Epistemic Conditions and Social Preferences in Trust Games

Gillies, Anthony S and Rigdon, Mary L

University of Michigan

13 July 2008

07.13.2008

Epistemic Conditions and Social Preferences in Trust Games*

Anthony S. Gillies
Department of Philosophy
University of Michigan

Mary L. Rigdon
Institute for Social Research
University of Michigan

Abstract

It is well-known that subjects in bilateral bargaining experiments often exhibit choice behavior suggesting there are strong reciprocators in the population. But it is controversial whether explaining this data requires a social preference model that invokes genuine strong reciprocity or whether some social preference model built on other-regarding preferences as a surrogate can explain it. Since the data precedes theory here, all the social preference models agree on most of it — making direct tests more difficult. We report results from a laboratory experiment using a novel method for testing between the classes of social preference models in the trust game that manipulates the distribution of payoff information in the game. We find evidence supporting the strong reciprocity hypothesis.

KEY WORDS: strong reciprocity, social preferences, trust game

* Corresponding author: Rigdon (mrigdon@umich.edu). Authors' names are listed alphabetically. This research was supported by a grant from the International Foundation for Research in Experimental Economics.

1 Introduction

It is a robust and well-known fact that subjects in bilateral bargaining experiments manage to reach off-equilibrium cooperative distributions. Something that stands most in need of explanation is choice behavior consistent with *strong reciprocity*.

Consider, as an example, the investment game (Berg, Dickhaut, and McCabe, 1995). Here Player i has some endowment M_i. Player 1 can invest any of that (X) at a growth rate r; Player 2 then decides a distribution of rX between them. Subjects playing this game anonymously in a one-shot environment under double-blind conditions still manage to reach distributions that Pareto dominate the equlibrium. In equilibrium, an investor (Player 1) invests nothing because he knows that a rational Player 2 would keep all of rX and return none of it; both players finish with their initial endowments. That is inefficient. But subjects famously do much better: with equal endowments at $\$M$, first-movers invest on average half of their endowment and second-movers, on average, return slightly more than the amount invested (Berg, Dickhaut, and McCabe, 1995; Ortmann, Fitzgerald, and Boeing, 2000; Casari and Cason, 2008). It is this second-mover behavior that suggests significant numbers of strong reciprocators in our midst.

What is true in the investment game is — again, famously — true for a wide class of bilateral bargaining games: contracting games (Rigdon, 2008), ultimatum games (Güth, Schmitteberger and Schwarze, 1982; Houser and Xiao, 2003; Oosterbeek, Sloof, and van de Kuilen, 2004), mini-ultimatum games (Falk, Fehr, and Fischbacher, 2003), and binary-choice trust games (McCabe, Rigdon, and Smith, 2002; Bohnet and Zeckhauser, 2004). In each case, of course, there are interesting boundary conditions and open emprical questions. But the broad result is impressive: a substantial portion of the population reveal social preferences in choice behavior. They care about, though perhaps not only about, the payoffs of relevant agents in the decision making environment.[1]

Just how they care about those payoffs is an open question. There are two broad classes of social preference models that answer the question in different ways. The first are models of strong reciprocity. Agents might care about the payoffs of others depending on the actions of those agents: they might value the payoffs negatively — and hence want *ceteris paribus* to see

[1] For an overview of some of the relevant recent literature on social preferences see Fehr and Fischbacher (2005).

them reduced — if they view the actions as hostile or desrving of retribution; they might value them positively — and hence want *ceteris paribus* to see them increased — if they view the actions as kind or deserving of approbation. Key to this way of trying to capture the data is that how agents see the motives and intentions of others is relevant to deciding what they themselves should do. Those agents are strong reciprocators, and models that aim at capturing this directly are models of *strong reciprocity*.[2]

But models of strong reciprocity are only one way of trying to explain this choice behavior. Other social preference models might just as well explain the observed choice behavior consistent with strong reciprocity *without* appeal to strong reciprocity at all. If so, then there is no empirical reason to think that there are strongly reciprocal types in the population at all and thus no empirical reason to favor models that truck in reciprocity talk. For instance, agents might care about the payoffs of others just because they are — conditionally or unconditionally — altruistic. This altrusim can take a variety of forms. Perhaps agents are *inequity averse*: they don't much care for unequal splits, but they might be willing to tolerate *some* inequality if it favors them (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). Or perhaps the relevant segment of the population simply prefers (*ceteris paribus* and subject to various constraints of course) distributions in which the relevant agents (say, their counterpart under certain conditions) do a bit better to other distributions (Levine, 1998; Cox, Friedman, and Sadiraj, 2008). Both kinds of models of altruism share a common thread: they both attempt to capture the relevant data not by appeal to strong reciprocity but by some altruism-inspired surrogate. We'll use the cover-term "other-regarding preferences" for that surrogate. That is the key idea for us, so we will largely ignore differences between the ways of doing that.

But, as in a lot of places in behavioral game theory, the data all of these social preference models are supposed to explain preceded the theory. Thus all of the theories — whether intentions-based or other-regarding preferences based — cover pretty much the same data. So novel tests that distinguish the classes of theories are welcome.

There are differences between the classes of models. One important difference is that models that try to capture the choice behavior consistent with strong reciprocity by appeal to some altruistic surrogate assume that prefer-

[2]Prominent examples of such models include Dufwenberg and Kirchsteiger (2004); McCabe and Smith (2000b); Falk and Fischbacher (2005). For overviews of the landscape, see Smith (2008); Gintis (2000).

ences over distributions are separable from the paths through the game tree that determine those distributions. The preferences are defined over bundles and those definitions are insensitive to the presence or absence of alternative bundles. That is not true for (direct) models of strong reciprocity since the whole point in that case is to make preferences sensitive in just that way. This asymmetry makes modeling strong reciprocity much harder. But it is also a difference that has been exploited in the laboratory to test the classes of models — by varying alternative distibutions avalilable, we can get direct tests of the theories (McCabe, Rigdon, and Smith, 2003; Charness and Rabin, 2002; Falk, Fehr, and Fischbacher, 2003; Ashraf, Bohnet and Piankov, 2006).

