SUSTAINING COOPERATION IN TRUST GAMES

KEVIN A. MCCABE, MARY L. RIGDON* AND VERNON L. SMITH

Abstract. It is well-known in evolutionary game theory that population clustering in Prisoner Dilemma games allows some cooperative strategies to invade populations of stable defecting strategies. We adapt this idea of population clustering to a two-person trust game. Players are typed based on their recent track record as whether or not they are trusting (Players 1) and whether or not they are trustworthy (Players 2). They are then paired according to those types: trustors with trustworthy types, and similarly non-trustors with untrustworthy types. The empirical question we address is whether this adaptation of clustering to bargaining environments sustains cooperative play analogous to the situation in finitely repeated PD games.

JEL Classification: C72, C91
Keywords: exchange, trust, reciprocity, cooperation, clustering, bargaining, experimental economics

Date: October 6, 2003.
*Corresponding author: Rigdon. Rigdon is at the Center for Basic Research in the Social Sciences, Harvard University, 34 Kirkland St., Cambridge, MA 02138 (mrigdon@latte.harvard.edu); McCabe and Smith are at the Interdisciplinary Center for Economic Science, George Mason University, 4400 University Dr.; MSN 1B2, Fairfax, VA 22030 ({kmccabe, vsmith2}@gmu.edu). A version of this paper appeared as Chapter 3 in Rigdon (2001). This work has been supported by NSF Grant No. SBR9510919 and by the Russell Sage Foundation. We would like to thank Sheryl Ball, Anthony Gillies, Glenn Harrison, Daniel Houser, Preston McAfee, Andreas Ortmann, Stergios Skaperdas, Brian Skyrms, Dale Stahl, the participants at the Economic Science Association meetings in Tucson (November 2001), and the participants at the American Economic Association session on Trust and Reciprocity in Games held during the Allied Social Science Association meetings in Atlanta (January 2002) for discussion and comments. All errors remain our own.

1. Introduction

There are two related problems of cooperation in bargaining environments. The first problem is to explain why and how people bargain their way to Pareto efficient, off-equilibrium path outcomes. This problem has received considerable attention in the recent literature (Guth, *et al.*, 1982; BDMc, 1995; Roth, 1995; Fehr and Gächter, 2000; McCabe, *et al.*, 2001). The second problem is to say how cooperation can be sustained once it emerges. The second problem has received comparably less attention than the first.

Even though sustaining cooperation has received less attention in bargaining situations, it has been a primary focus in Prisoner's Dilemma (PD) and public good games (Andreoni and Miller, 1993; Andreoni and Varian, 1999; Axelrod, 1984, 1997; Bohnet and Kübler, 2003; Kreps, *et al.*, 1982; Ledyard, 1995). Consider the analysis of the finitely repeated PD game in Axelrod (1984). In this game, always defecting is an evolutionary stable strategy (ESS) in the sense that it does not pay to cooperate in a population where everyone else always defects. Yet a small band of conditional cooperators (say, tit-for-tat players) can invade a population of unconditional defectors provided that the cooperators can *cluster*. That is, if these cooperators interact more often with each other than with the defectors (or if the result of two cooperators meeting is advantageous enough), then the population can be invaded. For clustering to work, though, it must be the case that the probability of two members of the relevant subpopulation meeting is not the same as the probability of two arbitrary members of the population at large meeting. The problem in populations without clustering is that the chance of members from a small band of conditional cooperators meeting each other is comparatively low.

We want to adapt this idea of population clustering to a simple two-person trust game. The clustering in our trust game will be a function of recent behavior in this bargaining environment. An agent's history of choices gives him a track record. Players can be typed based on their recent track record as whether or not they are trusting (for Players 1), and whether or not they are trustworthy (for Players 2). Once the players are typed, they can then be paired according to those types: trustors with trustworthy types, and similarly non-trustors with untrustworthy types. If some people are inclined to trust, this sort of matching protocol will induce self-selection clustering within the population. The empirical question that we want to address is whether this adaptation of clustering to bargaining environments can sustain cooperative play

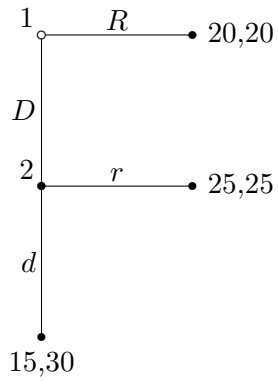

FIGURE 1. Trust game

analogous to the situation in finitely repeated PD games. That is, if cooperative play emerges in the trust game, can the level be maintained via an endogenous matching rule? This paper studies the effect of an experimental treatment controlling for the history of cooperation by procedures unknown to the subjects so that cooperation is not sustained by common knowledge and expectations about the particular clustering mechanism in the population.

In the next section we describe a two-person trust game and our mechanism for clustering the population. We then discuss the design and procedures used in our experiments (Section 3). Data analysis follows in Section 4 and concluding remarks are contained in Section 5.

2. Sustaining Trust

In the trust game pictured in Figure 1, Player 1 is asked to choose from the following: (1) You are given \$40, which you can split evenly with another person—Player 2—in which case the game is over or (2) You present Player 2 with two choices, either Player 2 can take \$30 out of \$45, leaving you \$15; or she can split \$50 evenly between the two of you.

A standard backward induction argument verifies that the unique subgame perfect equilibrium (SPE) of this game is the $(20, 20)$ outcome determined by the pure strategy (R, d): a rational Player 2 would strictly prefer \$30 to \$25, and so would choose *down* (d) at her decision node; knowing this a rational Player 1, who prefers \$20 to \$15, would therefore choose *Right* (R) at his decision node.

Although the pure strategy profile (R, d) is the unique SPE, it is not an ESS: (R, d) is a Nash equilibrium but it is not strict, and thus cannot be an ESS (Weibull, 1995, Proposition 5.1). Intuitively, the situation is this. Consider a population in which Players 1 all play R and Players 2 all play d. Players 2 are susceptible to a certain amount of drift: a mutant Player 2 who would play r were she given the chance has the same fitness as a non-mutant Player 2. So selection pressures cannot rule out that such mutant Players 2 will thrive equally as well as their non-mutant peers. Consequently, a mutant Player 1 who plays D instead of R may well meet a mutant Player 2. If the proportion of mutant Players 2 is high enough, then such a Player 1 will achieve a higher level of fitness than his non-mutant peers, namely a payoff of 20. And so it cannot be ruled out that the population of (R, d) players will be destabilized due to the drift of Player 2 and subsequent mutation of Player 1.

