Working Paper Series
No. 29

Testing static game theory with dynamic experiments: a case study of public goods

Anabela Botelho
Glenn W. Harrison
Lígia Pinto
Elisabet E. Rutström

November 2005

Núcleo de Investigação em Microeconomia Aplicada
Universidade do Minho

Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA

Testing Static Game Theory with Dynamic Experiments:

A Case Study of Public Goods

by

Anabela Botelho, Glenn W. Harrison, Lígia M. Costa Pinto & Elisabet E. Rutström †

August 2005

ABSTRACT

Game theory provides predictions of behavior in many one-shot games. On the other hand, most experimenters usually play repeated games with subjects, to provide experience. To avoid subjects rationally employing strategies that are appropriate for the repeated game, experimenters typically employ a "random strangers" design in which subjects are randomly paired with others in the session. There is some chance that subjects will meet in multiple rounds, but it is claimed that this chance is so small that subjects will behave as if they are in a one-shot environment. We present evidence from public goods experiments that this claim is not always true.

Corresponding author: Glenn Harrison, Department of Economics, College of Business, University of Central Florida, Orlando, FL 32816, USA. E-mail: gharrison@research.bus.ucf.edu.

JEL Classification Codes: C72, C92, H41
Keywords: Game theory, experiments, public goods

† Department of Economics, University of Minho and NIMA (Botelho and Pinto) and Department of Economics, College of Business, University of Central Florida (Harrison and Rutström). E-mail: botelho@eeg.uminho.pt, gharrison@research.bus.ucf.edu, pintol@eeg.uminho.pt, and erutstrom@bus.ucf.edu. Rutström thanks the U.S. National Science Foundation for research support under grants NSF/IIS 9817518, NSF/MRI 9871019 and NSF/POWRE 9973669. Botelho and Pinto thank the Fundação para a Ciência e Tecnologia for sabbatical scholarships SFRH/BSAB/489/2005 and SFRH/BSAB/491/2005, respectively. We are grateful to Ryan Brosette, Linnéa Harrison, James Monogan and Bob Potter for research assistance, and to R. Mark Isaac for comments. All data, instructions, and statistical code is available at the *ExLab* Digital Library at http://exlab.bus.ucf.edu.

Game theory provides predictions of behavior in many one-shot games. On the other hand, when testing one-shot games many experimenters conduct sequences of multiple games with subjects, to provide experience and to collect a larger set of observations. We consider the difficulty of drawing inferences about static game theory using repeated game experiments. To avoid subjects rationally employing strategies that are appropriate for the repeated game, experimenters often employ techniques that minimize or eliminate the probability that subjects meet more than once. In a "Random Strangers" design subjects are randomly paired with others in the session. There is some chance that subjects will meet multiple times, but it is believed that this chance is so small that subjects will behave as if they are in a one-shot environment. In fact, this belief is so strong that it has been referred to as a "repeated single-shot" design.[1] In a "Perfect Strangers" design a matching algorithm is used that guarantees that subjects meet only once. The polar opposite is a "Partners" design where the same subjects are pitted against each other for each round in a repeated game. We examine whether subjects perceive a Random Strangers experiment the same way that they perceive a Perfect Strangers experiment, such that the former can be used to reliably implement static games in the laboratory.

Comparisons of Partners and Random Strangers designs have been common since Andreoni [1988] reported the counter-intuitive result that the latter generates a greater amount of cooperation than the former. Most of the replications and variations of this experiment have found no significant difference in behavior, and a few even report the opposite, more intuitive, pattern. Nevertheless, if one is interested in testing one-shot theories, these experiments are not very informative since it is unclear to what extent Random Strangers implements sufficient control on the strategic environment perceived by subjects. To date, there have been no systematic test of Random Strangers versus

[1] Andreoni and Croson [2005].

Perfect Strangers, which is surprising given the popularity of the former. We provide evidence from public goods experiments that shows that *the assumption that Random Strangers is the same as Perfect Strangers is not always true*. We find that the fraction of subjects that play the game strictly by the non-cooperative Nash Equilibrium prediction is significantly higher in Perfect Strangers. The difference is 40 percentage points, a noticeable fraction of our subjects.

The Random Strangers design is by far the most popular operational counterpart of a one-shot environment in experiments. N subjects are recruited into a session in which subjects are paired into groups of K<N in each round. We assume without loss of generality that N is an integer multiple of K. In the Random Strangers design the K subjects in each group are picked at random from the N subjects, obviously without replacement. There is therefore some chance, varying in N and K, that any one subject will see the same subject in a later period providing that the subjects are not in the last round. This chance gets small very quickly as N increases in relation to K, and on the basis of this arithmetic it is usually assumed to be an adequate procedure that ensures that subjects behave as if the chance is actually zero. Occasionally, subjects are told that these chances are very small.

The Perfect Strangers design, on the other hand, picks pairings in a way that ensures that no subject will *ever* be paired with the same person in later rounds. The only problem with the Perfect Strangers design is that it is "subject hungry." For K>2, and N around 20 or so, it becomes difficult to run sessions with more than a few rounds. But the Perfect Strangers design is the one that literally matches the one-shot notion that underlies the theories being tested in the lab.

Experimenters know all of this, and yet it is surprising to find virtually no studies that systematically compare behavior in Perfect Strangers and Random Strangers settings. In addition to its popular use in comparisons between one-shot and repeated play games, the Random Strangers design is frequently used in other experimental investigations, particularly those involving public goods and auctions. We consider as an example the classic public goods voluntary contribution game. Apart from it's intrinsic importance, it has been the context for many valuable tests of the role of Random Strangers and Partners designs.

In section 1 we briefly review the literature on comparisons of behavior in one-shot versus repeated public goods settings, where the Random Strangers design has been used to model the one-shot setting. In section 2 we introduce a simple experimental design in which we compare the Random Strangers and Perfect Strangers designs. We find that *the assumption that subjects treat Random Strangers designs as if they were one-shot experiments is false*. Our subjects behave in a systematically different manner in the Perfect Strangers design. In fact, we can show that the Perfect Strangers design is associated with more subjects adopting a strict free-riding behavior consistent with the one-shot theory, rather than with subjects simply providing smaller contributions conditional on making some contribution. Thus the use of the Perfect Strangers design seems to *encourage a qualitative change in the way subjects view the game*, with more of them thinking the game through in the strategic manner assumed by game theory. Section 4 re-analyzes the data from two of the previous experiments that have examined Partners and Random Strangers, and shows that their results are consistent with these conclusions.