We report results of a laboratory experiment that exploits a novel way of distinguishing between classes of social preference models that does not depend on manipulating either the possible distributions available or the paths traversed through the game tree to arrive at those distributions. We focus — for simplicity — on binary-choice trust games, and in particular on second-mover behavior in those games. Rather than asking about alternative paths through the game tree, we instead ask about how varying the way *information* is distributed about the game affects social preference models. We show that different classes of models treat information about the material payoffs in different ways, and this difference implies differences in predictions across our treatments in different places. Our main result is that the observed choice behavior is compatible with strong reciprocity models but not the other-regarding preference surrogates. This is evidence in favor of the strong reciprocity hypothesis: opportunists and altruists do not exhaust the population; there are also genuine strong reciprocators.

2 Epistemic Conditions

It is common to think that there is a strong connection between what is common knowledge and the pure strategy equilibria in two-person non-cooperative games. After all, it is tempting to think that in equilibrium rational Player i does his part because rational Player j does her part — and j does her part because i does his. That kind of mutual expectation between i and j does sound like it should have something to do with common knowledge. But the truth here, as pointed out by Aumann and Brandenburger (1995), is much simpler: if the players know their own payoffs and the strategy choices of each other, then rational players' choices will be a Nash equilibrium. This way

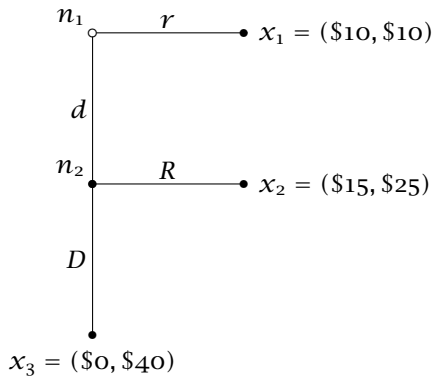


Figure 1: The $10 Trust Game

of distributing information is far short of common knowledge.[3] There is no similar interesting *necessary* condition on what information — higher-order or otherwise — players must have in order for their choices to be in equilibrium. The reason is simple: players can stumble into equilibrium for no reason in particular, without anyone knowing much of anything. Still, as Aumann and Brandenburger emphasize, there is no way to weaken the sufficient conditions here. It is in that sense that that distribution of information characterizes Nash equilibrium.

But suppose we care not so much about whole strategy profiles but only about second-movers' contribution to those profiles. And suppose we care not so much about Nash equilibria in classical game theory but about stable points in social preference amendments in behavioral game theory. There are still interesting epistemic conditions to be found. Those conditions characterize who needs to know what about the game in order for the model in question to explain second-mover behavior. In other words: fix a social preference model. What distribution of information about the material payoffs in the game is required for that model to explain second-mover choice behavior? Here is where we see the social preference models segregating themselves.

We will focus on a trust game with the payoffs as in Figure 1. Suppose x

[3] This, in fact, is only the preliminary observation that Aumann and Brandenburger (1995) make. Their two main results are a good deal more general (dealing with mixtures of strategies interpreted as conjectures) — as with the preliminary observation, common knowledge plays no role.

is a terminal node in the tree. Then, where $\pi(x)$ is the distribution (vector of payoffs) determined by terminal node x, let $\pi_i(x)$ be i's share of that.[4] Where n is a non-terminal choice node, let Π — with or without subscripts — give us the the set of possible distributions reachable from n:

$$\begin{aligned} \Pi(n) &= \{\pi(x)\colon x \text{ is a terminal node that is a descendant of } n\} \\ \Pi_i(n) &= \{\pi_i(x)\colon x \text{ is a terminal node that is a descendant of } n\} \end{aligned}$$

This bit of notation helps to state who needs to know what, given some model, to explain observed choice behavior in the trust game that is consistent with the presence of strong reciprocators.

To see this, let's label the two choice nodes in the simple trust game. Let's call the first, controlled by Player 1, n_1 and the second, controlled by Player 2, n_2. $\Pi(n_1)$ is the set of all three distributions possible in the game, $\Pi(n_2)$ is the set of distributions that Player 2 has to choose between. What we now want to do is consider three ways of distributing information about this game between Player 1 and Player 2. In particular, we look at three ways of distributing information about $\Pi(n_1)$ and $\Pi(n_2)$. (There are, of course, other ways of distributing the information as well. As will become clear, we care about these three ways for a reason.)

First, information about payoffs may be a matter of private knowledge. That is, suppose that while the space of *actions* for Player i is known by Player $-i$, the *payoffs* are merely privately known in the sense that each player only knows his own share of that distribution: for each terminal node x, Player i only knows $\pi_i(x)$. Hence, Player 1 only knows $\Pi_1(n_1)$ and Player 2 only knows $\Pi_2(n_1)$. This is the *private distribution* of payoff information.

Second, information about payoffs may be a matter of asymmetric knowledge. Of course, each player knows his own payoff for each terminal node. What we are interested in here is a distribution of information in which, in addition, Player 2 knows Player 1's payoffs and knows that Player 1 does *not* know this. That is, for each terminal node x, Player 2 knows $\pi_1(x)$ and knows that this fact is something Player 1 does not know. Under this distribution of

[4] Technically it is a bit better to first label the terminal nodes $x_1, \ldots, x_n$ and take $\pi(x_i)$ to be a vector of ordered pairs (x_i, m_j) where m_j is player j's share of the monetary payoff at x_i — that way we can always recover an association between a player's payoff and the node, and hence the path through the game tree, that yields that payoff. For simplicity, we suppress that detail here. When no confusion will arise, we omit reference to the terminal node and write π_i. Similarly, when convenient and when no confusion will result we simply identify terminal nodes with the distributions they determine.

information, Player 1 still only knows $\Pi_1(n_1)$ but Player 2 now knows quite a bit more: (i) she knows $\Pi(n_1)$ and (ii) she knows that Player 1 does not know (i). This is the *asymmetric distribution* of payoff information.

Third, information about payoffs may be a matter of common knowledge. This represents the standard assumption, and what usually gets operationalized in laboratory settings by reading the instructions aloud.[5] Under this distribution, for each terminal x, each player knows the entire distribution $\pi(x)$, each knows that each knows this, and so on. Here $\Pi(n_1)$ is common information.[6] This is the *common distribution* of payoff information.

Different models of Player 2 behavior treat different levels of payoff information in the trust game as relevant in explaining observed choice behavior that is consistent with strong reciprocity. Those differences are the minimal epistemic conditions for the respective models, and what we use to differentiate the theories.