This raises the empirical question with which we are concerned. Since (R, d) is not an ESS we know that it is *possible* for cooperation to emerge in this environment. What we want to know is what the behavioral and institutional preconditions are for such cooperation to actually emerge and be sustained. This is an empirical question. In particular, is the mere possibility of random drift enough to allow cooperation to emerge and be sustained, or can the level of cooperation and its stability be encouraged by population clustering?

We implement the idea of clustering by typing players based on their observed moves in the trust game above, and (in one treatment) match players based on their types. Types come in the form of a "trust score", τ_n^j, where j= 1 or 2 for player role, n indicates the round, and $\tau \in [0, 1]$. The idea is that each player will have a score which is updated each round and is essentially the relative frequency of the number of times the player cooperated to the number of chances the player has had to cooperate. τ is defined algorithmically. At the end of each period, the algorithm begins by looping through the decisions made by all the Players 1 and calculating their respective score and then does the same for all the Players 2. A move by Player i is a defection move just in case it is i's strategy in the subgame perfect strategy profile. A move by Player i is a cooperative move just in case it is not a defection move. We assume that both Player 1 and Player 2 begin with a trust score of zero.

Algorithm 1 (Player 1 Trust Score)**.** Let c_1 (d_1) indicate a cooperative (defection) move by Player 1. Then the trust score of a Player 1 after Round n, τ_n^1, is given by the following algorithm:

(1) If $n = 0$: $\tau_0^1 = 0$

(2) If $n \leq 5$: Let k be the number of c_1 moves through Round $n-1$. Then:

$$\tau_n^1 = \begin{cases} \frac{k}{n} & \text{if} \quad d_1 \text{ in Round } n \\ \frac{k+1}{n} & \text{if} \quad c_1 \text{ in Round } n \end{cases}$$

(3) If $n > 5$: Let k be the number of c_1 moves in Rounds $n-1, \ldots, n-4$. Then:

$$\tau_n^1 = \begin{cases} \frac{k}{5} & \text{if} \quad d_1 \text{ in Round } n \\ \frac{k+1}{5} & \text{if} \quad c_1 \text{ in Round } n \end{cases}$$

That the divisor (when $n > 5$) is always 5 puts a premium on the last five interactions of the players. Pre-theoretically, there is a recency effect of goodwill—recent acts of goodwill overshadow distant acts of ill-will and vice versa. The trust score algorithm for Player 1 codifies this intuition by only keeping track of the behavior over the most recent five rounds.

To compute the trust score of a Player 2 after Round n, we need to first compute the number of times that Player 2 has had an opportunity to make a choice—the idea being that her trust score should neither be incremented nor decremented in cases where Player 1 chooses his outside option.[1] This will be recorded as Player 2's *opportunity score*. We need to make a similar allowance to codify the recency effect of trust and trustworthiness. Instead of tracking the behavior of Player 2 (for the purposes of computing her trust score) over the most recent five rounds, we need instead track it over the most recent five rounds *in which she had an opportunity*

[1]Why would one adopt a prior that observing Player 1 defect would not affect Player 2's cooperative propensities? One might indeed think that a Player 2's trust score should be decremented in cases where Player 1 chooses his outside option; the idea being that Player 2's cooperative propensity, in such cases, gets infected by the observation of non-cooperative play by Player 1. Whether or not some portion of the population reacts in this way is an empirical question. But even assuming this view is correct, the result of using our trust score algorithms (which are not sensitive to this posited behavior) in the matching experiments would be that some Players 2 have an artificially high trust score. Thus, when matched according to trust scores, some such Players 2 may be matched with (real) trusting Players 1. But notice that this would make the observation of sustained cooperative play rather more difficult to achieve. Hence, if the experimental results indicate such sustained cooperative behavior even using our scoring algorithms, then those results should be thought of as rather robust.

to make a decision. We simply need to verify if Player 1 moved down (right), in which case Player 2's opportunity score is (not) incremented. We will call this queue her *omega queue.*

Algorithm 2 (Player 2 Opportunity Score, Omega Queue)**.** Let c_1 (d_1) indicate a cooperative (defection) move by Player 1, and let c_2 (d_2) indicate a cooperative (defection) move by Player 2. Then Player 2's *opportunity score in Round n*, ρ_n, is given by the following algorithm:

(1) If $n = 0$: $\rho_0 = 0$

(2) If $n \geq 1$:

$$\rho_n = \begin{cases} \rho_{n-1} & \text{if} \quad d_1 \text{ in Round } n \\ \rho_{n-1} + 1 & \text{if} \quad c_1 \text{ in Round } n \end{cases}$$

Where $n \geq 5$, let Ω_{n-1} be the four most recent rounds prior to Round n in which Player 2 has had a chance to move.

Player 2's trust score is very similar to Player 1's, but the denominator becomes opportunity score, rather than round.

Algorithm 3 (Player 2 Trust Score)**.** Let c_2 (d_2) indicate a cooperative (defection) move by Player 2. Then the trust score of a Player 2 after Round n, τ_n^2, is given by the following algorithm:

(1) If $n = 0$: $\tau_0^2 = 0$

(2) If $\rho_n = \rho_{n-1}$: $\tau_n^2 = \tau_{n-1}^2$

(3) If $\rho_n \neq \rho_{n-1}$, $\rho_n \leq 5$, and $n \leq 5$: Let k be the number of c_2 moves through Round $n-1$. Then:

$$\tau_n^2 = \begin{cases} \frac{k}{\rho_n} & \text{if} \quad d_2 \text{ in Round } n \\ \frac{k+1}{\rho_n} & \text{if} \quad c_2 \text{ in Round } n \end{cases}$$

(4) If $\rho_n \neq \rho_{n-1}$, $\rho_n \leq 5$, and $n > 5$: Let k be the number of c_2 moves in Ω_{n-1}. Then:

$$\tau_n^2 = \begin{cases} \frac{k}{\rho_n} & \text{if} \quad d_2 \text{ in Round } n \\ \frac{k+1}{\rho_n} & \text{if} \quad c_2 \text{ in Round } n \end{cases}$$

(5) If $\rho_n \neq \rho_{n-1}$ and $\rho_n \geq 5$: Let k be the number of c_2 moves in Ω_{n-1}. Then:

$$\tau_n^2 = \begin{cases} \frac{k}{5} & \text{if} \quad d_2 \text{ in Round } n \\ \frac{k+1}{5} & \text{if} \quad c_2 \text{ in Round } n \end{cases}$$

So at the end of each round, each player has a trust score, which essentially tracks the relative frequency of cooperative moves up to that round.