1. Partners, Random Strangers, and Perfect Strangers

A. Theoretical Issues

Why do we worry about Strangers designs at all, let alone whether they are Random Strangers or Perfect Strangers? The short answer is that any multiple round game can cause reputation effects, such that play in one round can be influenced by the expectation of meeting the same player in a later round. Strategically, of course, such reputation effects do not always affect predicted play. If the game has a finite and known number of repetitions, if the stage game in any single round has only one Nash Equilibrium (NE), and if it is common knowledge that all players can backward induct for the horizon of the game,[2] then the NE of the repeated game is just a "degenerate" succession of NE of the stage games. Nevertheless, if one relaxes any of the three conditions just noted, then there may be many NE of the repeated game that differ from degenerate, successive plays of the NE of the stage game.[3]

The public good games considered in the experimental literature are virtually identical in form to the prisoners' dilemma games considered in repeated game theory. In the standard form of both games there is invariably a single NE of the stage game. Most experiments provide subjects with a known and finite horizon. Some experiments leave the final horizon indeterminate, which can generate many of the same effects as having an infinite horizon. But the one thing we cannot easily control in experiments is the knowledge that subjects have about the other players. If the common knowledge assumption does not hold, the one-shot theory prediction is not the appropriate one to use since players may strategically want to create reputations. Thus, if the Random Strangers design introduces reputation effects, or even just perceived reputation effects, the appropriate theoretical domain is not one of static non-cooperative games.

We therefore hypothesize that behavior in Perfect Strangers will be significantly more like the non-cooperative one-shot NE prediction than Random Strangers. In addition we vary the size of the

---

[2] Chess reminds us that backward induction is not an "all or nothing" thing behaviorally.

[3] For textbook expositions, see Fudenberg and Tirole [1991; ch.4,5] or Binmore [1992; ch.8]. Obviously repeated games are interesting in their own right. Our concern is with the difficulty of drawing inferences about static game theory using repeated game experiments. The task of drawing inferences about repeated game strategy choices from observed actions in repeated game experiments is actually a delicate one: see Engle-Warnick and Slonim [2006] for a discussion of the issues and a proposed methodology

cohort from which subjects are matched in Random Strangers (i.e. we vary N) expecting to see behavior approaching that of Perfect Strangers as the probability of being re-matched with the same person declines.

B. Previous Experiments

The experimental literature on public goods has a long tradition of being concerned with the strategic importance of differentiating between "Partners and Strangers." However, it is striking that virtually all of the Strangers designs have been what we call Random Strangers.

The earliest public goods experiments were conducted exclusively with a Partners design.[4] However, the presumption was that the game would be viewed by subjects as a finite-horizon repeated game in which the sole NE was the same outcome as the NE of the stage game. For example, Isaac and Walker [1988; p.195] state this position clearly:

> The results across all periods are not supportive of the multi-period Nash equilibria prediction of zero contribution in every period (based upon a backwards induction argument). Instead, the experiments uniformly begin with positive contributions [...] followed by a tendency for contributions to decay. This decay pattern is consistent with the experimental results cited by Kreps et al. [1982], and it suggests that the incomplete information models should also be a fruitful line of theoretical inquiry for public goods research.

Of course, an alternative is to consider the effects of experimental designs that mitigate the role of reputation effects under incomplete information.

Andreoni [1988] initiated this approach in the experimental literature, explicitly contrasting what he termed Partners and Strangers. The Strangers design in his experiments were Random Strangers: 20 subjects were randomly assigned to 4 groups of 5 in each of 10 rounds. He reports that Random Strangers contributed more, on average, than Partners. This counter-intuitive result

[4] The same is true of the extensive early experimental literature on first-price sealed-bid auctions.

generated a flurry of interest, as discussed below. At an "eyeball" level the average contributions in each treatment are compared, round by round (Table 1, p.296). Partners contribute an average of 16.6 tokens over 10 rounds, and Random Strangers contribute an average of 20.7 tokens, for a difference of 4.1 tokens. The problem, noted by Croson [1996; p.30] and Palfrey and Prisbey [1996; p.413], is that there is a significant standard deviation in contributions, around 16 tokens per round in each treatment. The test used by Andreoni [1988; p.296, fn. 9] is a median test applied to all individual contributions over all rounds. The null hypothesis here is that the Partners and Random Strangers samples arise from populations with the same median, and he reports that this null can be rejected with a *p*-value of less than 0.01. Of course, this test assumes that the samples are random (Conover [1980; p.171]), and this is violated by the temporal dependence between rounds and subjects. We examine this hypothesis later, using an econometric specification which accounts for some of the features of these data and that is comparable to the analysis of our own data. To anticipate, we find that there is no statistically significant evidence of differences between Partners and Random Strangers in the data from Andreoni [1988].

Weimann [1994] undertook a replication of the Andreoni [1988] conclusion, but his experiments also changed the design in a way that makes them hard to compare. In his Strangers experiments the subjects were contacted by phone, after receiving instructions and a record form in the mail, rather than in some common setting. This has the advantage of complete anonymity, of course, but it also means that the subjects cannot verify that they are being randomly paired. As it happens, two of the variations on the baseline experiments employed deception, further clouding the credibility of inferences. In any event, Weimann [1994] concludes that he did not replicate the conclusion of Andreoni [1988].

Croson [1996] replicated the Andreoni [1988] design and also found different results.