Let us start, to illustrate, with the *homo economicus* model. This model says that Player 2's utility (like that of all agents) at terminal node x, $U_2(x)$, is identical to her material payoff at x and that Player 2 seeks to maximize her utility and thus her own payoff. And consider the private distribution of information under which Player 2 only knows, for each terminal node, her own payoff at that node. A self-interested Player 2 has no trouble seeing what to do: she maximizes, and we already know that she doesn't need to know anything about $\Pi_1(n_1)$ to do that. All she needs is to know $\Pi_2(n_2)$. Everything else is irrelevant. So, since $\pi_2(x_3) = 40 > \pi_2(x_2) = 25$, this model predicts that Player 2's choose the opportunistic D.

We can get this tight an explanation from the *homo economicus* model from modest means: Player 2 needs to know $\Pi_2(n_1)$ but nothing more. Anything less, of course, and we lose the prediction. Thus:

Observation (*Homo Economicus*)**.** The minimal epistemic condition (about material payoffs) for Player 2 choice behavior in the *homo economicus* model is that Player 2 know $\Pi_2(n_2)$.

We got that explanation without Player 2 knowing anything but her own pay-

[5]Public announcements are often, though not aways, a good way of making something common knowledge.

[6]Note that these three distributions do not form a tidy hierarchy in the sense that we only add to the knowledge of the agents as we add information: in particular, it is false that if Player 2 knows that p under the common distribution then she knows that p under the asymmetric distribution. That is because one of the things she knows in that latter case is that Player 1 does not know that Player 2 knows π_1. But Player 1 does know that in the former case.

off information in the bottom of the tree. That is the characteristic information for that explanation. Thus all the other higher-order information under the asymmetric and common information distributions is irrelevant to explaining Player 2 behavior given this model. Explaining her behavior, given this model, doesn't depend in any way on that information.

We can ask the same kinds of questions of social preference models. What we will see is that while they all allow for differences in choice behavior between our distributions of information, they predict those differences in different spots. Since our interest is not in particular theories but in the broad classes they represent, we will leave the discussion as informal as possible and just state the stylized trends that the theories predict.

The first thing to notice is that, for our purposes, there is no interesting difference between unconditional and conditional altruism models. Both models try to explain behavior consistent with strong reciprocity by appeal to some surrogate that allows $U_2(x)$, Player 2's utility at terminal node x, to be affected by some property of $\pi(x)$. The models differ on what that property is, and how it can and cannot affect $U_2(x)$; but those are family squabbles.

What is significant is that *all* such other-regarding preference models share a common characterizing epistemic condition. All such models say that Player 2 will act so as to maximize her utility, picking whichever of $U_2(x_2)$ and $U_2(x_3)$ is greater. But what information does she need to do that in the simple trust game? She clearly must know what she can earn in the two distributions she has to choose from so she must know $\Pi_2(n_2)$. But since these are other-regarding preference models and her utility for each distribution is also a function of what Player 1 gets at that distribution, she must also know $\Pi_1(n_2)$. So she must know $\Pi(n_2)$. In terms of what Player 2 needs to know about Π, that is it. As far as payoffs are concerned, $U_2(x_2)$ and $U_2(x_3)$ are only a function of $\Pi(n_2)$. In particular, it is manifestly *not* a function of what Player 1 knows about what Player 2 knows about $\Pi(n_2)$ or $\Pi(n_1)$.[7]

Consider as an example a simple form of an inequity aversion model (Fehr and Schmidt, 1999). Here, Player 2's utility at x $U_2(x)$ is $\pi_2(x)$ decremented by how much she dislikes inequality that favors Player 1 (if the distribution x determines does favor Player 1) or by how much she dislikes inequality in her

[7]Some have argued — see, e.g., Cox, Friedman, and Sadiraj (2008) — that there is room for other-regarding preference models that require Player 2 to know something of $\Pi(n_1)$ in order to figure $U_2(x_2)$. For our purposes, we can remain agnostic on that question since even if that were true our main point is unaffected: U_2 still wouldn't depend on what Player 1 knows about what Player 2 knows about $\Pi(n_1)$.

favor (if x does favor Player 2):

$$U_2(x) = \begin{cases} \pi_2(x) - \alpha_2(\pi_1(x) - \pi_2(x)) & \text{if } \pi_1 > \pi_2 \\ \pi_2(x) - \beta_2(\pi_2(x) - \pi_1(x)) & \text{otherwise} \end{cases}$$

The weights are constrained so that $\beta_2 \leq \alpha_2$ and $0 \leq \beta_2 < 1$. The thing to notice is that no matter what values α_2 and β_2 take, what this requires is that Player 2 know the material payoffs of the whole distribution $\pi(x)$ and for her to sensibly compare such weighted utilities, she must know — in the simple trust game — $\Pi(n_2)$. It does *not* require her to know anything about what Player 1 knows, and in particular it does not require her to know what Player 1 knows about what Player 2 knows about $\pi(x)$ or, in the case we are interested in, about what Player 1 knows about what Player 2 knows about $\Pi(n_2)$. And if she *does* have this information, there is no place the model can put it to use; it is just not relevant. This is an example, but the reasoning holds good for the class of other-regarding preference models.

Observation (Other-Regarding Preference Models)**.** The minimal epistemic condition (about material payoffs) for Player 2 choice behavior in other-regarding preference models is that Player 2 know $\Pi(n_2)$.

So the minimal epistemic condition for an other-regarding preference model to explain Player 2 choice behavior is simple: Player 2 must know the details of $\Pi(n_2)$. Thus any payoff information beyond this is redundant from the point of view of other-regarding preference models and thus irrelevant to any explanation of Player 2 choice behavior such models might have on offer.

Things are different for direct models of strong reciprocity. The idea behind these models is that behavior in bargaining environments like the simple trust game are like the execution of an implicit incomplete contract. Player 2 only has a choice if Player 1 passes on the outside option; Player 2 reasons that Player 1 would only do that, trusting her and putting himself at risk — *ceteris paribus* of course — if he thought Player 2 would reciprocate. That outcome Pareto dominates the equilibrium, Player 1 seems to be saying, so let's do that.

That is the idea behind such models. And while that is clear enough, it is still hard to make that precise enough to model in a way that is both illuminating and tractable. (Indeed, this is one reason why other-regarding preference models are so tempting: since they invoke separability, they are mathematically simple.) But all ways of making it precise take something like this form for the trust game: Player 2's utility at x depends not only on that entire distribution

$\pi(x)$ but also on the path through the tree that got the players to that point. This is often encoded as a kindness term that measures how kind or trusting Player 2 thinks Player 1 is.