The two treatments reported below differ according to their *matching protocol*. In the baseline condition (the Random treatment) subjects are *randomly* paired each period. Trust scores in the Random treatment are tracked, but not used in matching Players 1 and Players 2. The experimental treatment (the Sorted treatment) *pairs subjects according to their trust scores*. The matching protocol for the Sorted treatment is straightforward: At the end of Round n Players 1 are rank-ordered by their trust scores (high to low). Similarly for Players 2. Then the matching rule simply pairs the highest ranked Player 1 with the highest ranked Player 2 for interaction in Round $n+1$, the next to highest ranked Player 1 with the next to highest ranked Player 2 for interaction in Round $n+1$, and so on.[2]

3. Experimental Design and Procedures

Our experiments were conducted with undergraduate students from a variety of majors at The University of Arizona. A total of eight experimental sessions were run: four sessions of the Sorted treatment and four sessions of the Random treatment.[3] Each experimental session consisted of 16 subjects.[4]

A subject is paid \$5 for showing up on time and immediately (and randomly) seated at a computer terminal in a large room containing 40 terminals. Each terminal is in a separate cubicle, and the subjects are dispersed so that no subject can see the terminal screen of another. Each person is randomly assigned a role (Player 1 or 2) and keeps this role for the entirety of the experiment. The instructions for each experiment do not use words like 'game', 'play', 'player', 'opponent', 'partner, 'trust', etc.; rather neutral terms such as 'decision problem', 'decision maker 1 (DM1)', 'DM2', 'your counterpart', etc. are used in order to provide a baseline context.

The interactions in the experiment consist of anonymous pairings in a computerized game. By using a mouse, each Player 1 can click on the right or down arrows. A player confirms his choice by clicking on a "Send" button. This move information is then displayed on their counterpart's screen. If Player 1 moves down, Player 2 would be prompted to click on the right or down arrow (again confirming her choice by clicking on a "Send" button). This move information is then

[2]Ties in trust scores are broken randomly.

[3]In order to control for some variability we ran all of the sessions at the same time of day, taking two weeks to complete.

[4]Two randomized treatments only had 14 subjects due to no shows.

Sorted	4/64/1280*
Random	4/60/1200

*$a/b/c$ where a = number of sessions, b = number of subjects, c = number of observations.

TABLE 1. Experimental Design

displayed on Player 1's screen. Earnings are shown to both Player 1 and Player 2 after each period. The game is sequential in structure—i.e. we do not employ the strategy method to elicit choices. Subjects respond to actual move information when making a decision.

The payoffs represent the experimental dollar amounts the subjects could earn with an exchange rate of 20 experimental dollars equal to 1 U.S. dollar; both the payoffs and the exchange rate are common information. The games were played sequentially for 20 periods, although the subjects do not know the total number of periods until the session is complete.[5] At the end of the experiment, their accumulated earnings were paid to them privately (single-blind protocol). The experiments lasted on average a little under one hour, from arrival to completion. Subjects' earnings (not including the show-up fee) average \$21.00 ($s = 1.8$) in the Random treatment and \$23.00 ($s = 2.1$) in the Sorted treatment. The subjects did not have prior experience with this environment or others like it. Each subject participated in one and only one such experiment. See Table 1 for a summary of the experimental design.

The instructions stated the following about matching (see Appendix A for detailed instructions): "*Each period you will be paired with another individual: your counterpart for that period. You will participate for several periods, being re-paired each period.*" We did not reveal the exact assignment rule to any of the subjects because we were concerned that such information might generate a difference in strategic behavior.[6] This is especially the case in the Sorted environment—knowing that cooperators are being matched each period might lead individuals to alter their type for strategic reasons rather than due to reciprocity type motives.

Anonymously matched subjects in a single play trust game have a strong incentive to choose dominant strategies and to expect the same of their counterpart. They have no knowledge of

[5]The subjects did know that they were recruited for a one-hour experiment.

[6]Gunnthorsdottir, *et al.* (2001) faced similar considerations when sorting subjects in a public goods game based on their level of contribution. They also do not tell the subjects the sorting rule.

the types with which they are paired, yet many subjects exhibit trusting/trustworthy behavior. Since they make more money than if they play non-cooperatively, they can hardly be said not to be rational. If such behavior is deeply ingrained in a subset of every sample of subjects, then the greater experience of reciprocity in repeat interaction the greater should be the use of such strategies by these subjects. The sorting protocol enables clustering to occur while controlling for the information that would allow clustering to be the deliberate, constructively rational choice of those who otherwise would choose non-cooperatively.

4. Results

Table 2 provides the conditional outcome frequencies by blocks of five trials for the Sorted and Random conditions. Note that in the first trial block (rounds 1–5) roughly half of the play occurs at the SPE in both treatments and about half of the cooperative ventures by Player 1 are reciprocated. There is not a statistically significant difference between either the amount of play which reaches the SPE ($p = 0.4691$) or the amount of play which reaches the efficient outcome ($p = 0.5775$).[7] By the second trial block, however, there are significant differences in the mean proportion of outcomes across treatments. This is most pronounced in the last trial block. When subjects are sorted based on their trust scores there are far fewer pairs ending up at the SPE; when subjects are sorted, more pairs reach the cooperative outcome than when they are randomly matched each round. Players 1 reach the SPE 46.88% of the time in the Sorted treatment as compared to 72.67% in the Random treatment ($p = 0.0088$). Furthermore, Players 2 who are paired with trusting Players 1 respond in kind in the Sorted treatment 83.53% of the time compared with 51.22% of the time in the Random treatment ($p = 0.0128$).[8] One question is how well trusting Players 1 do compared to playing the SPE outcome in both treatments. In the Sorted treatment, the expected value of trust based on the average frequencies of cooperation

[7] p-values being reported are from two sample t-tests examining whether or not the means in question are different, unless otherwise noted.

[8] It was interesting watching the results come in from these experiments. What was easy to observe is that by Round 10 in the Sorted treatment around half of the Players 1 were playing SPE, so their trust scores began deteriorating rapidly and about half were playing down, keeping their trust scores near the maximum. Most of the trusting interactions were met with trustworthiness by their counterpart, keeping more than half of the Players 2 trust scores high as well. This was not the case in the Random treatment.