Although she does not report the average contributions, inspection of her Figure 1 (p.28) indicates that contributions in the Partners treatments were roughly 4 to 5 tokens *higher* than those in the Random Strangers treatments, relative to an endowment of 25 tokens in each round. But the same concern with the variation of individual contributions arises. She reports (Table 2, p.30) standard deviations in each treatment around 8 tokens in each round. Using a Wilcoxon test, despite its assumption that observations are independent, she rejects the hypothesis that the contributions are the same, in favor of higher contributions in the Partners treatment. We also re-consider these results later, using an econometric specification to help us identify the sources of these differences, and verify these conclusions.

Burlando and Hey [1997] also fail to replicate the conclusion claimed by Andreoni [1988]. They find no significant difference between Partners and Random Strangers overall, although there are some minor interaction effects depending on the national location of the experiments and the sequencing.[5] It would be useful to see how much of the differences across nations is due to national effects rather than from differences in individual characteristics, as noted by Botelho, Harrison, Hirsch and Rutström [2005] in the context of cross-national bargaining experiments, but the raw data on individual characteristics was not collected for the British subjects (p.47, fn.10).

[5] The classification of these results in Andreoni and Croson [2005; Table 1] does not match the conclusions of the original study. They classify the British subjects as contributing more on the Strangers design compared to the Partners design, whereas the original study finds no difference; they classify the Italian subjects as contributing more in the Partners design compared to the Strangers design, but this is due to some differences following a restart, rather than in initial rounds of behavior. Burlando and Hey [1997; p.53-4] note that "... for the UK subjects the partners percentage was 86.64 as compared to 85.65 for the strangers – a difference that is not statistically significant (p=0.2073); for the Italians, the partners percentage was 70.62 as compared with 73.38 for the strangers – a difference that is significant according to a [Wilcoxon rank-sum test] at 1% (p=0.0051). Interestingly, this difference is largely driven by the difference in behavior between the first and second sub-sessions among the Italian subjects – in the second sub-session partners free-rode much less than strangers [...]. Perhaps by then they had learned that co-operation was a good thing?" The percentages they refer to are the percentages of the *bad* that was placed in the public domain. So "dumping" *more* in the public domain here amounts to free riding *more* and contributing *less*.

Palfrey and Prisbey [1996] conduct an experiment that compares Partners and Random Strangers, along with other treatments. Their subjects participated in 40-round games, broken into 4 treatments. In each round the subject received a random "exchange rate" that would convert their tokens into points. Each subject received a different exchange rate each round, and the subjects in the same group of 4 received different exchange rates. These exchange rates were drawn uniformly at random as integers between 1 and 20. In the first 20 rounds each subject received a fixed group return, and then a new, fixed group return in the last 20 rounds. These returns were "low" and then "high." In each 10-round sequence each subject received a random private return. Subjects were either in a Partners treatment for the entire 40 rounds or in a Random Strangers treatment for the entire 40 rounds. They find average contributions of 3.46 tokens in the Partners treatment and 3.71 in the Random Strangers treatment, out of endowments of 9 tokens. The standard deviation of each is 3.68 tokens and 3.60 tokens. One cannot reject the null hypothesis of identical mean contribution levels using a *t*-test (p=0.13),[6] nor can one reject the null hypothesis of identical variances in contributions using an F-test (p=0.51).[7] These results are even stronger if one only considers the responses in the first round, although sample sizes become very small since there were only 24 subjects in each of the main treatments.

Keser and van Winden [2000] provide convincing evidence that Partners contribute more than Random Strangers in stationary public good experiments. A key feature of their design was simply to increase the number of replications in each treatment, so that they had 6 sessions with the

---

[6] Andreoni and Croson [2005; Table 1] and Keser and van Winden [2000; p. 24] claim that these data show that Strangers contribute more than Partners, but this may just be due to them being willing to accept a p-value this high.

[7] Keser and van Winden [2000; p. 24] claim that these data show that Strangers have a higher variance in *contributions* than Partners, but this is likely due to a misreading of a claim by Palfrey and Prisbey [1996; p.424] about the dispersion in the fitted *parameter* of a specific model estimated from these treatments.

Random Strangers treatment and 10 sessions with the Partners treatment, spanning 160 subjects. They found that average contributions were 1.9 tokens and 4.53 tokens, out of an endowment of 10, across the two treatments.

Andreoni and Croson [2005] review the literature on public goods contributions with Partners and Random Strangers. They discuss additional studies examining these treatments, but in which there was some other design change.

Fehr and Gächter [2000; fn.3] report the only evidence we know of comparing Random Strangers and Perfect Strangers in a public goods experiment. They note briefly that the results of a Perfect Strangers replication of their design generated essentially the same results as their Random Strangers experiments. However, they only considered one sequence of regimes (Punishment followed by Non-Punishment), and did not maintain the Perfect Strangers treatment after the first regime of 6 periods. In other words, subjects that were matched only once in rounds 1-6 might have been matched again in rounds 7-12, thereby reducing the Perfect Strangers control. Moreover, one would have to control for the history generated by following a related experiment, as we do below, to be able to draw any inferences about the effects of Perfect Strangers rather than Random Strangers in a Non-Punishment game. Rather than debate if such comparisons are conclusive, we prefer to ensure the control against any reputational effects afforded by a Perfect Strangers design.

Generally, the comparisons of behavior in Partners and Random Strangers in this literature has led to interpretations that involve preferences with social arguments such as altruism or "warm glow." We introduce the alternative hypothesis that some subjects did not perceive the Random Strangers experiments as one-shot game environments, but strategically based their decisions on the presence of reputation effects.

2. Experimental Design

Each subject participated in an experimental session in which there were 10 rounds of a traditional voluntary contribution public goods game. Subjects participate in groups of 2 in each round.[8] We explain to subjects how we ensure that there is no chance that they will meet the same person in any other round.

Our experiments also had some other task after the initial 10 rounds, or in a prior 10 rounds. This task was a "sanctions" public goods game in the spirit of Fehr and Gächter [2000]. The results of those experiments are not of interest here, but we control for them in the statistical analysis.

Table 1 summarizes the experimental design. Thirteen sessions were conducted. The first four used Perfect Strangers designs, and the last nine used Random Strangers designs. We varied the size of the cohort in the Random Strangers design from 6 to 16 participants. Subjects were aware that different participants were in cohorts of different size, making this a salient feature of the design. We hypothesize that increasing the size of the cohort in Random Strangers will reduce the perceived reputation effect and make behavior more similar to Perfect Strangers. Each subject received an endowment of 20 tokens at the outset of each round, and each token was worth 5 cents.