No matter how these details get spelled out, we can still say something useful for the whole class of models. What we care about is how these models work when Player 2 compares her kindness/intentions-adjusted utility for choosing R with her kindness/intentions-adjusted utility for choosing D. So in figuring how $U_2(x_3)$ compares to $U_2(x_2)$ she needs to weigh how kind or trusting she figures Player 1 is in choosing d instead of r. Our point here is simple: Player 2 cannot really attribute any trust to Player 1, and so cannot be acting so as to execute an implicit incomplete contract to get to the Pareto dominant distribution, unless the terms of that contract — that is, $\Pi(n_1)$ — are common information between them (or, at least, a good enough approximation thereof). The reason is just that in order to have a contract, both parties need to know what is at stake and what is at risk. You cannot, in the ideal case, have a contract between parties in which the terms to the contract are not a matter of common information between the parties. That just isn't a contract.

Observation (Strong Reciprocity Models)**.** The minimal epistemic condition (about material payoffs) for Player 2 choice behavior in (intentionality-based) strong reciprocity models is that $\Pi(n_1)$ is common information.

Obviously, since common knowledge can be hard to come by, these models do not assume that agents always have that. But they do say that the closer Player 2 comes to having good enough grounds for thinking that the details of $\Pi(n_1)$ is some good enough approximation of common knowledge, the more likely she will take Player 1's off-equilibrium behavior to be trusting behavior. Whether or not she feels much like reciprocating is a further question.

3 Design and Predictions

Our laboratory experiment simply systematically varies the distributions of payoff information in the simple trust game. That gives us three treatments:

- PRIVATE: Each Player i only knows $\Pi_i(n_1)$; that is, i only knows i's share of the distribution at each terminal node (the private distribution)
- ASYMMETRIC: Player 1 only knows $\Pi_1(n_1)$; Player 2 knows $\Pi(n_1)$ and knows that Player 1 does not know this (the asymmetric distribution)

- COMMON: $\Pi(n_1)$ is a matter of common knowledge between the players (the common distribution)

Each of these corresponds to the the epistemic conditions of our three broad classes of models (*homo economicus*, other-regarding preferences, and strong reciprocity). Putting the two together then gives three stylistic predictions about comparative cooperation rates by Player 2s in the population. We illustrate those predictions here, noting that the predictions do not depend on any particular model within a given class or on any particular parameterization of any particular model with a given class. The predictions are, in this sense, fully general.

For each class of models, the basic reasoning to get predicted cooperation rates is straightforward. Suppose a given class of models has E as a characterizing epistemic condition. Then any information about the material payoffs beyond that is redundant information. Thus if two treatments differ only in that one provides E and the other provides E plus some payoff information beyond E, then that class of models predicts Player 2 behavior, and in particular Player 2 behavior consistent with strong reciptocity, to be constant across those treatments.[8]

Take *homo economicus*. Obviously, we are going to get a very uninteresting limit case prediction here: no reciprocity in any treatment. That is simply because, first, the model predicts no cooperation by Player 2 under a private distribution of payoff information (and so none in PRIVATE) and that distribution is the minimal information for that model. All the other ways of distributing information that we care about — in ASYMMETRIC and COMMON — just provide extra information that the *homo economicus* model deems irrelevant. So, if that model were right, we would expect rates of cooperation to remain constant across private, asymmetric, and common distributions of payoff information. And, moreover, they should be low. Let $\Pr(R|X)$ be the proportion of Player 2s choosing R conditional on being in treatment X. Then the first prediction is:

Prediction (*Homo Economicus*)**.**

1. $\Pr(R|\ \text{PRIVATE}) = \Pr(R|\ \text{ASYMMETRIC}) = \Pr(R|\ \text{COMMON})$

[8] To be a little more precise: the correspondence between epistemic conditions and the models is that *homo economicus* treats information beyond what Player 2 knows about $\Pi_2(n_2)$, not that beyond $\Pi_2(n_1)$, as irrelevant; *mutatis mutandis* for other-regarding preference models. In our distributions Player 2 also has some information about the outside option. Still, since $\Pi(n_2) \subset \Pi(n_1)$, this does no harm to the predictions and makes our distributions, and hence our experimental design, more symmetric.

2. $\Pr(R|\ \textsf{PRIVATE})$ is low

Other-regarding preference models, of course, were built to explain behavior that is consistent with strong reciprocity. They say that Player 2's behavior is sensitive to the entire distributions she must choose between, not just her shares of those distributions. That is, for her to maximize her expected utility she needs to know $\Pi(n_2)$ not just $\Pi_2(n_2)$. In the kind of epistemic situation in PRIVATE, she lacks some of that crucial information and so we expect low cooperation. But when she has it — that is, in any kind of epistemic situation in which her information entails the minimal epistemic condition — we should expect much higher rates of cooperation. In particular, in ASYMMETRIC, Player 2 knows $\Pi(n_1)$. This straightaway entails that Player 2 knows $\Pi(n_2)$, and so we should expect more Player 2s opting for R here. The interesting comparison is whether we should expect *even more* plays of R by Player 2s in COMMON. Given this class of models, clearly not: the information under the common distribution that goes beyond what Player 2 has in the asymmetric distribution is strictly irrelevant, given an other-regarding preference model, to the adjusted utilities $U_2(x_2)$ and $U_2(x_3)$. Since that extra information is choice-irrelevant, it ought not make a difference. This second prediction can then be put this way:

Prediction (Other-regarding Preferences)**.**

1. $\Pr(R|\ \textsf{PRIVATE}) < \Pr(R|\ \textsf{ASYMMETRIC})$

2. $\Pr(R|\ \textsf{ASYMMETRIC}) = \Pr(R|\ \textsf{COMMON})$

And, finally, let's look at intentions-based models of strong reciprocity. Assuming some model in this class, the private and asymmetric distributions of payoff information are not sufficient epistemic conditions for robust trust by Player 1 and so not sufficient for strong reciprocity by Player 2. Still, no one thinks *all* Player 2s are strong reciprocators — some are opportunists, some altruists. In PRIVATE, opportunists have all the information they need but there is no chance for altruism, except by dumb luck. So we should expect very few Player 2s opting for R. What about in the richer epistemic situation in ASYMMETRIC? Under the asymmetric distribution, notice, there is enough information for the altruists in the population to act accordingly: once they know $\Pi(n_2)$ they can opt for R in non-random ways. But the strong reciprocity hypothesis is that the altruists and opportunists do not exhaust the population. So although we would expect more cooperation under the