Trials	$(20, 20)$	*Down*	$(25, 25)$	$(15, 30)$
Sorted				
1–5	$\frac{74}{160} = .4625$	$\frac{86}{160} = .5375$	$\frac{44}{86} = .5116$	$\frac{42}{86} = .4884$
6–10	$\frac{80}{160} = .50$	$\frac{80}{160} = .50$	$\frac{58}{80} = .725$	$\frac{22}{80} = .2750$
11–15	$\frac{77}{160} = .4813$	$\frac{83}{160} = .5188$	$\frac{70}{83} = .8434$	$\frac{13}{83} = .1566$
16–20	$\frac{75}{160} = .4688$	$\frac{85}{160} = .5313$	$\frac{71}{85} = .8353$	$\frac{14}{85} = .1647$
Random				
1–5	$\frac{73}{150} = .4867$	$\frac{77}{150} = .5133$	$\frac{36}{77} = .4675$	$\frac{41}{77} = .5325$
6–10	$\frac{94}{150} = .6267$	$\frac{56}{150} = .3733$	$\frac{24}{56} = .4286$	$\frac{32}{56} = .5714$
11–15	$\frac{93}{150} = .62$	$\frac{57}{150} = .38$	$\frac{22}{57} = .386$	$\frac{35}{57} = .614$
16–20	$\frac{109}{150} = .7267$	$\frac{41}{150} = .2733$	$\frac{21}{41} = .5122$	$\frac{20}{41} = .4878$

TABLE 2. Conditional Outcomes by Trial Block

and defection moves by Players 2 is \$22.29.[9] In the Random treatment, the expected value of trust based on the average frequencies of cooperation and defection moves by Players 2 is \$19.49.[10] So in the Sorted treatment, it pays for the Players 1 to be trusting; this is not the case in the Random treatment.[11]

The above is aggregated in trial blocks. The dynamics of play over time reveals the same trends, albeit more graphically. Figures 2 and 3 show the mean fraction of each type of play over the 20 rounds for both treatment conditions. The trends are unmistakable: as play proceeds through the later rounds, cooperation emerges and is sustained among the sorted subjects, but there is no similar round-effect for the randomly paired subjects.

Along these same lines, it is interesting to look at the average trust score by blocks of five rounds (See Table 3). Remember that in both the Random and Sorted treatment a trust score

[9] $\text{EV}(trust|sorted) = 0.7288(\$25) + 0.2712(\$15) = \$22.288$.

[10] $\text{EV}(trust|random) = 0.4485(\$25) + 0.5515(\$15) = \$19.485$.

[11] The subjects did not know the end period in the experiment. Hence, one might conclude that the outcomes observed can be supported by the Folk Theorem. But this, by itself, cannot explain the data for at least two reasons. First, the Folk Theorem predicts far too many equilibria. Second, and more importantly, it cannot explain why we observe the *differences in the level* of cooperative behavior across the two treatments.

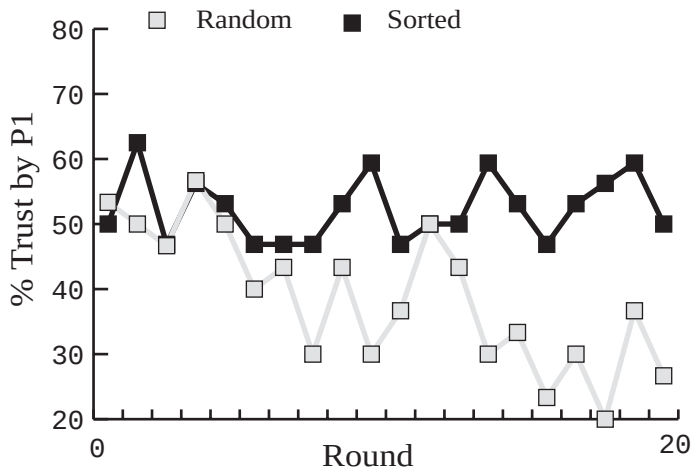

FIGURE 2. Percent of Players 1 Trusting Over Time

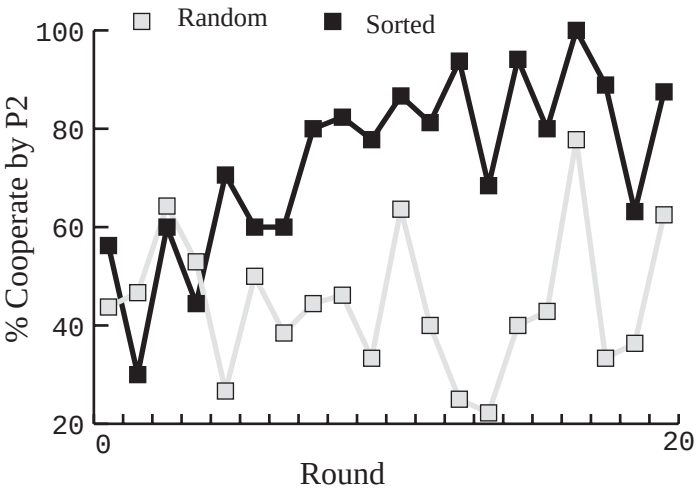

FIGURE 3. Percent of Players 2 Cooperating Over Time

is calculated for each player based on their decisions, but only the Sorted treatment matches players according to their score. Since the trust scores track the behavioral data, it is not surprising that an examination of the scores tells a very similar story to that of the outcome frequencies. The average trust score over the first 10 rounds is statistically the same for the two treatments ($p = .5246$ for 1–5 and $p = .1331$ for 6–10). However, in the last 10 rounds the trust

Trials	1–5	6–10	11–15	16–20
Mean (Sorted)	0.45	0.48	0.50*	0.52**
Mean (Random)	0.43	0.44	0.40	0.36
s **(Sorted)**	0.4030	0.3787	0.4337	0.4406
s **(Random)**	0.4165	0.3659	0.3531	0.3472

* p-value = 0.0008, ** p-value = .0000.

