributions Mechanism. Mimeo, working paper (Caltech).

Baumol, William J. and Wallace E. Oates (1979), Economics Environmental Policy and the Quality of Life. Prentice Hall, Englewood Cliffs, NJ.

[18] The presumption that gives the probabilistic analysis normative power is that the movement of the voters' probabilities in their vote choices are normatively meaningful and comparable (again, see Mueller).

[19] For an example, consider the paper by Jasmina Arifovic & John Ledyard (2009). They are integrating a probabilistic learning response, altruism, fairness of treatment of self, and in a model of behavior that should expand the possibilities of modeling these problems.

Cain, Michael (1998). "An Experimental Investigation of Motives and Information in the Prisoners' Dilemma Game" Advances in Group Processes, 15:133-60, John Skvoretz and Jacek Szmatka editors, New York: JAI Press.

Charness, G. and M. Rabin, 2002. Understanding Social Preferences with Simple Tests, The Quarterly Journal of Economics, 117, 3, 817-869.

Coughlin, Peter (1984) "Probabilisitc Voting Models," in Encyclopedia of the Statistical Sciences, ed. Sam. Kotz, Norman Johnson, and Campbell Read, Vol. 6, NY: Wiley.

Cox, James C. Klarita Sadiraj and Vjollca Sadiraj (2001). "A Theory of Competition and Fairness without Equity Aversion," Paper presented at the International meetings of the Economics Society of America, Barcelona, Spain, June.

Damasio, Antonio (1999) The Feeling of What Happens. New York: Harcourt Brace and Company.

Dibble, C. (2006) Computational Laboratories for Spatial Agent-Based Models. In: L. Tesfatsion and K.L. Judd, (Eds.) Handbook of Computational Economics, Volume 2 edn. pp. 1511-1548. Amsterdam: Elsevier, North-Holland.

Edelman, Gerald M. (1992) Bright Air, Brilliant Fire: On the Matter of the Mind. New York: Basic Books.

Fehr, E. and Schmidt, K.M. (1999) A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114, 817-868.

Friedman, M. "The Methodology of Positive Economics," (1953) in his Essays in Positive Econcomics, Chicago: Univ of Chicago Press. pp. 3-43. Reprinted i n Wm. Breit and Harold M. Hochman (ed.), Readings in Microeconomics. Holt Rinehart and Winston. New Yo

Frohlich, Norman (1974) "Self-Interest or Altruism: What Difference?" Journal of Conflict Resolution, 18, March, pp. 55-73.

Frohlich, Norman and Joe A. Oppenheimer (1996)."Experiencing Impartiality to Invoke Fairness in the n-PD: Some Experimental Results." Public Choice, 86 (117 - 135).

Frohlich, Norman and Joe A. Oppenheimer (2000). "How People Reason about Ethics and the Role of Experiments: Content and Methods of Discovery." in Elements of Political Reason: Cognition, Choice

and the Bounds of Rationality. pp. 85 - 107. (eds. Arthur Lupia, Matthew McCubbins, and Sam Popkin). Cambridge University Press.

Frohlich, Norman and Joe Oppenheimer (2001) "Choosing from a Moral Point of View," Interdisciplinary Economics, (February) Vol. 12: 89-115.

Frohlich, Norman and Joe A. Oppenheimer, (2003), "Optimal Policies and Socially Oriented Behavior: Some Problematic Effects of an Incentive Compatible Device," in Public Choice V. 117, No. 3-4 (December.): 273-293.

Frohlich, Norman and Joe Oppenheimer (2006) Skating on Thin Ice - Cracks in the Public Choice Foundation. Journal of Theoretical Politics 18 (3):235-266.

Frohlich, Norman, Joe Oppenheimer, Anja Kurki (2004). "Modeling Other-Regarding Preferences and an Experimental Test," Public Choice, Volume 119, Issue 1-2, April: 91 - 117.

Frohlich, Norman and J. Oppenheimer, w Pat Bond and Irvin Boschman. (1984) "Beyond Economic Man." Journal of Conflict Resolution v. 28, no. 1, March, 1984: 3-24.

Green, Donald P. and Ian Shapiro (1994) Pathologies of Rational Choice Theory: A Critique of Applications in Political Science. New Haven, Conn: Yale University Press.

Grether, David M. and Charles R. Plott, (1979) "Economic Theory of Choice and the Preference Reversal Phenomenon," American Economic Review, 69 (September): 623 - 638.

Gunnthorsdottir, Anna, Daniel Houser, Kevin Mccabe, and Holly Ameden (2001) Disposition, History and Contributions in Public Goods Experiments. Mimeo.

Hempel, Carl G. (1965) Aspects of Scientific Explanation. NY: Macmillan Free Press.

Janssen, M. A., and T. K. Ahn, 2006. Learning, Signaling, and Social Preferences in Public Good Games. Ecology and Society 11(2): 21. URL: http://www.ecologyandsociety.org/vol11/iss2/art21/

Kahneman, D. and Tversky, A. (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47 (2):263-291.

Ledyard, J.O. (1995) Public Goods: A Survey of Experimental Research. In: Kagel, J.H. and Roth, A.E., (Eds.) The *Handbook of Experimental Economics*, pp. 111-194. Princeton: Princeton University Press.

Lindblom, Charles E. and David K. Cohen. (1979) Usable knowledge: social science and social problem solving New Haven: Yale University Press.