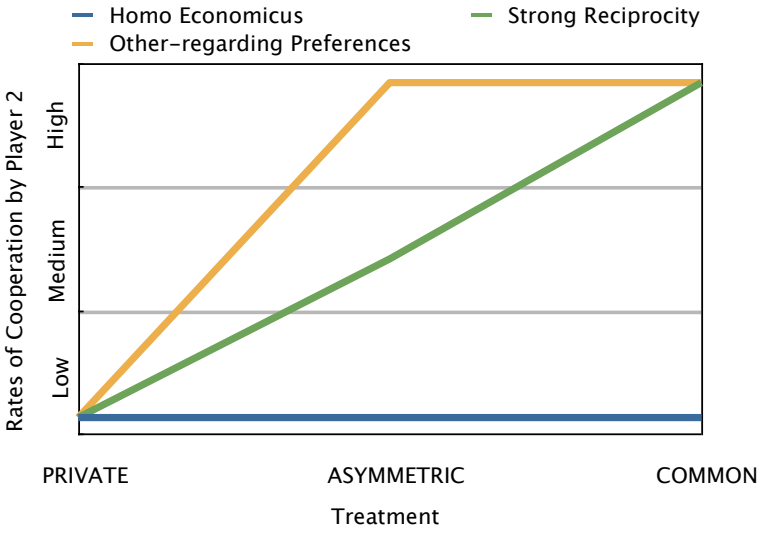


Figure 2: Stylized Predictions: Rates of Cooperation by Player 2

asymmetric distribution, we should expect *even more still* when we move from ASYMMETRIC to COMMON. According to models in this class, only the common distribution provides payoff information that is sufficient for the execution of an implicit contract between the agents.

Prediction (Strong Reciprocity)**.**

1. $\Pr(R|\ \textsf{PRIVATE}) < \Pr(R|\ \textsf{ASYMMETRIC})$
2. $\Pr(R|\ \textsf{ASYMMETRIC}) < \Pr(R|\ \textsf{COMMON})$

Different social preference models require Player 2 to have different information about the possible distributions in order to explain off-equilibrium behavior. What we have seen is that this can be exploited — even without considering any particular model — to segregate the classes of models in virtue of their emprical commitments. These stylized predictions are summarized in Figure 2.

4 Experimental Procedures

All sessions were run in the Robert Zajonc's Laboratory at the University of Michigan's Institute for Social Research.[9] The sessions were run December 2007 through February 2008. The subjects played the trust game once and only once, and this fact was common information.[10] Subjects were randomly assigned to the role of Player 1 or Player 2. Player 1s and Player 2s were kept separate for the entire experiment using two rooms. Once an even number of subjects had arrived, instructions were handed out, and then read aloud. Subjects were allowed to ask questions individually. When there were no additional questions, the experiment began. All subjects received a large envelope which contained two smaller envelopes: one had the decision sheet and the other had a short demographic survey to be completed after all subjects had made their decisions. Subjects were asked to remove the decision sheet, record their move of right or down with an arrow, place the sheet back in the envelope, and drop the contents in a box at the back of the lab.

The trust game was implemented using the strategy method. Player 2s are asked to assume that Player 1 has selected down and make their choice of right or down accordingly. This method allows observation of behavior by all Player 2s regardless of whether Player 1s choose d.[11]

Most sessions had 20 subjects and took less than 1 hour to complete. Each subject received $5 for showing up on time. Average earnings (excluding the show-up payment) were $9.17 for Player 1s and $17.20 for Player 2s, and varied from $0 to $40.

The only difference in the instructions for the treatment conditions was in the description of the information available to the players about potential payoffs. In PRIVATE, instructions for Player i show question marks that obscure the payoffs for Player $-i$ at terminal nodes and explain that these question marks hide the other player's payoff. In ASYMMETRIC, instructions for Player 1 are as in PRIVATE; Player 2 sees the full unobscured game tree and also sees the game tree and discussion as it appears in Player 1's instructions. In COMMON, there are no obscuring question marks. Subjects also completed a post-instruction quiz that was checked for accuracy: Player 1 had to demon-

[9] Data are avalable from the authors at request.

[10] Subjects who had participated in similar experiments were excluded from recruitment.

[11] Some evidence exists that trust games implemented in this manner generate differences in behavior (McCabe, Smith, and LePore, 2000; Casari and Cason, 2008; Solnick, 2007). We will return to this below in Section 6.

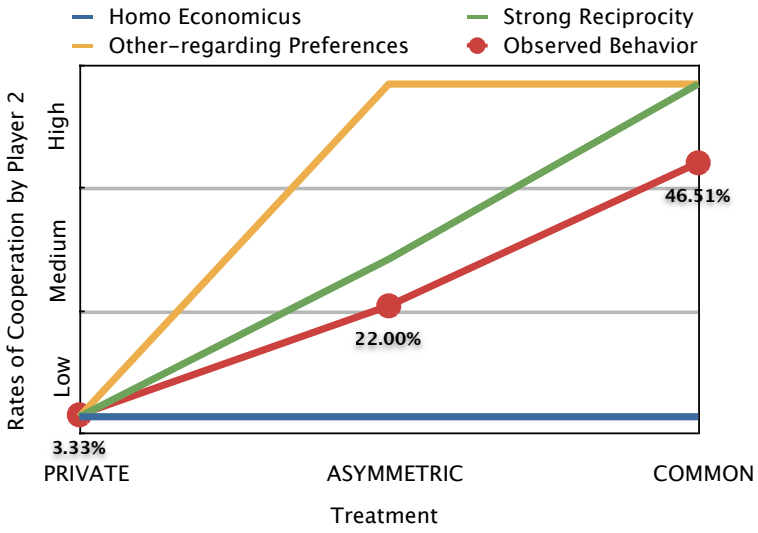


Figure 3: Rates of Cooperation by Players 2

strate an understanding of the relevant information distribution; and Player 2 had to register a prediction about Player 1's response. Subjects showed no difficulty with the quiz. At the completion of the experiment, subjects filled out a short demographic survey.[12]

5 Main Results

Our main result is that Player 2s choose R at low levels in PRIVATE, at medium levels in ASYMMETRIC, and at high levels only in COMMON. These differences are incompatible with the stylized predicitions of the *homo economicus* model and other-regarding preference models, but consistent with the predictions of strong reciprocity models. We present this main result in three different ways: Figure 3 plots the data against the predictions of the classes of models, Table 1 reports statistical tests, and Table 2 reports regression results.