TABLE 3. Trust Score by Condition

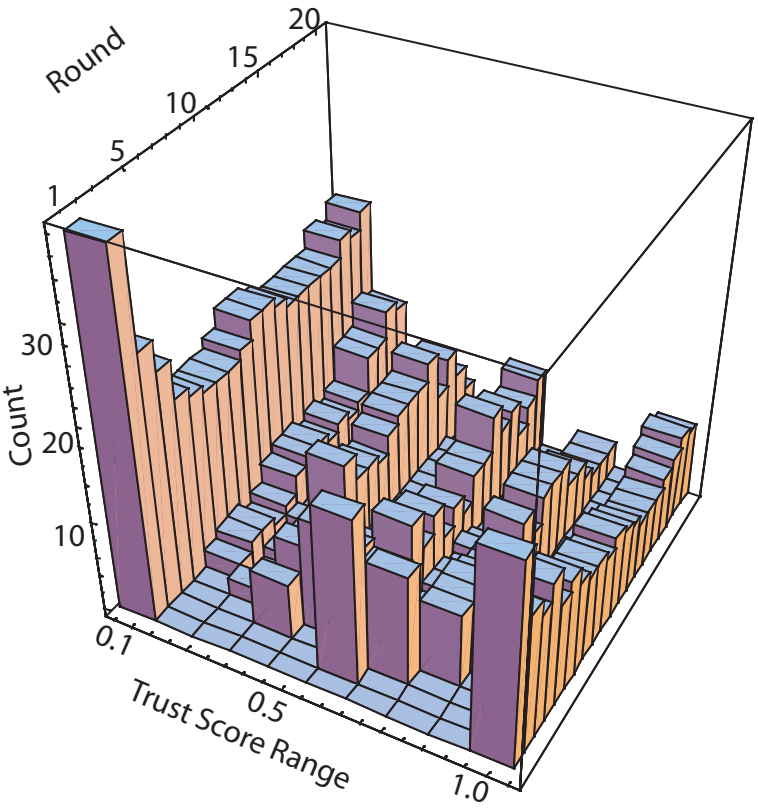

FIGURE 4. Trust Score Landscape: Random

scores are significantly higher under the Sorted condition than in the Random ($p = .0008$ for 11–15 and $p = .0000$ for 16–20).

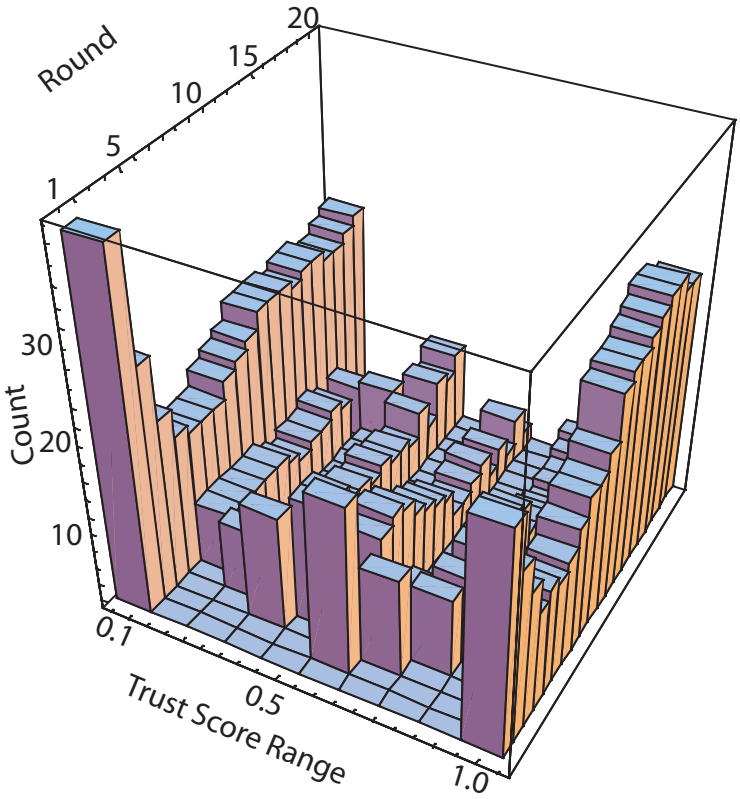

FIGURE 5. Trust Score Landscape: Sorted

One possible concern about our particular trust scoring algorithm is that given the sequential nature of the game, Players' 2 trust scores are slow to increment.[12] If this were the case, then the algorithm chosen could actually encourage the growth of cooperation by screening out large numbers of second movers from interacting in the environment. If these second movers are defectors, then this screening would prevent them from possibly infecting first movers who are cooperative. However, an examination of the data indicates that the first chance and the second

[12]One can think of employing other trust score algorithms where the initial value of trust scores is a neutral value of $\frac{1}{2}$, rather than 0. Such a rule may be more sensitive to what our default expectations about others' trust and trustworthiness tendencies might actually be. Our algorithms, on the other hand, seem rather to model the situation in Hobbes' "state of nature" in which the default expectation is that everyone is maximally untrusting/untrustworthy. Pre-theoretically, the emergence of sustained cooperative behavior under a sorting mechanism built on top of trust score algorithms like ours would seem rather unexpected.

	Sorted		Random	
	1st	**2nd**	**1st**	**2nd**
Mean	2.34	4.56	2	4
Median	1.5	3	1	3
Mode	1	2	1	3
s	2.91	4.3	1.34	2.04

TABLE 4. Descriptive Statistics for 1st and 2nd Chance to Move for Players 2 by Treatment

chance for a Player 2 to move occurs rather early in the experiment and these chances to move are not statistically different between the two treatments ($p = 0.28$ and $p = 0.26$, t-tests assuming unequal variances). Descriptive statistics are presented in Table 4 for both matching protocols.

Here we report logit regression results which examine Players 1 trust and Players 2 trustworthiness over time (see Table 5). The data columns for Trust shows the logit coefficient estimates and t-statistics for the regression of $\ln \frac{p(t)}{1-p(t)}$, $t = 1, 2, \ldots, 20$, where $p(t)$ is the probability a Player 1 is trusting in Round t, and $1 - p(t)$ is the probability that Player 1 will not be trusting (i.e., moves right). The independent variables are Round, which takes on the values of $1, 2, \ldots, 20$; Treat, a dummy variable with a value of 1 for the Sorted treatment, 0 otherwise; and Trust_{t-1}, a dummy variable with a value of 1 if Player 1 moved down in the previous round, 0 otherwise. The Round coefficient is negative, small and also not statistically significant. The highly significant coefficient for treatment indicates that the odds of trusting behavior by Players 1 is 40% higher in the Sorted treatment than the Random treatment. The largest effect on the odds of Player 1 trusting is whether Player 1 trusted last round, Trust_{t-1}, and this coefficient is highly significant.

The data columns for Trustworthy shows the logit coefficient estimates and t-statistics for the regression of $\ln \frac{p(t)}{1-p(t)}$, $t = 1, 2, \ldots, 20$, where $p(t)$ is the probability a Player 2 is trustworthy in Round t, and $1 - p(t)$ is the probability that Player 2 is not trustworthy (i.e., moves down). The independent variables are Round, which takes on the values of $1, 2, \ldots, 20$; Treat, a dummy variable with a value of 1 for the Sorted treatment, 0 otherwise; and Coop_{t-1}, a dummy variable with a value of 1 if Player 2 moved right in the previous round, 0 otherwise. The Round coefficient

	Trust			Trustworthy	
	Coeff.	t-stat		Coeff.	t-stat
Constant	-1.369	-8.056^*	Constant	-2.09	-11.69^*
Round	$-.0028$	$-.254$	Round	$-.0016$	$-.0134$
Treat	.407	3.138^*	Treat	.78	5.26^*
Trust_{t-1}	1.989	15.18^*	Coop_{t-1}	2.09	13.96^*

* p-values ≤ 0.001.