May, K.O. (1952), "A Set of Independent, Necessary, and Sufficient Conditions for Simple Majority Decision. Econometrica, vol.20 (October, 1952): 680-4.

Mackie, Gerry (2003). Democracy Defended. Cambridge University Press: Cambridge, England.

Mueller, Dennis (2003) Public Choice III. Cambridge Univ Press, Cambridge, UK.

Nola, Robert (2004) "Pendula, Models, Constructivism and Reality." Science and Education 13: 349-377.

Quattrone, George A. and Amos Tversky, (1988) "Contrasting Rational and Psychological Analyses of Political Choice." AMERICAN POLITICAL SCIENCE REVIEW. (82, NO. 3 SEPT.) 719-736.

Regenwetter, Michel; Jason Dana, and Clintin P. Davis-Stober (2008), "Transitivity of Preferences" University of Illinois at Urbana-Champaign, Working Paper.

Rabin, Matthew (1998). "Psychology and Economics," Journal of Economic Literature, vol 36 (March): 11 - 46

Rabin, Matthew (1993). "Incorporating Fairness into Game Theory and Economics." American Economic Review. v. 83, #5 (December): 1281 - 1302.

Roth, Alvin, E. (1995) "Bargaining Experiments," in Kagel, John H. and Alvin E. Roth, eds. (1995) The Handbook of Experimental Economics. Princeton, Princeton University Press: 253-342.

Saijo, Tatsuyoshi and Toru Yamaguchi, "The 'Spite Dilemma in Voluntary Contribution Mechanism Experiments." Paper presented at the 1992 Public Choice Meetings, March: New Orleans.

Shafir, Eldar and Amos Tversky (1995). "Decision Making," in Edward E. Smith and Daniel N. Osherson, eds. An Invitation to Cognitive Science: Thinking Volume 3, Second Edition, Cambridge, Mass.: MIT Press. 77-100

Simon, Herbert A. (1986) "Rationality in Psychology & Economics," The Journal of Business. v. 59, no. 4, Part 2 (October): pp. S209-224. a specially edited volume (by Robin M. Hogarth and Melvin W. Reder)

Simon, Herbert A. (1995) in Rationality and Society. vol. 7, No. 4, Oct.:

Tversky, Amos and Daniel Kahneman (1986), "Rational Choice and the Framing of Decisions," Journal of Business, v. 59, no. 4 pt. 2, pp. s251-s278. Reprinted in Karen Schweers Cook and Margaret Levi, eds. The Limits of Rationality. Chicago:

Valavanis, Stefan. (1958) "The Resolution of Conflict When Utilities Interact," The Journal of Conflict Resolution. v. 2, pp. 156 - 69.

Wendel, Stephen and Joe Oppenheimer (2007) "An Analysis of Context-Dependent Preferences in Voluntary Contribution Games with Agent-Based Modeling." Presented at the Conference on Behavioral Economics And Experimental Economics of the French Economic Association in Lyon, France (May 23-25), 2007.

Zeckhauser, Richard and Elmer Shaefer (1968). "Public Policy and Normative Economic Theory." in Raymond A. Bauer and Kenneth J. Gergen, eds. The Study of Policy Formation. Macmillan, The Free Press: New York: 27 - 101.

Appendix: Further Analysis of the Oscillation & Convergence Result Without Probabilistic Responses

The Three Person Case -- With three, or more, individuals the differences each individual experiences are not the difference between herself and one other player. Now it is the difference between the individual and the others in the group. Again, we use our concept of the individual *neutral contribution levels*, depicted as $N_{1, 2, 3}$. After the first round, where each participant will be giving their neutral contribution level, the initial difference between individual *i*'s contribution and the average of others is $N_i - (.5)(N_j + N_k)$, which we can still refer to as G_i. These differences will vary across individuals, but they will necessarily sum to zero. Using the following definitions:

$N_i = X_{i,t}^{opt}[Y_{i,t-1}=X_{i,t-1}] = (k-1) + f[0] + a_i$ Neutral contribution level of player *i*

$G_i = N_i - (.5)(N_j + N_k)$ Gap between neutral contribution levels

Based on these definitions, and equation 10, the following equations can be derived:

$G_i + G_j + G_k = 0$ Neutral contribution level gaps sum to zero

$X_{i,t} = N_i - r(d_{i,t-1})$ Player *i*'s Contribution in round *t*

$d_{i,t-1} = X_{i,t-1} - (.5)(X_{j,t-1} + X_{k,t-1})$ Difference between player *i*'s contribution

With 3 players, we can specify the maximizing contribution level in round t^* with a model of behavior similar to the 2 person case:

$$X_{i,t^*} = N_i - G_i r + \sum_{t=3}^{t^*} (-1)^{t-1}(3/2)^{t-2}(r)^{t-1} G_i.$$

$$X_{i,t^*} = N_i - G_i r(1 + \sum_{t=3}^{t^*} (-3/2r)^{t-2}).$$

Once again there are 3 ranges of the response rate, r, that determine the general shape of the responses. With two persons, the nature of the response pattern was determined by whether r was greater than, equal to, or less than ½. In the case of 3 individuals, the response cut points depend on whether r is greater than, equal to, or less than 2/3. When r = 2/3 the interaction will exhibit a steady state oscillation. The oscillations will continually be between N and the average of the N's that were contributed in the first round, paralleling the two person case. When r < 2/3 the size of the oscillations dampens, and each participant's contribution will converge toward a value below N_i - .5rG_i when G is positive, and above that when G is negative. The smaller r, the quicker will be the dampening. And finally, when r > 2/3 the oscillations will get bigger, until they reach the maximum possible given the resources the individuals have available. Again, the larger r, the faster would be the explosion in the oscillations.