The predictions we summarized in Figure 2 say what we should expect, given a type of model, as we add to the payoff information Player 2s have. The *homo economicus* model says we should expect "low", and invariant, rates of cooperative play by Player 2s across the treatments. Other-regarding preference

[12]The quiz and survey are available from the authors upon request.

models say we should expect low levels in PRIVATE, but high (and identical) proportions of Player 2s choosing R in both ASYMMETRIC and COMMON. The strong reciprocity hypothesis is that altruists (conditional or otherwise) do not exhaust the population of cooperative types. So models of strong reciprocity want to predict an increase for each stage as we move from PRIVATE to ASYMMETRIC to COMMON. Assume that, in this context, "low" correspond to rates below 5%, "moderate" to rates in the 20% – 25% range, and "high" to rates above 45%. Then we can graph the mean rates of observed cooperative play by Player 2s in our treatments and lay this on top of the qualitative predictions in Figure 2 to get Figure 3. And doing that — even before conducting more statistical tests — makes it pretty clear that neither *homo economicus* models nor other-regarding preference models can cover the data.

Table 1 reports the number of Player 2s who choose off-equilibrium path cooperation (R), the number of pairs ending up at the cooperative outcome (x_2), and the number of observations for each treatment. The rate of R-play in PRIVATE is statistically lower than in ASYMMETRIC (p-value= 0.0236).[13] This is evidence against any model — in particular, against the *homo economicus* model — that predicts that $\Pr(R|\text{PRIVATE}) = \Pr(R|\text{ASYMMETRIC})$. The data instead conform to what we would expect given an other-regarding preference model or a strong reciprocity model. The rate of R-play in ASYMMETRIC is statistically lower than in COMMON (p-value= 0.0124). This is evidence against any model that predicts — as other-regarding preference models do — that $\Pr(R|\text{ASYMMETRIC}) = \Pr(R|\text{COMMON})$. The data instead conform to what we would expect given a strong reciprocity model. That is, compatible with the strong reciprocity hypothesis, we see evidence that $\Pr(R|\text{PRIVATE})$ is lower than $\Pr(R|\text{ASYMMETRIC})$ and that $\Pr(R|\text{ASYMMETRIC})$ is lower than $\Pr(R|\text{COMMON})$.

This is confirmed by logitistic regression analysis. All of the relevant predictions from the social preference models are pairwise predictions, comparing Player 2 cooperation between PRIVATE and ASYMMETRIC and comparing Player 2 cooperation between ASYMMETRIC and COMMON. So, similarly, we want to compare data in this pairwise way, estimating the following simple model:

$$\text{Coop} = \beta \times \text{Treatment}$$

[13] The results reported here are based on difference in proportion tests, unless otherwise noted.

Cooperation Rates By Treatment			
	PRIVATE	ASYMMETRIC	COMMON
P2s Choosing R (%)	3.33	22.00	46.51
Pairs Reaching x_2 (%)	3.33	8.00	32.56
# of Pairs	30	50	43

Difference in Proportions Tests		
	$\Pr(R\|\text{PRIV}) \ll \Pr(R\|\text{ASYM})$	$\Pr(R\|\text{ASYM}) \ll \Pr(R\|\text{COMM})$
P2s Choosing R	Yes*	Yes**
Pairs Reaching x_2	No	Yes***

p-values: * ≤ 0.05, ** ≤ .01, *** ≤ .001

Table 1: Cooperation Rates

where

$$\text{Coop} = \begin{cases} 1 \text{ if Player 2 chose } R \\ 0 \text{ otherwise} \end{cases}$$

We estimate the above regression twice: once when making the pairwise comparison between $\Pr(R|\text{PRIVATE})$ and $\Pr(R|\text{ASYMMETRIC})$. In that case we set Treatment = ASYMMETRIC and drop data from COMMON. And we estimate it again when making the pairwise comparison between $\Pr(R|\text{ASYMMETRIC})$ and $\Pr(R|\text{COMMON})$. In that case we set Treatment = COMMON and drop data from PRIVATE.

	$\Pr(R\|\text{PRIV}) \ll \Pr(R\|\text{ASYM})$	$\Pr(R\|\text{ASYM}) \ll \Pr(R\|\text{COMM})$
Treatment	0.1222*	0.3244**
	(0.1312)	(.1487)
N	80	93
pseudo R^2	0.0913	.0532

std. errors in parentheses; p-values: * ≤ 0.05, ** ≤ .01, *** ≤ .001

Table 2: Player 2 Cooperation Pairwise Logits

The results of the analysis are summarized in Table 2. The first column reports that the estimation of the net odds of R play by a Player 2 in ASYMMETRIC are about 12% greater than those under PRIVATE. This is evidence against any model that predicts constant cooperation rates between PRIVATE and ASYMMET-

RIC. The second column reports that the estimation of the net odds of R play by a Player 2 in COMMON are about 32% greater than those under ASYMMETRIC. This is evidence against any model — and so against the class of other-regarding preference models — that predicts constant cooperation rates between ASYMMETRIC and COMMON. The strong reciprocity hypothesis is compatible with both odds estimates. To summarize:

Result (Strong Reciprocity)**.**

1. Player 2 observed behavior satisfies both inequalities predicted by strong reciprocity models:
   a) $\Pr(R|\text{PRIVATE}) < \Pr(R|\text{ASYMMETRIC})$
   b) $\Pr(R|\text{ASYMMETRIC}) < \Pr(R|\text{COMMON})$
2. Moreover, the cooperation rates are low in PRIVATE, moderate in ASYMMETRIC, and high only in COMMON

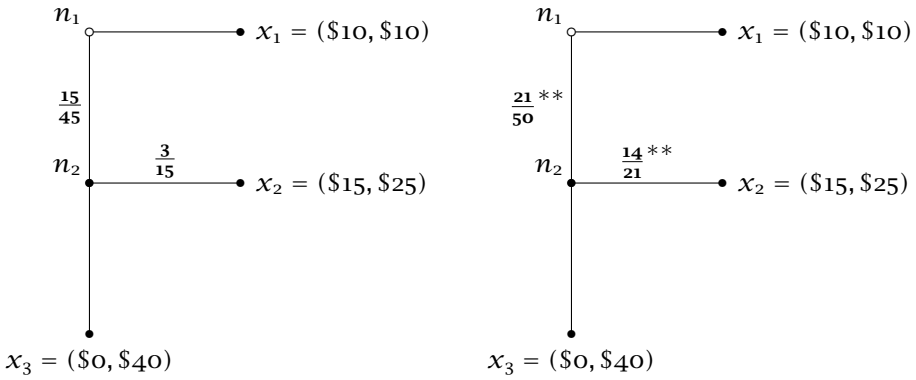
This is strong evidence in favor of the hypothesis that opportunists and altruists do not exhaust the population; there are also strong reciprocators.