In two sessions we used a relatively low return on contributions to the public good, and in all other sessions we used a relatively high return. The low return was 0.6 of token: hence every token contributed to the public good by one subject would decrease their private endowment by 1 token and return 0.6 of a token for herself. Of course, it would also generate 0.6 of a token for the other player, so the social return was 1.2 tokens for every 1 token invested. In the high return treatment we changed the public good return from 0.6 to 0.8, thereby increasing the social return from 20% to

[8] Most public goods experiments use four subjects per group, although the effect of larger group sizes has been studied by Isaac and Walker [1988] and others. Harrison and Hirshleifer [1989] and Goeree, Holt and Laury [2002] employed groups of 2 in their public goods experiments.

60%. The objective of this treatment was to see the effects of making the environment more rewarding to any strategy that would increase contributions to the public good. In terms of the marginal per capita return (MPCR) to contributing, which is just the ratio of the return from the public good contribution to the return from the implicit private good contribution, these are 0.60 and 0.80, respectively.[9]

We used a linear payoff schedule which was constant for all contributions, so the dominant strategy is simple: a subject that only seeks to maximize individual earnings in a single period should contribute nothing to the public good.[10]

We recruited 180 subjects from the University of Central Florida (UCF) in 2005.[11] Subjects were randomly assigned to each session, with no prior knowledge of the parameters or treatments. The sessions were all conducted at the Behavioral Research Lab of the College of Business Administration of UCF. This facility is a standard, computerized laboratory: each station has a "sunken" monitor, and we employed personal "cubicle-style" screens to ensure even more privacy. Instructions were provided in written form and orally, and the experiment was implemented using version 2.1.4 of the *z-Tree* software developed by Fischbacher [1999].[12] The same experimenter (Rutström) delivered the oral instructions for all sessions, to ensure comparability.[13] The oral

---

[9] Isaac and Walker [1988] carefully discuss the relationship between changes in group size and the implied MPCR. They use MPCR values of 0.3 and 0.75, and refer to the latter as high. So our values tend to be "high" in relation to the ones they consider.

[10] Alternative assumptions about the factors motivating subjects to contribute in public goods experiments have long been studied. See, in particular, Palfrey and Prisbrey [1996][1997] and Goeree, Holt and Laury [2002].

[11] UCF is located in Orlando, Florida. It has a large student body, with Fall 2004 enrollment of 42,837. The entering class in 2004 had an average SAT of 1,186. The student body is also ethnically divers: in 2004 8.5% stated that they were Black and Non-Hispanic; 70% stated that they were White and Non-Hispanic; 5.0% stated that they were Asian; and 12.2% stated that they were Hispanic.

[12] All instructions, scripts, and software are available at http://exlab.bus.ucf.edu. The latest version of the *z-Tree* software and documentation is available at http://www.iew.unizh.ch/ztree/index.php.

[13] A digital recording of the oral instructions in one typical session is available at the ExLab archive.

instructions also utilized a large-screen display that could be easily seen by all subjects, to ensure that certain information was common knowledge. Training rounds were included prior to each regime, to ensure that subjects understood the task.

The average subject earned $39 in these experiments, including a standard $5 show-up fee. No session lasted more than 2 hours, and most were at least 1½ hours in length.

3. Results

Figure 1 displays average contributions over each round of our experiments, pooling data for each of the Perfect Strangers and Random Strangers treatments. Maximum token contributions in each round could be 20, and we observe average contributions starting out at around 5 and 8 tokens and steadily declining. Average Perfect Strangers contributions are consistently lower than those in the Random Strangers treatment, consistent with our hypothesis that the former is more like the one-shot non-cooperative environment than the latter. However, these raw results do not control for a number of possible confounds. Two of our Perfect Strangers sessions had lower returns to the public good, there could be some effect from the previous history of the experiment, and there might be sampling differences across treatments that are associated with individual characteristics. To account for these possible effects we turn to a statistical analysis that conditions on them.

Our analysis employs a likelihood function that is constructed to be appropriate for this type of experiment. Theory tells us that there may be some individuals that gravitate to one particular contribution level: contribute zero. The raw data also flag this as a "spike" that needs to be addressed explicitly. Figure 2 shows the distribution of fractional contributions in each treatment, pooled over all periods. The mode at zero is evident. Figure 3 shows the same distribution in the initial round with a similar mode at zero. More subjects seem to focus on the zero contribution from the outset in the

Perfect Strangers environment. Our statistical analysis therefore considers the process by which some subject decides to contribute zero or some positive amount as separate from the process by which the subject decides how much to contribute.

The natural specification to capture this intuition from theory and the raw data is a "hurdle model." This specification is common in health economics, for example, where it is used to capture the idea that the factors that cause someone to seek medical care are distinct from the factors that cause the doctor and patient to decide how much to spend.[14] In this case going to the doctor is the hurdle that must be passed before expenditures would be positive. In our case the subject has to decide whether to contribute any amount at all, and only then does the process determining the positive contribution level apply.

The likelihood function for the overall hurdle model is constructed as the product of two likelihoods.[15] The first component is the likelihood that the subject contributed zero or not, and uses a standard probit specification defined over an index function $x_i\alpha$, where α is a parameter vector to be estimated and x_i is a vector of explanatory variables for observation i. The second component is the *conditional* likelihood that the subject contributed a certain fraction of the endowment. This likelihood function is constructed using the specification developed by Papke and Wooldridge [1996] for fractional dependant variables, since the dependant variable in this case is the fraction of the endowment contributed (conditional on any positive contribution).[16] Thus the log-likelihood of observation i is defined as $l_i(\beta) = C_i \times \log[G(x_i\beta)] + (1-C_i) \times \log[1-G(x_i\beta)]$ for contribution fraction C_i, parameter vector β, and some convenient cumulative distribution function $G(\cdot)$. We use the

[14] See Coller, Harrison and McInnes [2002] for an application.