The General, N-person Case **--** The general, N-person case follows a similar logic.

- For round 0, assume that each of M players starts at their Neutral Contribution Level: $x_{1..M,1} = N_{1..M}$.
- The resulting difference, $d_{i,1}$, between each player and that of the average of the others in round 1 is

  $d_{i,1} = N_i - \sum_{j \neq i} (N_j)/(n-1)$.

- As before, we define the gap in Neutral Contribution Levels, G_i $(=d_{i,1})$
- In any given round, *t*, observe that $\sum_{k=1..M} d_{k,t} = 0$
  - $d_{i,t} = x_{i,t} - \sum_{j \neq i} (x_{j,t})/(n-1)$

- $\sum_{i=1..M} d_{i,t} = \sum_{i=1..M} (x_{i,t} - \sum_{j \neq i} (x_{j,t})/(n-1)) = 0$

- In any given round, *t*, we now show that $d_{i,t} = G_i - rd_{i,t-1}\, n/(n-1)$
  - $x_{i,t} = N_i - r\, d_{i,t-1}$
  - $d_{i,t} = N_i - rd_{i,t-1} - \sum_{j \neq i} (N_j - rd_{j,t-1})/(n-1)$
  - $d_{i,t} = N_i - \sum_{j \neq i} (N_j)/(n-1) - rd_{i,t-1} + \sum_{j \neq i} (rd_{j,t-1})/(n-1)$
  - $d_{i,t} = G_i - rd_{i,t-1} + r\sum_{j \neq i} (d_{j,t-1})/(n-1) + r(d_{i,t-1})/(n-1) - r(d_{i,t-1})/(n-1)$
  - $d_{i,t} = G_i - rd_{i,t-1} + r \sum_{k=1..M} (d_{k,t-1})/(n-1) - r(d_{i,t-1})/(n-1)$
  - $d_{i,t} = G_i - rd_{i,t-1} + r*0/(n-1) - r(d_{i,t-1})/(n-1)$
  - hence, $d_{i,t} = G_i - rd_{i,t-1}\, n/(n-1)$

- Given that $d_{i,1} = G_i$, in any given round, *t*,
  - $x_{i,t} = N_i - rd_{i,t-1}$
  - $x_{i,t} = N_i - r(G_i - rd_{i,t-2}\, n/(n-1))$
  - $x_{i,t} = N_i - r(G_i - r(G_i - rd_{i,t-3}\, n/(n-1))\, n/(n-1))$
  - $x_{i,t} = N_i - G_i r(1 + \sum_{h=3..t} (-nr/(n-1))^{h-2})$

As in the previous 2 and 3 person cases, there are 3 ranges of the response rate, r, that determine the general shape of the responses. With n persons, the response cut points depend on whether r is greater than, equal to, or less than (n-1)/n. When r = (n-1)/n the interaction will exhibit a steady state oscillation. The oscillations will continually be between N and the average of the N's that were contributed in the first round, paralleling the two person case. When r < (n-1)/n the size of the oscillations dampens, and each participant's contribution will converge toward a value below $N_i - .5rG_i$ when G is positive, and above that when G is negative. The smaller r, the quicker will be the dampening. And finally, when r > (n-1)/n the oscillations will

get bigger, until they reach the maximum possible given the resources the individuals have available. Again, the larger r, the faster would be the explosion in the oscillations.

Table and figures

Table 1: 5-Person Prisoner's Dilemma (Showing Payoffs Only to One Player)	Amount Given by Others				
1 Person's Strategies	**40**	**30**	**20**	**10**	**0**
give 0	26	22	18	14	10
give 10	20	16	12	8	4

Table 2: The 2 Person Non Probabilistic Response Case

Round (t)	$X_{i,t}$ = Player *i*'s Contribution	Y_t = *Player j*'s Contribution	*Difference*
0	0	0	0
1	N_i	N_j	$N_i - N_j = G_i$
2	$N_i - r(G_i)$	$N_j + r(G_i)$	$G_i - 2r(G_i)$
3	$N_i - r(G_i - 2r(G_i))$	$N_j + r(G_i - 2r(G_i))$	$G_i - 2r(G_i - 2r(G_i))$
4	$N_i - r(G_i - 2r(G_i - 2r(G_i))$	$N_j + r(G_i - 2r(G_i - 2r(G_i))$	$G_i - 2r(G_i - 2r(G_i - 2r(G_i)))$

