ELSEVIER

Journal of Public Economics 61 (1996) 409–427

JOURNAL OF PUBLIC ECONOMICS

Altruism, reputation and noise in linear public goods experiments

Thomas R. Palfrey[a,*], Jeffrey E. Prisbrey[b]

[a]*California Institute of Technology, Division of Humanities and Social Sciences 228-77, Pasadena, CA 91125, USA*

[b]*Universitat Pompeu Fabra, Barcelona, Spain*

Received March 1994; revised version received July 1994

Abstract

We report an experiment using a design that permits the direct measurement of individual decision rules in voluntary contribution games. We estimate the distribution of altruism in our subjects and find that observed 'overcontribution' is attributable to a combination of random variation in behavior and a few altruistic players. We also employ Andreoni's partners/strangers design to measure reputation effects. The only difference observed is that the strangers treatment produces slightly more random variation in behavior. Our results explain some anomalies about contribution rates, and support past findings that reputation-building plays a minor role in such experiments.

Keywords: Voluntary contributions; Public goods; Experiments; Reputation; Learning; Errors

JEL classification: O26; 215

1. Introduction

The most common public goods experiment examines the extent to which contributions occur when individuals have a dominant strategy not to contribute. This mechanism of public good provision is called the voluntary

* Corresponding author.

[1] Present address: Federal Communications Commission, Washington, D.C.

SSDI 0047-2727(95)01544-2

contribution mechanism. In these experiments, a subject who is a member of a small group, is endowed with an amount of a good that may either be consumed privately or contributed to the public good of the group. Incentives are usually designed so that a self-interested subject has a strict dominant strategy to contribute nothing, but the efficient outcome for the group is for each subject to contribute all their input to the public good.

A common finding in these experiments is that subjects often contribute, thereby violating their dominant strategy. In addition, contribution rates have been found to be correlated with a number of treatment variables such as experience and induced preferences for the public good. However, to date there is no coherent theory that can account for the variety of findings that have been reported. A number of casual explanations for some of these findings have been offered in the literature, some suggesting a type of altruism that contaminates the experimentally induced incentives,[1] and/or that the subjects are trying to establish a reputation in order to influence play later in the experiment.

In Palfrey and Prisbrey (1992) we proposed an alternative explanation, namely that most of the observed anomalies could be accounted for simply as background noise, and that the appearance of altruistic behavior or strategic reputation-building is illusionary or, at best, of minor importance in explaining the data. As a result of the usual experimental designs in which errors can only be manifested as overcontribution, the importance of systematic findings such as altruism and strategic play have been overstated.[2] To a limited extent, recent experiments have been conducted that lend some credence to this view,[3] but a careful study that is designed to precisely measure the relative contribution of each of the various proposed explanations has not yet been carried out. Unfortunately, the typical experimental designs do not permit precise measurement of the separate contribution of these diverse effects: altruism, reputation-building, and noise. In this paper we present the results of an experiment that was specifically designed to sort out these effects and accurately measure the separate contribution of each.

A basic premise of our study is that individual behavior can be decomposed statistically into a systematic component and a residual component. We call the systematic component a *decision rule*, and the residual component *noise*, or error. In the context of a linear voluntary contribution game it is natural to limit attention to very simple decision rules, called cutoff decision rules, in which an individual contributes if and only if his

[1] See the survey by Ledyard (1993).

[2] Overcontribution is small in magnitude or non-existent in other public goods experiments where errors can be made in both directions. See Palfrey and Rosenthal (1988, 1991), and the references they cite.

[3] See, for example, Andreoni (1988, 1992) and Saijo and Nakamura (1993).

marginal rate of substitution between the private good and the public good is less than or equal to some critical value. This includes as a special case perfectly self-interested behavior,[4] where the critical value is 1. However, altruistic behavior or reputation-building behavior would be consistent with decision rules where the critical value is set higher than 1. 'Spiteful' behavior (Saijo and Nakamura, 1993) corresponds to a critical value less than 1. The noise component of individual behavior is modelled as statistical deviation from a cutpoint rule. One way to think of this is that the *observed* decision rule of a subject has some random variation over time due to extraneous factors that are essentially impossible to measure. These factors would include computational errors, errors associated with learning by doing, and so forth. With this interpretation of the noise component, we expect experience to lead to a decrease in noise.[5] We interpret such decreases in noise as evidence of *learning*.

Past experimental designs make it virtually impossible to accurately identify the decision rule component from the noise component. In those experiments, there is little if any variation of the marginal rate of substitution. Typically, everyone has the same marginal rate of substitution throughout the experiment, and it is greater than 1. The focus of attention is on the aggregate frequency of violations of a deterministic version of the self-interest model of behavior. In the context of our non-deterministic two-component model of individual behavior, contribution could be due to altruism or reputation-building, or it could be due to noise. In those experiments, noise leads to systematic bias in the data, in that (at least relative to the self-interested model) only noise that leads to contribution can possibly be observed.