TABLE 5. Trust and Trustworthiness Logits

is negative, small and also not statistically significant. The highly significant coefficient for treatment indicates that the odds of trustworthy behavior by Players 2 is 78% higher in the Sorted treatment than the Random treatment. The largest effect on the odds of Player 2 being trustworthy is whether Player 2 was trustworthy last round, Coop_{t-1}, and this coefficient is highly significant.

In every interaction in this environment there is more than merely one's own monetary costs and benefits at stake. Each player's trust score is at stake in every interaction. Also at stake are the gains from exchange, and in particular we can think of whether or not the players actually achieve the efficient allocation. These are "social variables" in the sense that they are sensitive to more than just one's own payoffs and actions. There are two ways for players to incur social costs. If they defect then their trust scores are decremented; if they are defected upon then they incur the cost of being trusting when they ought not have been. Similarly, there are two ways to incur social gain. One is through making a cooperative choice; the other is when a pair of players actually reach the cooperative outcome. We would like to track how efficient choices are with respect to these potential social gains. We can introduce an *efficiency score* at round n for each player i, $\nu_n^i \in [0,1]$, as follows:

$$\nu_n^i = \frac{\tau_n^i + d}{2}$$

Trials	1–5	6–10	11–15	16–20
Mean (Sorted)	0.3624	0.4219	0.4698	0.4838
Mean (Random)	0.3344	0.2981	0.2738	0.2493
Mode (Sorted)	0	0	1	1
Mode (Random)	0	0	0	0
s **(Sorted)**	0.3802	0.3968	0.4507	0.4501
s **(Random)**	0.3725	0.3069	0.2911	0.2952

TABLE 6. Efficiency Score, ν, by Condition

Source of Variation	SS	DF	MS
Between	13.1225	1	13.1225
Within	350.19	2478	0.1416
Total about Grand Average	364.0415	2479	

TABLE 7. ANOVA for Efficiency Score ν

where $d = 0$ if Player i did not reach the cooperative outcome in Round n and $d = 1$ if she did. This variable tracks how efficient play is with respect to the potential social gains, in the sense described above, to be had from exchange.

Table 6 displays some descriptive statistics for the efficiency score variable under each condition in five trial blocks. Figure 6 plots the average efficiency score for both treatments at each round. The efficiency scores begin in Round 1 at less than 0.40 for both treatments and remain similar in magnitude through, roughly, the first nine rounds. However, in the later rounds, the efficiency being achieved in each condition is significantly different. In fact, the null hypothesis that there is no difference in the average efficiency score of the two treatments can be easily rejected (see Table 7; $p = .0000$). The level of efficiency being achieved is greater when subjects are being matched according to their trust score.

Is cooperation being "crowded out" in the Random treatment? That is, supposing that the initial propensity to cooperate among subjects is the same across treatments, then the fact that behavior tends toward high levels of repeated cooperative play in the Sorted treatment, and the

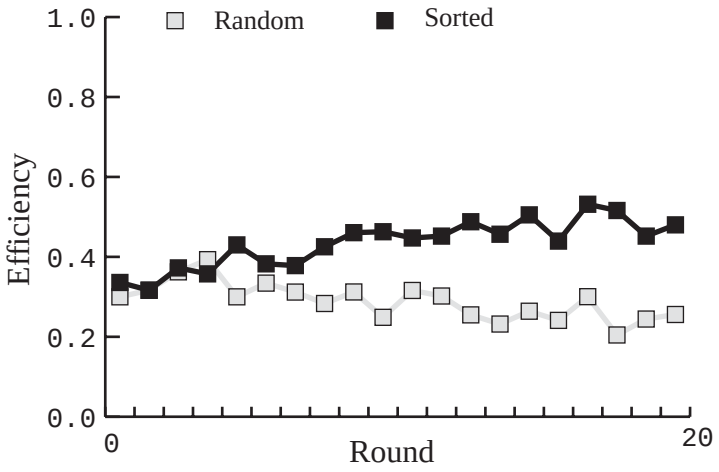

FIGURE 6. Average Efficiency Scores Over Time

	Random	**Sorted**
Coop	16/14*	17/17
NonCoop	14/16	15/15

*a/b where a = number of P1s, b = number of P2s.

TABLE 8. Distribution of Initial Player Types

fact that behavior tends toward subgame perfect play in the Random treatment would indeed be evidence that cooperative behavior is reinforced in the Sorted treatment and crowded out (or undermined) in the Random treatment. To examine this question, we can classify subjects as either a non-cooperator or cooperator based on their first observed move.[13] Players 1 are a non-cooperating type if in Round 1 they chose $(20, 20)$ and a cooperating type if they chose to play down, passing the game to their counterpart. Similarly, for Players 2. A Player 2 is a non-cooperating type if when her counterpart first played down, she chose the defection outcome $(15, 30)$, and a Player 2 is a cooperating type if she chose the cooperative outcome $(25, 25)$ on

[13]Basing type on only the first observed move attempts classification of agents according to their innate tendencies toward cooperation. The first observed move by a Player 2 occurs early: on average in the Sorted treatment after 2.34 rounds (median is 1.5 and mode is 1) and on average after 2 rounds (median is 1 and mode is 1) in the Random treatment.

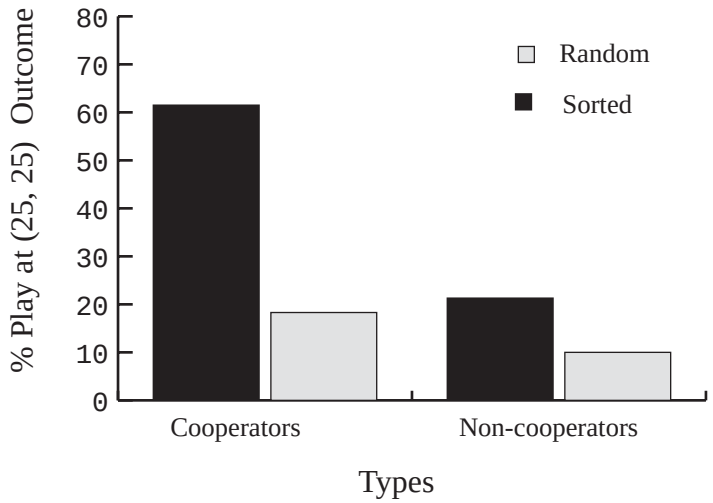

FIGURE 7. Cooperators versus Non-cooperators: Percent of Each Type Reaching the Cooperative Outcome of $(25, 25)$ in the Last 10 Rounds

her first available move. See Table 8 for the distribution of initial player types, in which rows indicate initial player types and columns indicate the matching protocol. Note that the initial distribution of player types is the same across treatments. Once we establish this typing, we can analyze how play differs among these groups depending on whether they are being sorted by their trust scores or simply being randomly re-paired. We want to focus on the last 10 rounds in particular (see Figure 7). Initial cooperators fare much better when they are meeting other cooperators under the sorting mechanism than when they randomly meet their counterparts—the last 10 interactions result in an outcome of $(25, 25)$ 62% of the time in the Sorted treatment compared to only 18% of the time in the Random treatment ($p = .0000$). This is not the case for initial non-cooperative types. In fact, there is no treatment effect for the defecting types: the percentage of cooperative outcomes reached in the last 10 rounds is not statistically different between the Random and Sorted treatments ($p = .1187$). This suggests that cooperation is crowded out in the Random treatment and fostered in the Sorted treatment.