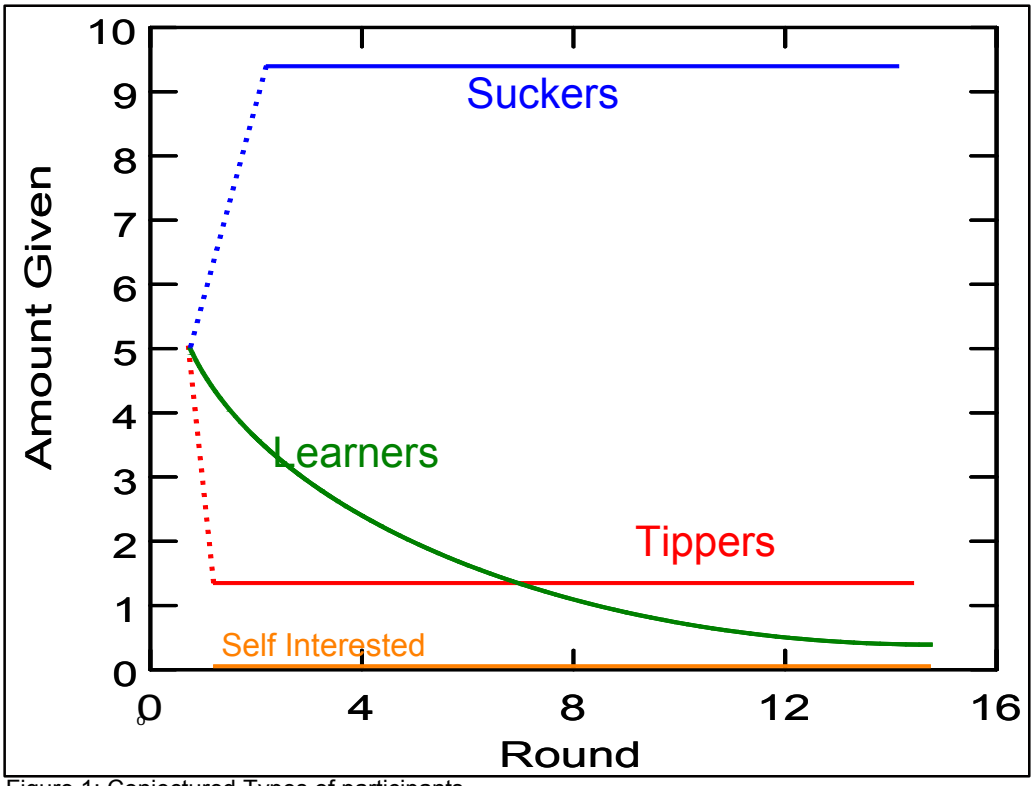

Figure 1: Conjectured Types of participants

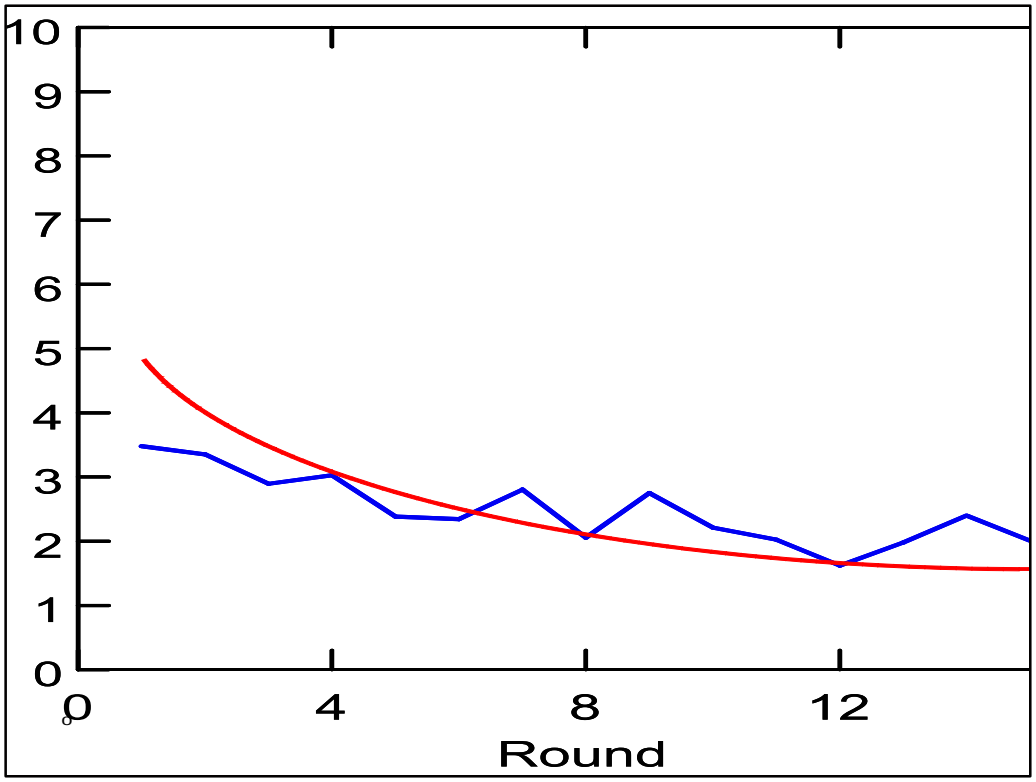

Figure 2: Average contribution among 5 person groups in a 'typical' VCM experiment (blue) and behavior with 'learning' conjecture (red).