[15] As is well known, one can alternatively estimate the two parts of the hurdle model separately and obtain consistent and efficient estimates (McDowell [2003]).

[16] In order to compare the estimation results based on our data to that of Andreoni [1988] and Croson [1996] we use fractions, since the initial endowments used across these three studies differ.

standard normal cumulative distribution function $G(z) = \Phi(z)$. Thus the overall likelihood function for the hurdle model requires the estimation of α and β. Since our data is a panel we use a specification that treats observations as independent across subjects but not within.

Explanatory variables include individual demographics and treatment effects. In addition to a binary dummy variable for the Perfect Strangers designs (Pstrangers), we also include dummy variables for the size of the cohort conditional on the use of a Random Strangers design (Csize).[17] If differences in behavior between Perfect Strangers and Random Strangers is due to perceived reputation effects, we expect a smaller cohort in Random Strangers to be correlated with a stronger difference in behavior from Perfect Strangers. We also include treatment dummies for the structure of the game during periods 1-10 (np_p), and the use of high rewards to contributing to the public good (High). Demographics include a measure of age in years (Age), binary indicators for sex (Male), race (Black, Asian, Hispanic or Other Race), academic major (Business), class standing (PreSenior), cumulative GPA below 3¼ (GPAlow), cumulative GP above 3¾ (GPAhigh), number of people in the subject's household (Hhsize), and a binary indicator of those that work part-time or full-time (Work). Table 2 lists descriptive statistics for these variables.

Table 3 provides maximum likelihood estimates of the hurdle model for these data. All estimates for the α parameter represent the calculated marginal effect of that variable on the probability of contributing. The reported estimates for the β are the marginal effects in terms of the positive fraction of tokens contributed.

The focus variable is the Pstrangers binary dummy. It clearly has a large and statistically significant effect on the decision to contribute something or nothing, and virtually no impact on the

[17] This variable takes on the value 0 for the Perfect Strangers treatment and the size of the cohort (from Table 1) for the Random Strangers treatments. Thus it can be viewed as an interaction between the Perfect Strangers treatment and cohort size.

level of positive contributions. This is striking evidence that the *Perfect Strangers treatment affects qualitative behavior, in the sense that it elicits more subjects to focus on the zero contribution response.* Subjects are on average 30 percentage points more likely to be free riders in the Perfect Strangers treatments, and this effect is statistically significant (*p*-value = 0.055). This is clear evidence that a Random Strangers environment does not elicit the same behavior as a comparable Perfect Strangers environment, and that the direction of the change in behavior is consistent with the Perfect Strangers environment being more conducive to subjects viewing the stage game as one-shot.

The effect of increasing the size of the cohort from which subjects are matched is to lower the probability of contributing by 2.4 percentage points for every extra member of the cohort. The difference between Random Strangers and Perfect Strangers is therefore diminishing in the size of the Random Strangers cohort. Since the reference cohort is the smallest one, consisting of 6 individuals, we predict that an increase in the size of the cohort by 11, to a size of 17, would approximate the Perfect Strangers environment for this experiment where we have pair-wise matching and a total of 20 periods.

We also find that the ordering of the experiments, as first or second in the real-time sequence, also makes a large difference to contributions. When the experiment comes before the other task, and variable np_p equals 1, the probability of being a free rider is 0.25 *lower* on average. Perhaps this is related to the fact that subjects in the Random Strangers experiment may underestimate the total number of rounds during which they will be randomly rematched within their cohort, thus leading to a lower perceived reputation effect than when they have already played 10 rounds.[18]

[18] This is a possibility since in the np_p treatment the experiment comes before the other task and subjects are not made explicitly aware of an additional 10 rounds in the second task, although they know of the second task itself. In the alternative treatment the experiment comes after the other task, so subjects know for sure that they are matched up with people from their cohort over a total of 20 periods.

Increases in the reward to contributing to the public good are associated with significant reductions in free riding, and significant increases in the amount contributed when someone does contribute something. This is consistent with previous observations in public goods experiments, as is the fact that the passage of time increases the likelihood of a subject becoming a free rider. The marginal effects of the period dummies on α in Table 3 show a significant and steady decline after period 3.[19]

There are some clear demographic effects on contributions, particularly in terms of the fraction of contributions conditional on making any contribution. Men are actually more generous once they decide to contribute, even though there is an offsetting (and less statistically significant) effect on the decision to contribute. Those with a lower GPA are much more likely to contribute something, although they tend to contribute less once they decide to contribute something.

4. Comparisons to the Previous Experiments

Our econometric model of contributions allows us to re-examine data from the previous experiments of Andreoni [1988] and Croson [1996] using statistical methods that are comparable to our own. Using data on individual contributions,[20] we estimate the same hurdle model with controls for the key Partners versus Random Strangers treatment. We also include fixed effects for each round, and interact those with the Partners/Strangers treatment.

Our estimated model based on the data from Andreoni [1988] finds no significant difference between Partners and Random Strangers. There is generally no difference in terms of whether subjects decide to contribute anything at all, or in terms of what level of contribution they would

---

[19] The size of the marginal effect for each period appears to be too large, but is appropriate given that each dummy has an average sample value of 0.1, and the effects are each measured relative to period 1.

[20] Generously provided by James Andreoni and Rachel Croson.

make if positive. We do find statistically significant round effects, but these are common to the Partners and Random Strangers treatments.[21]

The data from the experiments of Croson [1996], tell an even stronger story. Estimating the same statistical model with her data, over all 20 rounds, we find a large effect from Random Strangers on the propensity to free ride, compared to Partners, and no effect at all on the level of contributions conditional on making any. Random Strangers are 42.5 percentage points less likely overall to make any contribution, and this is a significant effect (p-value = 0.004). They are estimated to contribute 11.1 percentage points more conditional on making any contribution at all, but this is not statistically different from zero (p-value = 0.31).

Thus, these statistical results show a qualitative effect on behavior such that Random Strangers elicits more subjects to focus on the zero contribution response, but with no effect on the conditional contributions.