6 Other Results

Our main result raises some additional issues that are, strictly speaking, independent of the main question we have been pursuing. One issue is that it looks like payoff privacy encourages competitive, equlibrium-path play. Another issue is that our experiment used the strategy method to elicit choices while most other trust game experiments use the game method — is choice behavior in trust games sensitive to elicitation method? In this section we briefly discuss some additional results from a second set of treatments aimed at probing these issues further.

We added two additional treatments to our design:

- PRIVATE-GAME: Distribution of payoff information as in PRIVATE, but Player 2 offers a choice at a node n_2 only if, and only after being informed that, Player 1 chose d
- COMMON-GAME: Distribution of payoff information as in COMMON, but Player 2 offers a choice at a node n_2 only if, and only after being informed that, Player 1 chose d



(a) PRIVATE-GAME ($N = 45$) (b) COMMON-GAME ($N = 50$)

p-value: ** ≤ 0.01

Figure 4: Game Method Results

Figure 4 reports the basic results from these two additional treatments. The first thing to notice is that in PRIVATE-GAME there is little trust by Player 1s: only $\frac{15}{45}$ Player 2s (33.33%) even get a chance to choose; all others are funneled to the inefficient subgame perfect distribution (Figure 4a). This is in contrast to COMMON-GAME, in which $\frac{21}{50}$ Player 2s (42%) have a chance to determine the gains from exchange (Figure 4b). This difference in rates of trusting d play is significant (p-value 0.011). Similarly, Player 2s choose R in only $\frac{3}{15}$ cases (20%) in PRIVATE-GAME while in COMMON-GAME $\frac{14}{21}$ Player 2s (66.67%) choose R. This difference is significant (p-value: 0.0057). These results are broadly consistent with those reported by McCabe, Rassenti, and Smith (1998) who find that payoff privacy in their trust/punishment "Game 1" funnels subjects into subgame perfect, but sub-optimal, distributions.[14]

Result (Payoff Privacy)**.** Cooperative choice behavior of Player 1s and Player 2s in COMMON-GAME resemble known results for binary-choice trust games. In PRIVATE-GAME both Player 1s and Player 2s are far more likely to play on the equilibrium path.

We began the paper by noting that, despite the temptation to think otherwise,

[14]This is also consistent with the early result in Smith (1962): he reports market convergence where private information of values, even with few traders, is sufficient to drive outcomes to competitive equilibrium.

Strategy Method		
	PRIVATE	COMMON
Player 1s Choosing d (%)	20.00	44.19
Player 2s Choosing R (%)	3.33	46.51
# of Pairs	30	43

Game Method		
	PRIVATE-GAME	COMMON-GAME
Player 1s Choosing d (%)	33.33	42.00
Player 2s Choosing R (%)	20.00**	66.67***
# of Pairs	45	50

p-values: * ≤ 0.05, ** ≤ .01, *** ≤ .001

Table 3: Cooperation Rates by Method

Nash equilibrium has little to do with common knowledge. There is some poetic justice, then, in the fact that the opportunism of Nash equilibrium does so well at predicting behavior in this game when we move away from common information to private information.

Clearly, as we see with such low numbers of Player 1s choosing d in PRIVATE-GAME, we had good reason to look for data on our comparative hypotheses using the strategy method. Even though we had good reason, and even though our main argument and result is independent of this issue, it is an interesting question whether our observed rates of cooperative play in the trust game are sensitive to the elicitation method. So we want to compare our data within each distribution of payoff information (private distributions, common distributions) across methods (strategy, game).

Some previous comparisons of elicitation method report little difference in choice behavior for some games (Brandts and Charness, 2000; Cason and Mui, 1998; Oxoby and McLeish, 2004). Our interest is in the binary-choice trust game and so the previous research most relevant is Casari and Cason (2008). They report results on elicitation method from an experiment using a trimmed-version of the investment game in which Player 1 must make a binary trust choice to invest all of his endowment or none of it; Player 2 then has the usual message space for dividing the gains from exchange. They find that Player 1 behavior is invariant across elicitation methods, but that Player 2 behavior is not — Player 2s display choice behavior significantly less trustworthy when

those choices are elicited by the strategy method compared to the game method. We see much the same thing.

The variable of interest here is elicitation method more than information distribution, so let's look first at COMMON/COMMON-GAME (Table 3). While there is no difference between rates of d play by Player 1s across methods, Player 2s choose R at higher rates in the game method: 46.51% compared to 66.67% (p-value = 0.0098).[15]

Result (Strategy Method and Strong Reciprocity)**.** Eliciting choices by strategy method reduces rates at which we observe choices consistent with strong reciprocity.

We see a somewhat different trend in PRIVATE/PRIVATE-GAME: Player 2s choices are reciprocal only 3.33% of the time when those choices are elicited by the strategy method compared to 20% in the game method (p-value = 0.0375). But this should be treated very cautiously: the difference in rates of R play is likely just residue from the fact that we have so few observations of Player 2 behavior in PRIVATE-GAME: only 15 Player 1s opt to move d (we have no idea why they do nor why any Player 2s would then choose R, but 3 of those 15 do). We conjecture that if the sessions were backed by a large enough supply of funds for subject payment to get a reasonable number of observations at Player 2s' decision node (in this case, "large enough" likely means unlimited), we would see no difference in either Player 1 behavior or in Player 2 behavior across methods under private distributions of payoff information.

7 Conclusion

We began with a fairly simple question: Given a social preference model, what information about possible distributions does Player 2 have to have in order for that model to explain off-equilibrium choice bahvior? Those epistemic conditions then serve as a way of segregating theories since different classes of models say that different information about possible distributions is relevant. That fact alone generates a simple design and set of hypotheses. Our main result suggests that forms of other-regarding preferences are not empricially adequate models. That holds quite generally for that entire class of models. Of course, this result does not confirm any *particular* strong reciprocity model; it

[15]This result is further supported by regression analysis, which we have omitted.

merely says that what we see is compatible with the minimal requirements of all such models.