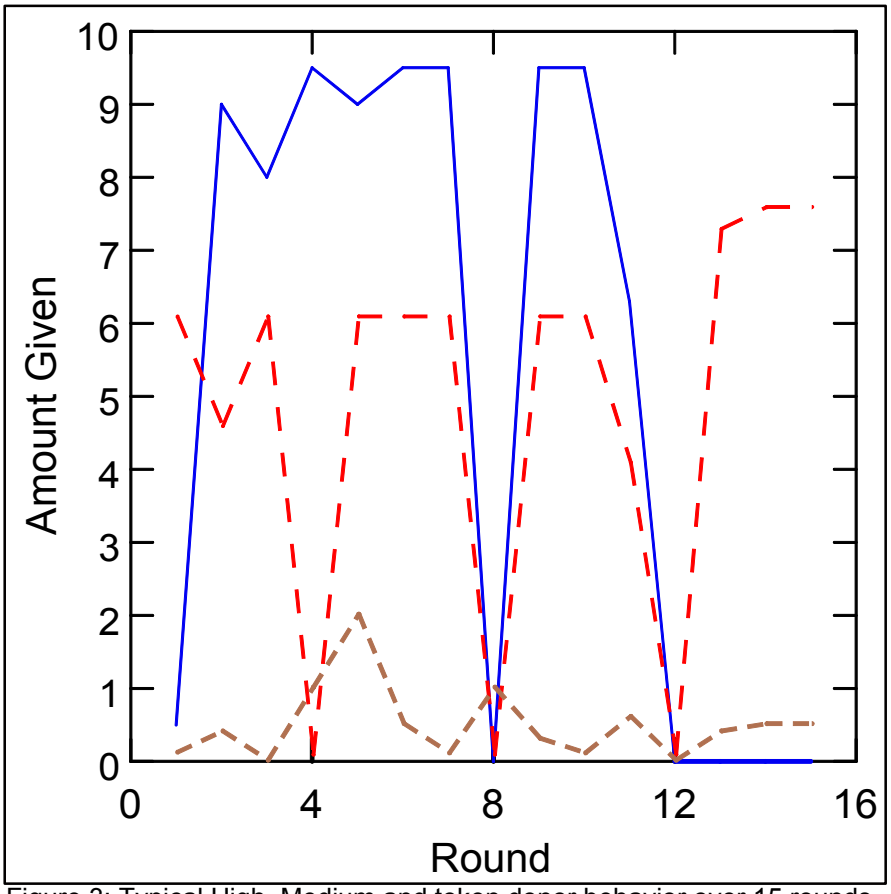

Figure 3: Typical High, Medium and token donor behavior over 15 rounds from experimental data.