5. Conclusions

We find a significant effect from the use of a design that ensures a zero probability of re-encounters between subjects, the Perfect Strangers design. Our experiment and statistical analysis provide evidence that it not only moves subjects towards the prediction of standard theory for one-shot games, but it significantly increases the fraction of subjects that behave exactly according to the one-shot prediction from the outset. Further, in our Random Strangers design we find that the size of

---

[21] The only effect that we observe is a fascinating one in terms of the underlying static game theory: an end-period effect. Andreoni [1988; p.295] explains that "... we expect that giving by Partners will be greater than giving by [Random] Strangers, especially early in the game (before the Partners begin to 'bail out'). In the tenth round, however, both Partners and Strangers are playing an end-game, hence both are predicted to free ride." However, in the last round we find that Random Strangers are 17 percentage points *more* likely to contribute some amount, and this effect is statistically significant (p-value = 0.028). This end-game effect is not sufficient overall to offset the conclusion that Random Strangers and Partners are behaving similarly in these experiments.

the cohort plays a significant role, increasing the fraction of NE players as the group size increases, causing the probability of re-encounters to decrease. We conclude that Random Strangers will not necessarily implement a one-shot environment. It appears that the fraction of subjects that perceive the environment as having reputation effects is smaller in Random Strangers than in Partners, but still larger than in Perfect Strangers. In addition, this fraction increases as the cohort size increases in the Random Strangers design. We conjecture that Perfect Strangers designs will have comparable effects in other strategic experimental tasks in which there may be effects from the subjects behaving as if in a repeated game.

Table 1: Experimental Design

Each experiment had 10 rounds of one regime, followed by 10 rounds of the other regime

Session	Return to Public Good	Anonymity	N in Session	History
A	Low	Perfect	26	NP-P
B	Low	Perfect	24	P-NP
C	High	Perfect	26	NP-P
D	High	Perfect	26	P-NP
E	High	Random	10	P-NP
F	High	Random	16	P-NP
G	High	Random	8	P-NP
H	High	Random	6	P-NP
I	High	Random	8	NP-P
J	High	Random	6	NP-P
K	High	Random	6	NP-P
L	High	Random	8	NP-P
M	High	Random	10	NP-P