It is useful to think of experimental tests as a means of segregating and falsifying models. But it is also useful to think more abstractly about what the explanatory force of the classes of models is. If the *homo economicus* model explained observed behavior, then that would say something important about the distribution of types of agents in the population: that the opportunists exhaust the population. Similarly, if other-regarding preference models really could explain all that needed explaining, then that would say something important about the distribution of types of agents in the population: that the opportunists and altruists exhaust the population. When we interpret our main result at this level, it thus says something important: opportunists and altruists — conditional or otherwise — do not exhaust the population; there are also strong reciprocators in our midst.

Bibliography

Ashraf, N. I. Bohnet and N. Piankov (2006). “Decomposing Trust and Trustworthiness," *Experimental Economics* **9**:193–208.

Aumann, R. and A. Brandenburger (1995). “Epistemic Conditions for Nash Equilibrium," *Econometrica* **63**: 1161–1180. Reprinted in: R. J. Aumann, *Collected Papers, vol. 1* (Cambridge, MA: MIT Press, 2000).

Berg, J., J. Dickhaut, and K. McCabe (1995). “Trust, Reciprocity, and Social History," *Games and Economic Behavior* **10**: 122–142.

Bohnet, I. and R. Zeckhauser (2004). “Trust, Risk, and Betrayal," *Journal of Economic Behavior and Organization* **55**(4): 464–484.

Bolton, G. and A. Ockenfels (2000). “ERC: A Theory of Equity, Reciprocity, and Competition" *American Economic Review* **90**(1): 166–193.

Bowles, S. and H. Gintis (2003). “Origins of Human Cooperation," in P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation* (Cambridge, MA: MIT Press), pp. 429–443.

Brandts, J. and G. Charness (2000). “Hot vs. Cold: Sequential Responses and Preference Stability in Experimental Games," *Experimental Economics* **2**: 227–238.

Camerer, C. F. (2003). *Behavioral Game Theory* (Princeton, NJ: Princeton University Press).

Casari, M. and T. N. Cason (2008). “The Strategy Method Biases Measured Trustworthy Behavior," Working Paper, Purdue University.

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

Charness, G. and M. Rabin (2002). “Understanding Social Preferences With Simple Tests," *The Quarterly Journal of Economics* **117**(3): 817–869.

Cox, J. C., D. Friedman and V. Sadiraj (in press). “Revealed Altruism," *Econometrica* **76**(1): 31–69.

Dufwenberg, M. and G. Kirchsteiger (2004). “A Theory of Sequential Reciprocity," *Games and Economic Behavior* **47**(2): 268–298.

Falk, A. and U. Fischbacher (2005). “Modeling Strong Reciprocity," in H. Gintis, S. Bowles, R. Boyd, and E. Fehr (eds.) *Moral Sentiments and Material Interests* (Cambridge, MA: MIT Press).

Falk, A., E. Fehr, and U. Fischbacher (2003). “On the Nature of Fair Behavior," *Economic Inquiry* **41**(1), 20–26.

Fehr, E. and U. Fischbacher (2005). “The Economics of Strong Reciprocity," in H. Gintis, S. Bowles, R. Boyd, and E. Fehr (eds.) *Moral Sentiments and Material Interests* (Cambridge, MA: MIT Press).

Fehr, E. and K. M. Schmidt (1999). “A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics* **114**(3): 817–868.

Gintis, H. (2000). “Strong Reciprocity and Human Sociality,” *Journal of Theoretical Biology* **206**: 169–179.

Güth, W., Schmittberger, and Schwarze (1982). “An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* **3**(4): 367–388.

Houser, D. and E. Xiao (2003). “Emotion Expression in Human Punishment Behavior," *Proceedings of the National Academy of Sciences 102*(20): 7398–7401.

Levine, D. (1998). “Modeling Altruism and Spitefulness in Experiments,” *Review of Economics Dynamics* **1**: 591–622.

McCabe, K., S. Rassenti, and V. Smith (1996). “Reciprocity, Trust, and Payoff Privacy in Extensive Form Bargaining," *Games and Economic Behavior* **24**: 10–24.

McCabe, K., M. Rigdon, and V. Smith (2002). “Cooperation in Single Play, Two-Person Extensive Form Games between Anonymously Matched Players," in R. Zwick and A. Rapoport (eds.), *Experimental Business Research*, (Boston, MA: Kluwer), pp.49–67.

McCabe, K., M. Rigdon, and V. Smith (2003). “Positive Reciprocity and Intentions in Trust Games," *Journal of Economic Behavior and Organization* **52**(2): 267–275.

McCabe, K. and V. Smith (2000a). "A Two-person Trust Game Played by Naïve and Sophisticated Subjects," *Proceedings of the National Academy of Sciences* **97**(7): 3777–3781.

McCabe, K., Smith, V., LePore, M., 2000. Intentionality Detection and 'Mindreading': Why Does Game Form Matter? *Proceedings of the National Academy of Sciences* 97(8), 4404–4409.

McCabe, K. and V. Smith (2000b). "Goodwill Accounting in Economic Exchange," in G. Gigerenzer and R. Selten (eds.), *Bounded Rationality: The Adaptive Toolbox* (Cambridge, MA: MIT Press).

Oxoby, R. and K. McLeish (2004). "Sequential Decision and Strategy Vector Methods in Ultimatum Bargaining: Evidence on the Strength of Other-regarding Behavior," *Economics Letters* **84**: 399–405.

Ortmann, A., J. Fitzgerald, and C. Boeing (2000). "Trust, Reciprocity, and Social History: A Re-examination," *Experimental Economics* **3**(1): 81–100.

Oosterbeek, H., R. Sloof, and G. van de Kuilen (2004). "Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis," *Experimental Economics* **7**: 171–188.

Rigdon, M. (2008). "Trust and Reciprocity in Incentive Contracting," *Journal of Economic Behavior and Organization*, in press.

Smith, V. (2008). *Rationality in Economics: Constructivist and Ecological Forms* New York, NY: Cambridge University Press.

Smith, V. (1962). "An Experimental Study of Competitive Market Behavior," *The Journal of Political Economy* **70**(2): 111–137.

Solnick, S. (2007). "Cash and Alternate Methods of Accounting in an Experimental Game," *Journal of Economic Behavior and Organization* **62**: 316–321.