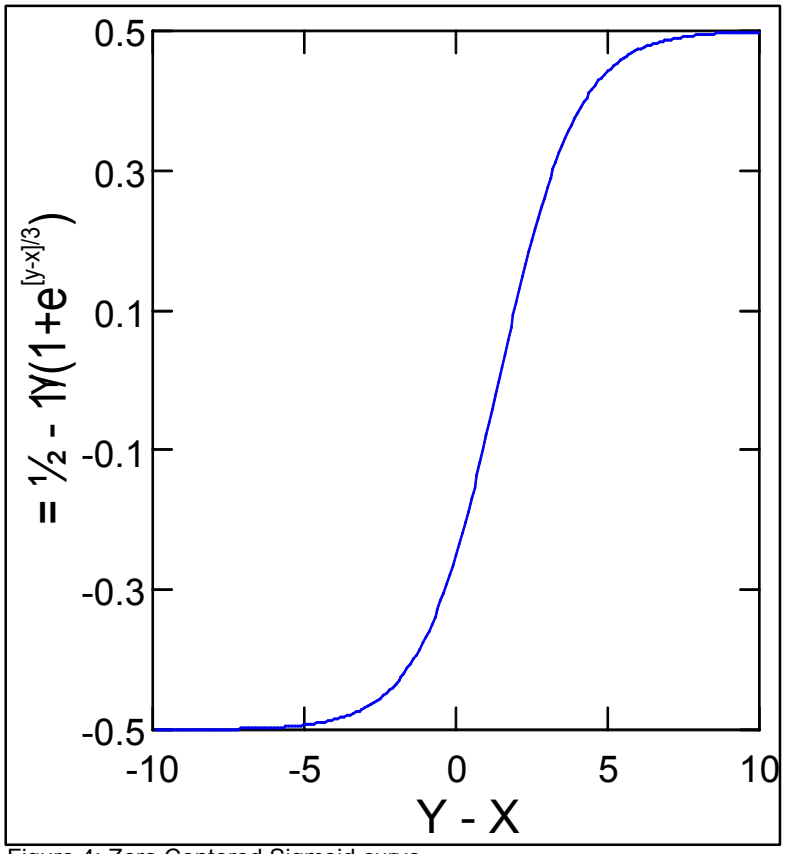

Figure 4: Zero Centered Sigmoid curve

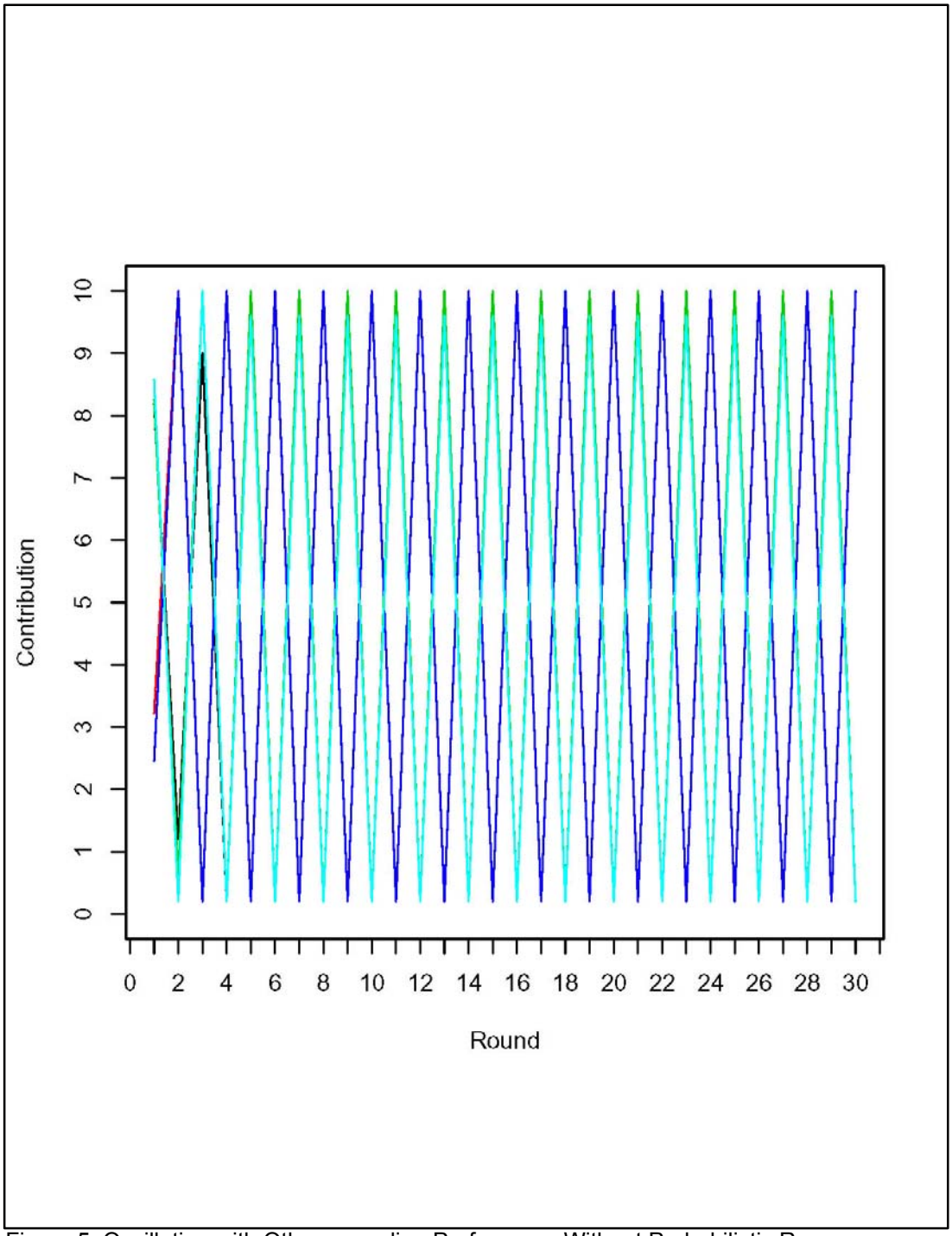

Figure 5: Oscillation with Other-regarding Preferences Without Probabilistic Response