In summary form, here are the five central results from this sorting experiment:

Result 1. In the last 10 rounds, the fraction of subjects reaching the SPE is dramatically lower in the Sorted treatment than in the Random treatment.

Result 2. In the last 10 rounds, the fraction of subjects reaching the cooperative outcome is significantly higher in the Sorted treatment than in the Random treatment.

Result 3. In the last 10 rounds, the average trust scores are much higher in the Sorted treatment than in the Random treatment.

Result 4. The average efficiency score, i.e. how efficient play is with respect to the potential social benefit, is higher in the Sorted treatment than in the Random treatment.

Result 5. In the last 10 rounds, the number of cooperative player types reaching cooperative outcomes is far greater in the Sorted treatment than in the Random treatment. There is no treatment effect for non-cooperative types.

5. Conclusions

It is well-known in evolutionary game theory that population clustering in PD games allows for some cooperative strategies to invade populations of stable defecting strategies. Similarly, in the experimental community there are results which suggest that a similar "clustering" phenomenon can be induced among subjects in public goods games to sustain high levels of contributions. Gunnthorsdottir, *et al.* (2000) match subjects in a standard public goods game based on their contribution level in the previous round, with the four highest contributors grouped, the next four highest grouped and so on. The sorting mechanism helps keep contribution levels high over time among initially cooperative types as compared to a random grouping. The results of the sorting experiments here suggest a similar story about behavior in simple two-person bargaining games. Although the SPE in our trust game is not an ESS, we find no behavioral evidence of significant cooperative play which can be attributed to random drift and mutation in the population. This is because in the Random treatment the level of efficient outcomes is low and initial cooperators seem to be crowded out of the environment. On the other hand, we do find evidence for a behavioral clustering phenomenon in this bargaining game. Sorting subjects by trust scores accomplishes two tasks. First, it allows cooperative play which is Pareto-superior to the SPE to emerge. Second, once cooperative play emerges, sorting subjects does not allow this behavior to be "infected" and compromised by either defecting Players 2 or by untrusting Players 1.

References

Ahn, T.K., Elinor Ostrom, David Schmidt, Robert Shupp, and James Walker (2001). "Cooperation in PD Games: Fear, Greed, and History of Play," *Public Choice* **106**: 137–155.

Andreoni, James and John H. Miller (1993). "Rational Cooperation in a Finitely Repeated Prisoner's Dilemma Game: Experimental Evidence," Economic Journal **103**(418): 570-585.

Andreoni, James and Hal Varian (1999). "Preplay Contracting in the Prisoners' Dilemma," *Proceedings of the National Academy of Sciences* **66**: 10933–10938.

Axelrod, Robert (1984). *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, Robert (1997). *The Complexity of Cooperation*. Princeton: Princeton University Press.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995). "Trust, Reciprocity, and Social History," *Games and Economic Behavior* **10**(1): 122–142.

Bohnet, Iris and Dorothea Kübler (2002). "Compensating the Cooperators: Is Sorting Possible in the Prisoner's Dilemma Game?," Working Paper, Harvard University.

Camerer, Colin and Mark Knez (2001). "Increasing Cooperation in Prisoner's Dilemmas by Establishing a Precedent of Efficiency in Coordination Games," *Organizational Behavior and Human Decision Processes*, forthcoming.

Charness, Gary (2000). "Bargaining Efficiency and Screening: An Experimental Investigation," *Journal of Economic Behavior and Organization* **42**: 285–304.

Fehr, Ernst and Simon Gächter (2000). "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives* **14**(3): 159–181.

Gunnthorsdottir, Anna, Daniel Houser, Kevin McCabe, and Holly Ameden (2000). "Disposition, History, and Contributions in Public Goods," Working Paper, George Mason University.

Guth, Werner, Rolf Schmittberger, and Bernd Schwarz (1982). "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* **3**: 367–388.

Harrison, Glenn W. (2001). "Experimental Behavior as an Algorithmic Progress: An Introduction," Working Paper, University of South Carolina.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith (1994). "Preferences, Property Rights, and Anonymity in Bargaining Games," *Games and Economic Behavior* **7**: 346–380.

Hoffman, E., K. McCabe, and V. Smith (1996). "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review* **86**: 653–660.

Kreps, David, Paul Milgrom, John Roberts, and Robert Wilson (1982). "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma Game," *Journal of Economic Theory* **17**: 245-252.

Ledyard, John O. (1995). "Public Goods," in J. Kagel and A. Roth (eds.), *The Handbook of Experimental Economics* (Princeton: Princeton University Press).

McCabe, Kevin, Stephen Rassenti, and Vernon Smith (1996). "Game Theory and Reciprocity in Some Extensive Form Experimental Games," *Proceedings of the National Academy of Sciences* **93**: 13421–13428.

McCabe, Kevin, Stephen Rassenti, and Vernon Smith (1998). "Reciprocity, Trust, and Payoff Privacy in Extensive Form Bargaining," *Games and Economic Behavior* **24**: 10–24.

McCabe, Kevin, Mary Rigdon, and Vernon Smith (2001). "Cooperation in Single Play, Two-Person Extensive Form Games between Anonymously Matched Players," in R. Zwick and A. Rapoport (eds.), *Experimental Business Research* pp.49–68 (Boston, MA: Kluwer).

Miller, John H. (1996). "The Coevolution of Automata in the Repeated Prisoner's Dilemma," *Journal of Economic Behavior and Organization* **29**: 87–112.

Nash, John F., Jr. (1950). "The Bargaining Problem," *Econometrica* **18**: 155–162. Reprinted in: H. W. Kuhn (ed.), *Classics in Game Theory* (Princeton: Princeton University Press, 1997).