Figure 1: Average Contributions

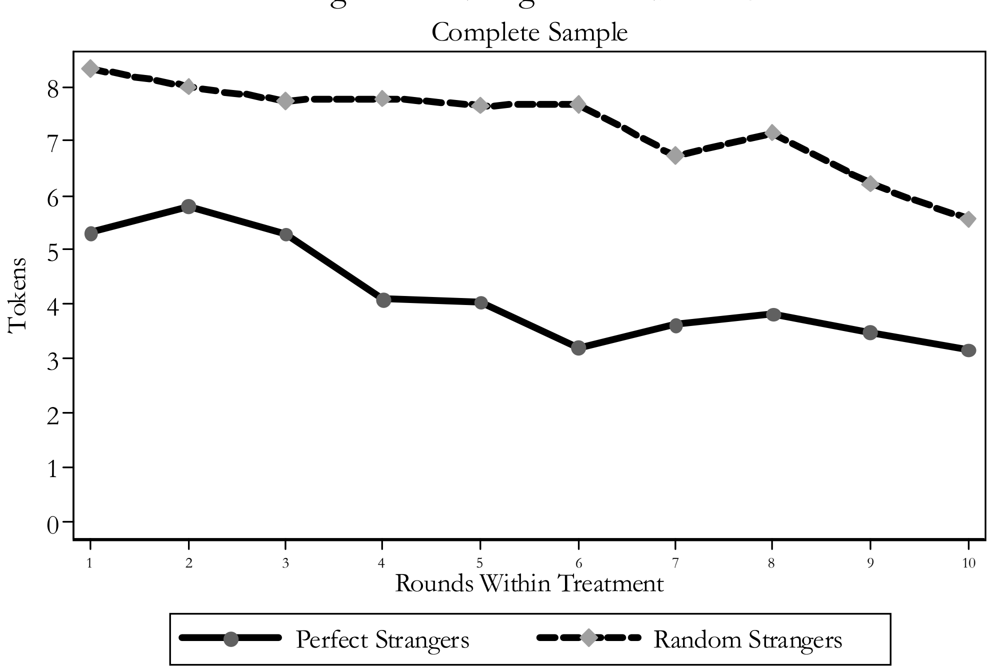

Figure 2: Distribution of Fractional Contributions

Pooled Over All Rounds

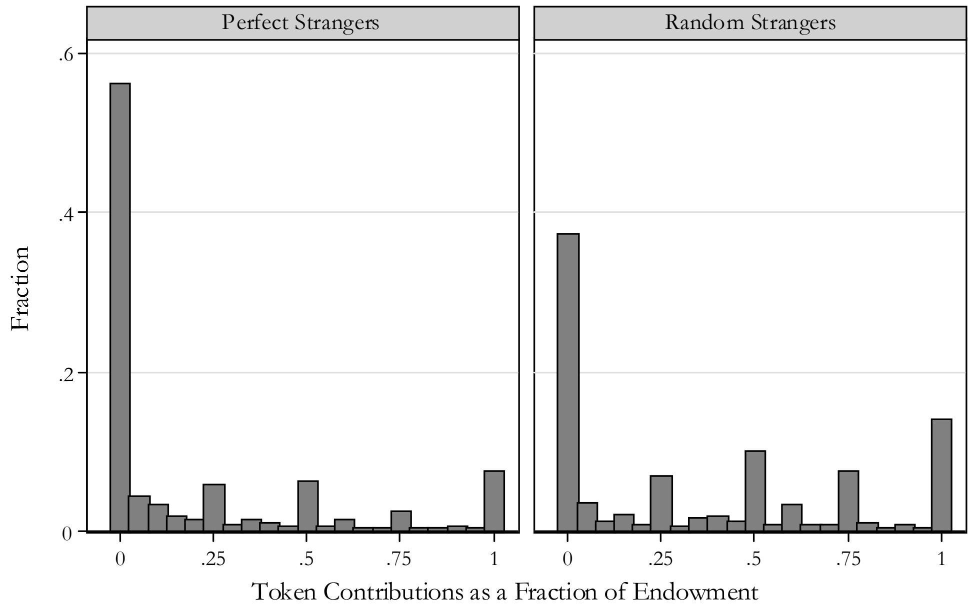

Figure 3: First Round Distribution of Fractional Contributions

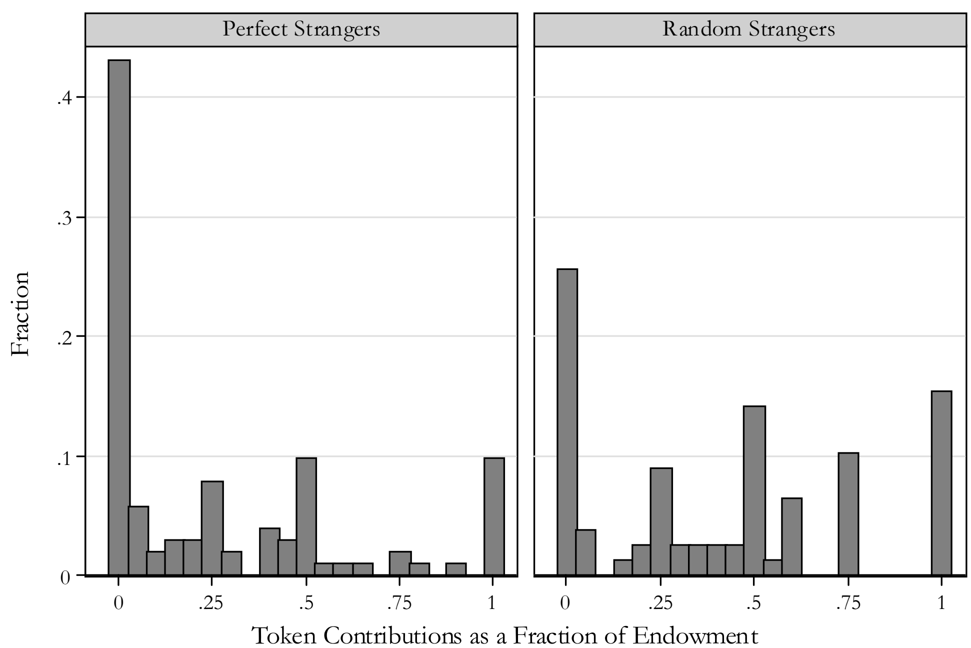

Table 2: Descriptive Statistics for Explanatory Variables

Variable	Description	N	Mean	Std. Dev.	Minimum	Maximum
Pstrangers	Perfect strangers	180	0.57	0.50	0	1
Csize	Cohort size for RS	180	4.20	5.34	0	16
np_p	Sequence	180	0.50	0.50	0	1
High	High reward for contributions	180	0.72	0.45	0	1
Age	Age	180	21.51	2.65	18	36
Male	Male	180	0.64	0.48	0	1
Black	Black	180	0.08	0.28	0	1
Asian	Asian	180	0.08	0.28	0	1
Hispanic	Hispanic	180	0.13	0.33	0	1
White	White	180	0.66	0.48	0	1
OtherRace	Other Race	180	0.04	0.21	0	1
Business	Business major	180	0.43	0.50	0	1
PreSenior	Pre-senior	180	0.47	0.50	0	1
GPAlow	Low GPA	180	0.47	0.50	0	1
GPAhigh	High GPA	180	0.15	0.36	0	1
HHsize	Size of household	180	1.65	1.26	1	7
Work	Any work	180	0.72	0.45	0	1

Table 3: Maximum Likelihood Estimates of the Hurdle Model of Contributions

Marginal effects for α and β parameters
N=1800 responses from 180 subjects; Wald test of H_0: α=β=0 has χ^2_{25} = 115.5 (*p*-value<0.001)

Parameter	Variable	Description	Estimate	SE	*p*-value	95% Confidence Intervals	
α	Pstrangers	Perfect strangers	-0.299	0.156	0.055	-0.605	0.007
	Csize	Cohort size for Random strangers	-0.024	0.015	0.119	-0.054	0.006
	np_p	Sequence	0.255	0.063	0.000	0.132	0.378
	High	High reward for contributing	0.373	0.070	0.000	0.236	0.511
	Age	Age	0.021	0.012	0.089	-0.003	0.045
	Male	Male	-0.041	0.071	0.565	-0.181	0.099
	Black	Black	0.082	0.125	0.509	-0.162	0.327
	Asian	Asian	-0.041	0.103	0.694	-0.243	0.162
	Hispanic	Hispanic	0.130	0.082	0.111	-0.030	0.291
	OtherRace	Other Race	0.109	0.137	0.425	-0.159	0.378
	Business	Business major	-0.092	0.068	0.176	-0.225	0.041
	PreSenior	Pre-senior	0.088	0.067	0.186	-0.043	0.219
	GPAlow	Low GPA	0.180	0.069	0.009	0.045	0.315
	GPAhigh	High GPA	-0.081	0.096	0.401	-0.270	0.108
	HHsize	Size of household	-0.021	0.025	0.415	-0.071	0.029
	Work	Any work	0.078	0.064	0.222	-0.047	0.204
	PeriodNP2	Period 2	-0.028	0.037	0.454	-0.100	0.045
	PeriodNP3	Period 3	-0.081	0.037	0.031	-0.153	-0.008
	PeriodNP4	Period 4	-0.109	0.038	0.004	-0.183	-0.035
	PeriodNP5	Period 5	-0.133	0.038	0.001	-0.208	-0.058
	PeriodNP6	Period 6	-0.195	0.042	0.000	-0.278	-0.111
	PeriodNP7	Period 7	-0.239	0.041	0.000	-0.319	-0.159
	PeriodNP8	Period 8	-0.212	0.043	0.000	-0.296	-0.128
	PeriodNP9	Period 9	-0.257	0.040	0.000	-0.336	-0.178
	PeriodNP10	Period 10	-0.301	0.040	0.000	-0.378	-0.223
β	Pstrangers	Perfect strangers	-0.141	0.100	0.160	-0.338	0.056
	Csize	Cohort size for Random strangers	-0.015	0.010	0.130	-0.034	0.004
	np_p	Sequence	0.045	0.050	0.374	-0.054	0.143
	High	High reward for contribution	0.288	0.055	0.000	0.181	0.395
	Age	Age	-0.003	0.013	0.842	-0.028	0.023
	Male	Male	0.139	0.045	0.002	0.050	0.228
	Black	Black	-0.016	0.095	0.863	-0.202	0.170
	Asian	Asian	-0.043	0.080	0.589	-0.199	0.113
	Hispanic	Hispanic	-0.104	0.074	0.159	-0.248	0.041
	OtherRace	Other Race	-0.009	0.077	0.905	-0.161	0.142
	Business	Business major	0.074	0.051	0.142	-0.025	0.174
	PreSenior	Pre-Senior	-0.005	0.055	0.930	-0.112	0.103
	GPAlow	Low GPA	-0.087	0.050	0.083	-0.185	0.012
	GPAhigh	High GPA	-0.042	0.079	0.600	-0.197	0.114
	HHsize	Size of household	0.015	0.019	0.429	-0.022	0.051
	Work	Any work	0.118	0.063	0.061	-0.005	0.241
	PeriodNP2	Period 2	0.026	0.025	0.290	-0.022	0.075
	PeriodNP3	Period 3	0.026	0.030	0.378	-0.032	0.085
	PeriodNP4	Period 4	-0.027	0.029	0.341	-0.083	0.029
	PeriodNP5	Period 5	-0.027	0.033	0.422	-0.092	0.038
	PeriodNP6	Period 6	-0.029	0.034	0.392	-0.095	0.037
	PeriodNP7	Period 7	-0.023	0.037	0.533	-0.094	0.049
	PeriodNP8	Period 8	-0.006	0.037	0.876	-0.077	0.066
	PeriodNP9	Period 9	-0.036	0.038	0.347	-0.110	0.039
	PeriodNP10	Period 10	-0.039	0.038	0.305	-0.114	0.036