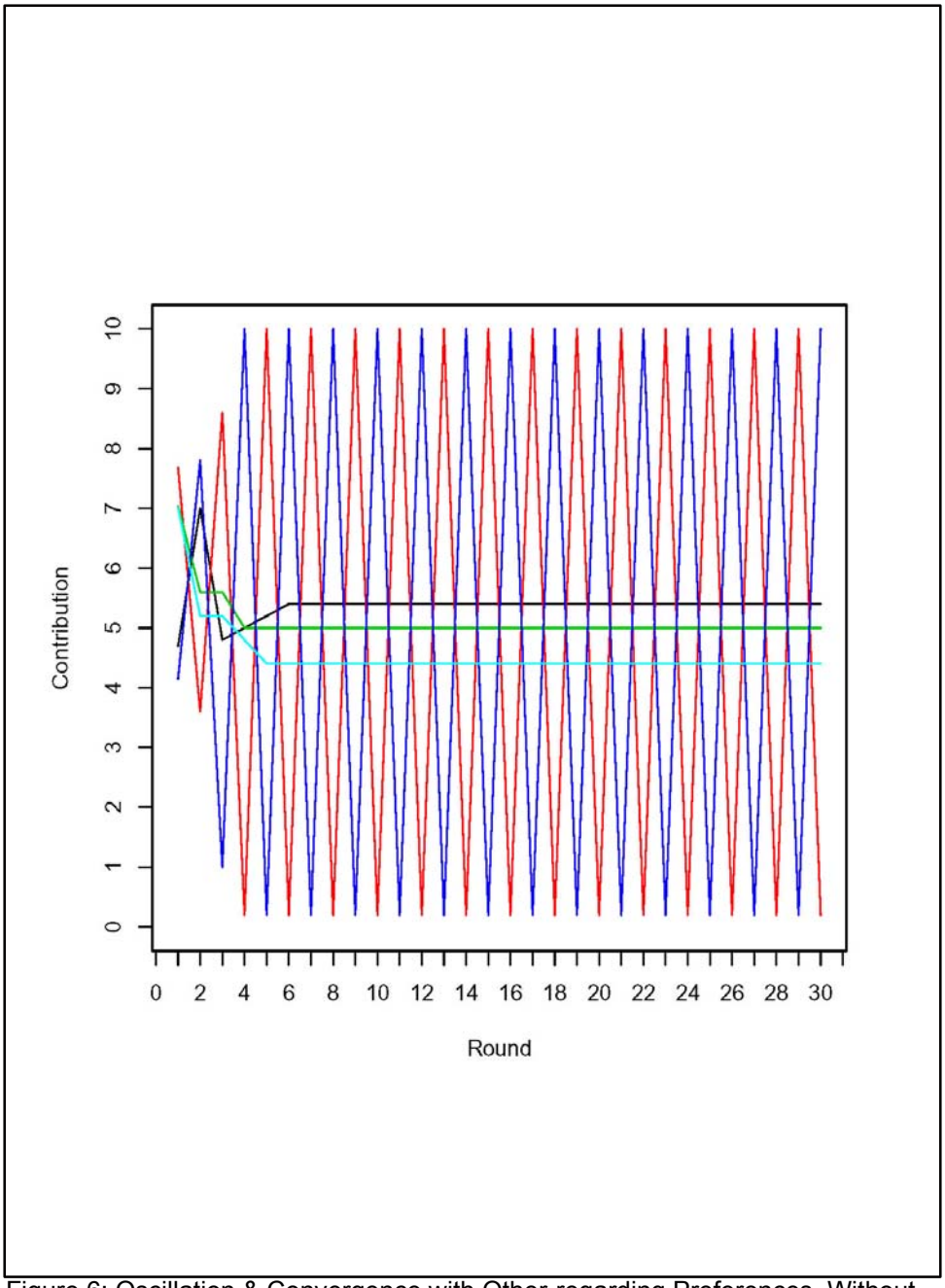

Figure 6: Oscillation & Convergence with Other-regarding Preferences Without Probabilistic Response8

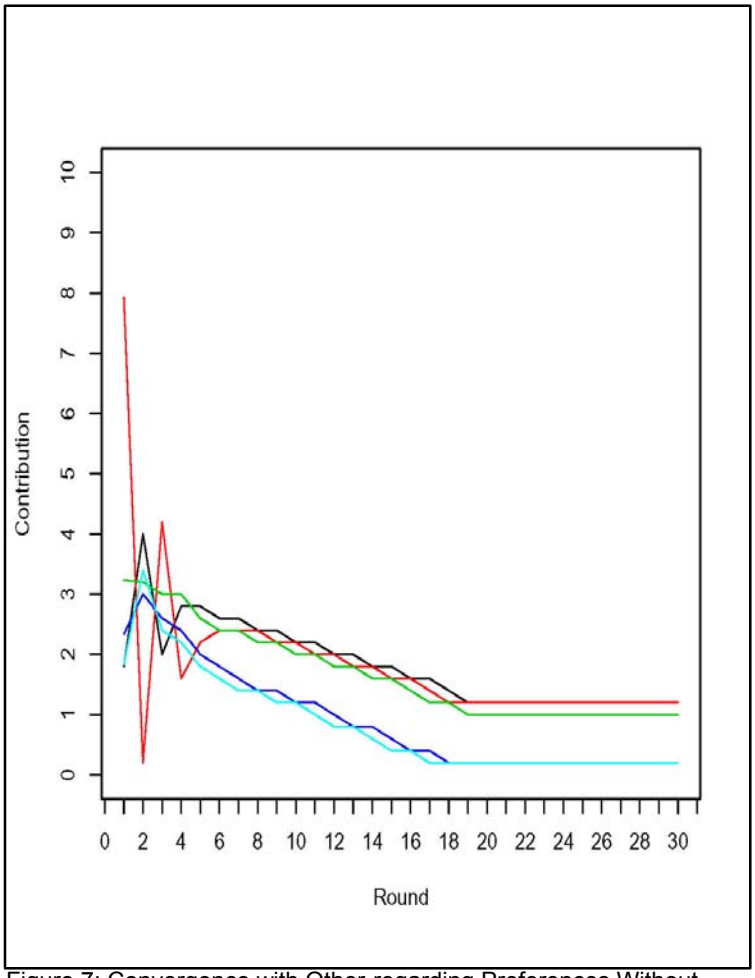

Figure 7: Convergence with Other-regarding Preferences Without Probabilistic Response