Orbell, John M. and Robyn M. Dawes (1993). "Social Welfare, Cooperators' Advantage, and the Option of Not Playing the Game," *American Sociological Review* **58**: 787–800.

Rigdon, Mary L. (2001). *Cooperation and Intentions in Experimental Bargaining Games*. Ph.D. dissertation, Department of Economics: The University of Arizona.

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

Schmidt, David, Robert Shupp, James Walker, T. K. Ahn, and Elinor Ostrom (2001). "Dilemma Games: Game Parameters and Matching Protocols," *Journal of Economic Behavior and Organization* **46**: 357–377.

Selten, Rheinhard and Rolf Stoecker (1986). "End Behavior in Sequences of Finite Prisoner's Dilemma Supergames: A Learning Theory Approach," *Journal of Economic Behavior and*

Organization **7**: 47–70.

Silverstein, Albert, David Cross, Jay Brown, and Howard Rachlin (1998). "Prior Experience and Patterning in an Prisoner's Dilemma Game," *Journal of Behavioral Decision Making* **11**: 123–138.

Smith, Vernon L. (2003). "Experimental Methods in (Neuro)Economics," *Encyclopedia of Cognitive Science* (forthcoming).

Weibull, Jörgen W. (1995). *Evolutionary Game Theory*. Cambridge: MIT Press.

Wiseman, Thomas and Okan Yilankaya (2001). "Cooperation, Secret Handshakes, and Imitation in the Prisoners' Dilemma," *Games and Economic Behavior* **37**: 216–242.

Appendix A. Computerized Instructions for the both treatments

Page 1

In this experiment you will participate in a series of two person decision problems. The experiment will last for several periods. Each period you will be paired with another individual: your counterpart for that period. The joint decisions made by you and your counterpart for that period will determine how much money you will earn in that period. After each period you will be re-paired.

Your earnings will be paid to you in cash at the end of the experiment. We will not tell anyone else your earnings. We ask that you do not discuss your earnings with anyone else.

Please read the following instructions carefully. If you have a question at any time, please raise your hand and someone will come by to help.

Page 2

Notice that another button, "Back", has appeared at the bottom of the page. If at any time you wish to return to a previous page, click "Back". To continue reading the directions, click "Next".

Page 3

You will see a diagram similar to this one at the beginning of the experiment. You and another person will participate in a decision problem like the diagram below. We will refer to this other person as your counterpart.

SCREEN DIAGRAM

One of you will be DM 1. The other person will be DM 2. Beside the diagram we show whether you are DM 1 or DM 2. In this example, for now, you are DM 1. Please click "Next" to continue.

Page 4

Notice the boxes with letters in them. These letters will be replaced by numbers representing Experimental Dollars during the experiment. For 20 Experimental Dollars you will earn 1 U.S. dollar. The boxes with numbers show the different earnings in Experimental Dollars that you and your counterpart can make. There are two numbers in each box. The number on the top

(which is indented now) is DM 1's earnings if this box is reached. The number on the bottom is DM 2's earnings.

SCREEN DIAGRAM

You and your counterpart will jointly determine a path through the diagram to an earnings box. Please click "Next" to continue.

Page 5

A path is defined as sequence of moves through the diagram.

A move is a choice of direction in the diagram.

SCREEN DIAGRAM

The arrows in the diagram show the possible directions of moves that can be made. Notice that the moves for both DM 1 and DM 2 are always DOWN or RIGHT. When you click on either arrow, the path is highlighted.

The circles in the diagram with numbers in them indicate who gets to move at that point in the diagram. Please click "Next" to continue.

Page 6

For example, DM 1 starts the process at the top of the diagram by moving right or down. If DM 1 moves right the experiment is over. DM 1 earns 'zig' and DM 2 earns 'zog'.

SCREEN DIAGRAM

If DM 1 moves down, it is DM 2's turn to move. DM 2 can move right or down. If DM 2 moves right, DM 1 earns 'wig' and DM 2 earns 'wog'. If DM 2 moves down, DM 1 earns 'xig' and DM 2 earns 'xog'.

The decision path that was chosen will be highlighted. Please click "Next" to continue.

Page 7

We will now show you what the decisions look like from the point of view of DM 1. When you are DM 1 you move first. The arrows show you can move right or down. In order to move, click on the arrow for your choice. DM 2 will only see your decision when you click the "Send"

button to finalize your decision. To see how this works, click the RIGHT ARROW now. Be sure to click "Send" to finalize your move.

SCREEN DIAGRAM

At this point the moves are over. The path taken is highlighted white and earnings received are highlighted. Please click 'Next' to continue.

Page 8

As another example as DM 1, move DOWN by clicking on the arrow. To confirm your move click the "Send" button.

SCREEN DIAGRAM

Once the subject makes the choice, the following appears: Since you moved Down as DM 1, DM 2, seeing your move, now has a decision to make. If DM 2 moves right then you would earn 'wig' and DM 2 would earn 'wog'. If DM 2 moves down then you would earn 'xig' and DM 2 would earn 'xog'. Please click Next to continue.

Page 9

We will now show you what decisions look like from DM 2's point of view. Notice that your earnings are indented and this is the BOTTOM NUMBER in the boxes. You will only have a move if DM 1 moves down. Suppose DM 1 has moved down. You have to decide to move right or down. Please make a choice now by clicking on the arrow of your choice. Then click "Send" to confirm your move.

SCREEN DIAGRAM

Either the subject moves Right as DM 2 in which case she sees the following: Since you moved Right as DM 2, DM 1's earnings are 'wig'. Your earnings are 'wog'. Please click "Next" to continue.

OR the subject moves Down as DM 2 in which case she sees the following: Since you moved Down as DM 2, DM 1's earnings are 'xig'. Your earnings are 'xog'. Please click "Next" to continue.

Page 10

IMPORTANT POINTS:

* Each period you will be paired with another individual: your counterpart for that period.
* You will participate for several periods, being re-paired each period.
* If you are DM 1, your counterpart will be DM 2. In this case, you will make a decision first. On the other hand, if you are DM 2, your counterpart will be DM 1. If this is the case, you will have a decision to make if DM 1 chooses down.
* If you are DM 1, your payoff in Experimental Dollars is the top number in the box. If you are DM 2, your payoff in Experimental Dollars is the bottom number in the box. You will receive that amount of money if the box is reached. For every 20 Experimental Dollars you earn, you will receive 1 U.S. Dollar.

This concludes the directions. If you wish to return to them please click the "Back" button. If you have any questions please raise your hand. Otherwise, to begin the experiment, please click the green button, "Finished with directions".