References

Andreoni, James, "Why Free Ride? Strategies and Learning in Public Goods Experiments," *Journal of Public Economics*, 37, 1988, 291-304.

Andreoni, James, and Croson, Rachel T.A., "Partners versus Strangers: Random Rematching in Public Goods Experiments," in C.R. Plott and V.L. Smith (eds.), *Handbook of Experimental Economics Results* (North-Holland: Amsterdam, 2005).

Binmore, Ken, *Fun and Games: A Text on Game Theory* (D.C. Heath & Co.: Lexington, MA, 1992).

Botelho, Anabela; Harrison, Glenn W.; Hirsch, Marc A., and Rutström, Elisabet E., "Bargaining Behavior, Demographics and Nationality: What Can the Experimental Evidence Show?" in J. Carpenter, G.W. Harrison and J.A. List (eds.), *Field Experiments in Economics* (Greenwich, CT: JAI Press, Research in Experimental Economics, Volume 10, 2005, 337-372).

Burlando, Roberto, and Hey, John D., "Do Anglo-Saxons Free-Ride More?" *Journal of Public Economics*, 64, 1997, 41-60.

Coller, Maribeth; Harrison, Glenn W., and McInnes, Melayne Morgan, "Evaluating the Tobacco Settlement: Are the Damages Awards Too Much or Not Enough?"*American Journal of Public Health*, 92(6), June 2002, 984-989.

Conover, W.J., *Practical Nonparametric Statistics* (New York: Wiley, Second Edition, 1980).

Croson, Rachel T.A., "Partners and Strangers Revisited," *Economics Letters*, 53, 1996, 25-32.

Engle-Warnick, Jim, and Slonim, Robert L., "Inferring Repeated-Game Strategies from Actions: Evidence from Trust Game Experiments," *Economic Theory*, 28, 2006, 603-632.

Fehr, Ernst, and Gächter, Simon, "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, 90(4), September 2000, 980-994.

Fischbacher, Urs, "z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual," *Working Paper Nr. 21*, Institute for Empirical Research in Economics, University of Zurich, 1999.

Fudenberg, Drew, and Tirole, Jean, *Game Theory* (MIT Press: Cambridge, MA, 1991).

Goeree, Jacob K.; Holt, Charles A., and Laury, Susan K., "Private costs and public benefits: unraveling the effects of altruism and noisy behavior," *Journal of Public Economics*, 83, 2002, 255–276.

Harrison, Glenn W., and Hirshleifer, Jack, "An Experimental Evaluation of Weakest-Link/ Best-Shot Models of Public Goods," *Journal of Political Economy*, 97, February 1989, 201-225.

Isaac, R. Mark and Walker, James M., "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism," *Quarterly Journal of Economics*, 53, 1988, 179-200.

Keser, Claudia, and van Winden, Frans., "Conditional Cooperators and Voluntary Contributions to Public Goods," *Scandinavian Journal of Economics*, 102, 2000, 23-39.

Kreps, David M.; Milgrom, Paul; Roberts, John, and Wilson, Robert, "Rational Cooperation in the Finitely-Repeated Prisoners' Dilemma," *Journal of Economic Theory*, 27, 1982, 245-252.

McDowell, Allen, "From the help desk: hurdle models," *Stata Journal*, 3(2), 2003, 178-184.

Palfrey, Thomas R., and Prisbrey, Jeffrey E., "Altruism, Reputation, and Noise in Linear Public Goods Experiments," *Journal of Public Economics*, 61, 1996, 409-427.

Palfrey, Thomas R., and Prisbrey, Jeffrey E., "Anomalous Behavior in Linear Public Goods Experiments: How Much and Why?" *American Economic Review*, 87, 1997, 829–846.

Papke, Leslie E., and Wooldridge, Jeffrey M., "Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates," *Journal of Applied Econometrics*, 11, 1996, 619-632.

Weimann, Joachim, "Individual Behaviour in a Free Riding Experiment," *Journal of Public Economics*, 54, 1994, 185-200.