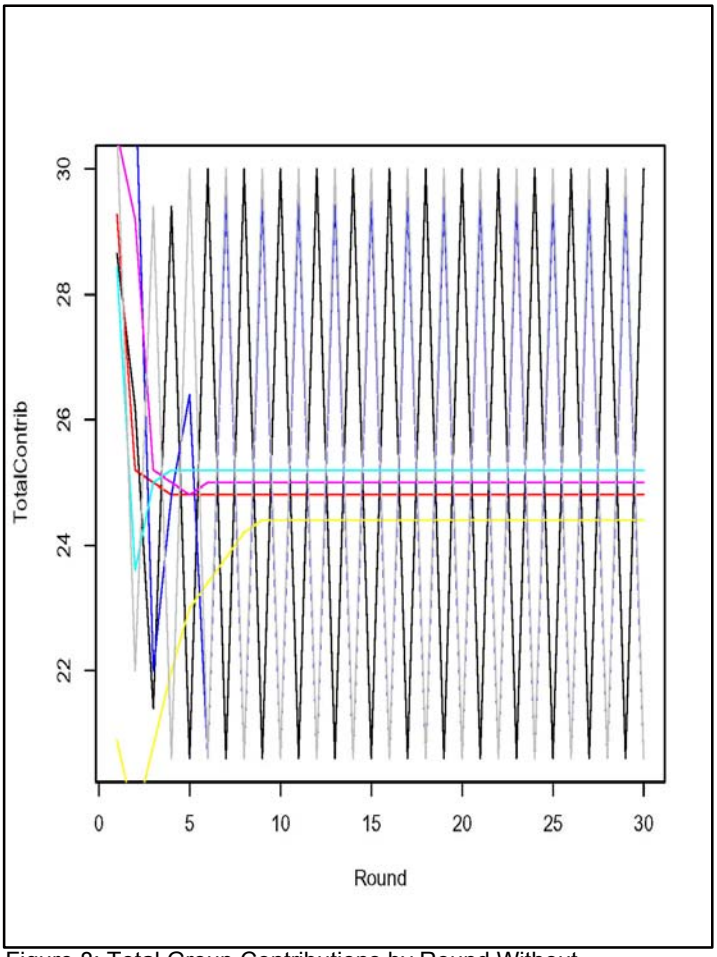

Figure 8: Total Group Contributions by Round Without Probabilistic Response

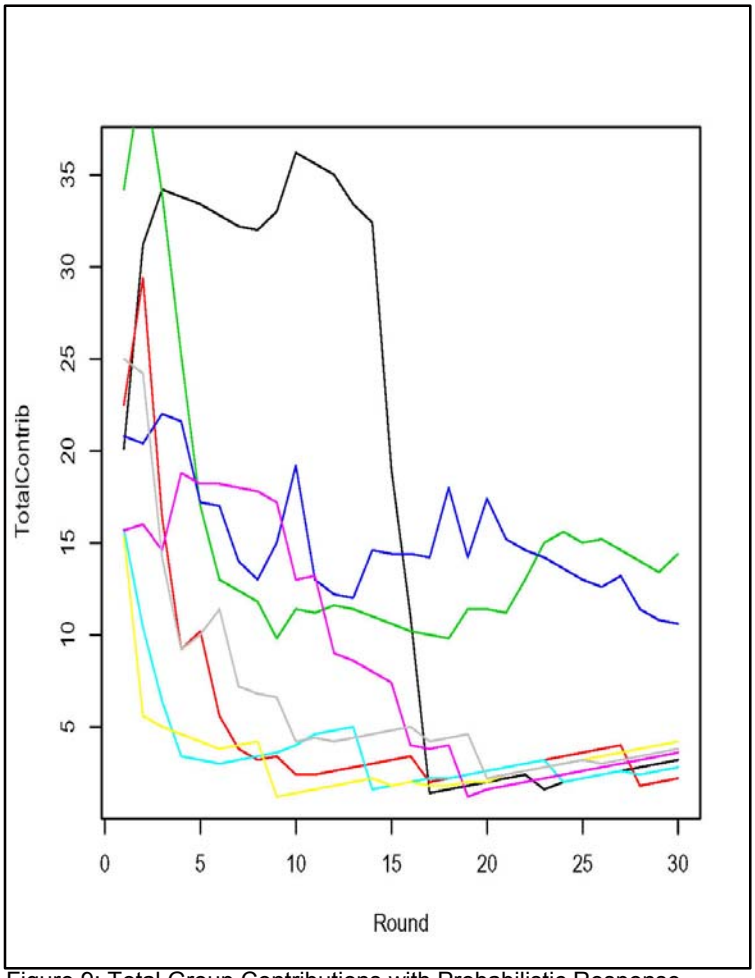

Figure 9: Total Group Contributions with Probabilistic Response

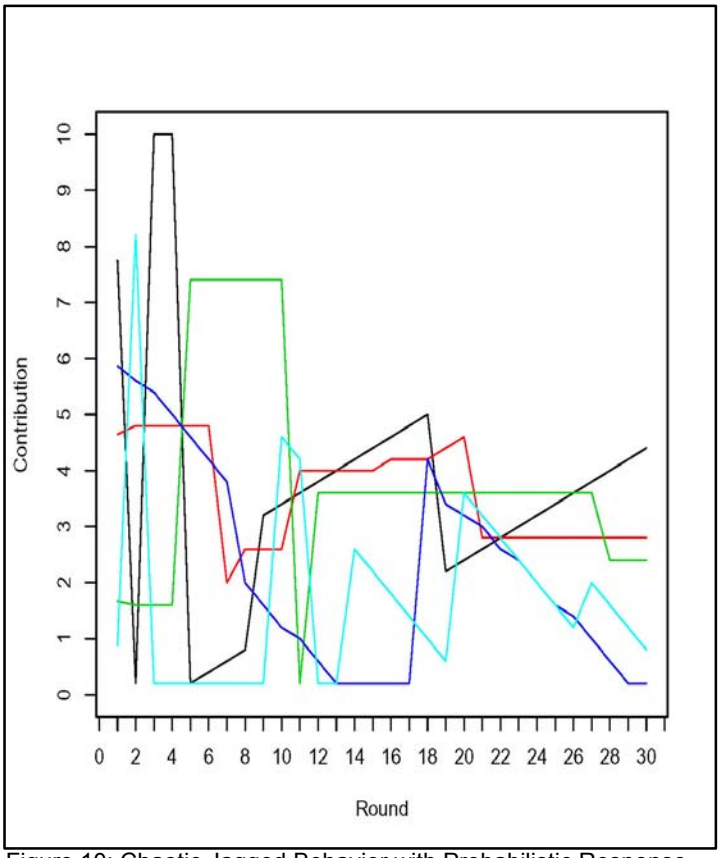

Figure 10: Chaotic Jagged Behavior with Probabilistic Response