An accurate measurement of a subject's decision rule and the magnitude of the noise component is possible in a heterogeneous and changing environment, i.e. an environment where a subject faces a number of different marginal rates of substitution, and yet his information is otherwise the same. It is then possible, by a variety of methods (Palfrey and Rosenthal, 1991), to estimate the subject's decision rule. As well as estimating the extent to which cut-point rules deviate from 1, these methods also calibrate the noise component. The design reported here systematically varies each subject's marginal rate of substitution in order to estimate the distribution of decision rules and the distribution of the error rates. This allows us to measure the extent to which altruism or reputation-building explains the commonly observed overcontribution and the extent to which

[4] Reputational play could also involve more complicated decision rules where the cut-point changes over time or as a function of history.

[5] Experience could also lead to adaptation of the decision rule, although we find little evidence for this.

these observations can be accounted for simply as noise. This also allows us to measure the extent to which players learn by experience.

Once the noise component and the systematic component of individual choice behavior have been separated, the next step is to break down the systematic component of decision rules and to identify the relative importance of altruistic behavior and strategic reputation-building behavior. Following the approach of Andreoni (1988), we do this by conducting half of the experimental contribution games as a sequence of one-shot encounters with changing group membership (the 'strangers' treatment) and half the contribution games as a sequence of encounters where group membership remains fixed (the 'partners' treatment).

The difference between the decision rule in a series of one-time encounters and the decision rule in a similar number of encounters repeated within the same group could be attributed exclusively to reputation-building. Accordingly, a comparison between the decision rules measured under the two treatments is then made. If reputation-building is an important part of the explanation, we should observe decision rules with higher points in the partners treatment than in the strangers treatment. In addition, we should observe significantly more decay (declining contribution rates over the course of an experiment) in the partners treatment. The ability of our method to measure error rates means that we are able to draw firm conclusions about whether decay in previous experiments was due to learning or was evidence of reputation-building.

The rest of the paper is organized as follows. Section 2 discusses the relevant findings from past experiments. Section 3 describes our experimental environment. Section 4 explains the details of the design. Section 5 analyzes the data. We make concluding remarks in Section 6.

2. Previous research

The experimental study of public good provision by the voluntary contribution mechanism has a history that is well detailed in Dawes (1980) and in Ledyard (1993). Almost all past research, including the influential works of Marwell and Ames (1979, 1980, 1981), Isaac and Walker (1988), Issac et al. (1984), and Andreoni (1988), examine situations in which each subject's marginal rate of substitution is fixed for all periods of the experiment; usually all subjects are assigned identical valuations.

A number of general findings have emerged from the literature:

- aggregate contribution rates range between 20% and 50%;
- at some point in time and in violation of dominant strategy incentives, nearly all players contribute to the public good;

- there is a strong negative relationship between the marginal rate of substitution and the rate of contribution; and
- contribution rates fall with repetition and with experience (where repetition represents a sequence of decisions within the same group, and experience represents another similar sequence of decisions with a different group).

And, with regard to learning and reputation effects, Andreoni (1988) finds that:

- subjects in repeated encounters contribute less to the public good than subjects in one-time encounters;
- the proportion of *free riders*, or subjects that consistently use the dominant strategy decision, is greater in repeated encounters than in one-time encounters; and
- experience effects are greater for subjects in one-time encounters than for subjects in repeated encounters.

A number of papers (Ledyard, 1993, and references therein) have tried to attribute the contributions to altruism on the part of the subjects. It is argued that the experimentally induced monetary incentives do not fully control for all aspects of a subject's utility, and that utility may partly depend on the welfare or efficiency of the group outcome as well as monetary payoff. If the amount of consideration given to the group outcome is high enough, contribution to the public good is consistent with utility maximization.

However, the presence of altruism does little to explain the counter-intuitive results in Andreoni (1988). After all, with the additional assumption of incomplete information, the ability to establish reputations is known, at least theoretically, to justify the use of dominated strategies; see Kreps et al. (1982). The work of Kreps et al. suggests that, if anything, the contribution rates in repeated encounters should be higher, not lower, than the contribution rates in one-time encounters.

In addition to the systematic qualitative features of the data noted above, there is also much statistical variation across trials. This suggests yet another explanation, which is simply that the data are noisy.[6] Because of the experimental designs that are used, 'noise' (in the sense of statistical deviation from the theoretical prediction) can only manifest itself as contribution. None of the past studies is designed to collect data that enable accurate measurement of the separate effects of 'noise' and 'altruism' on voluntary contributions.

Recently, Andreoni (1992) and Palfrey and Prisbrey (1992) have designed experiments that enable differentiation. Andreoni proceeds by comparing

[6] One can imagine many reasons why the data might be noisy: incomplete subject understanding of the rules; low payoff salience; boredom; experimenter effects; demand effects; etc.

data collected from a standard environment with data collected from a similar environment in which group efficiency was no longer important to the individual. Andreoni attributes actions which help the group in the manipulated environment to 'confusion', and he attributes the additional contribution in the standard environment to altruism.

Building on Palfrey and Prisbrey (1992) we use a heterogeneous environment in which each individual's marginal rate of substitution is varied over the course of an experiment. By observing a subject's decisions at a number of different marginal rates of substitution, instead of at just one, and by assuming that subjects make errors at some non-negative rate (possibly zero), the subject's entire response function can be estimated. Using the separate techniques of probit and classification analysis, they are able to directly measure the rate of errors in the subject pool, and also to directly measure contributions due to altruism.

The research presented here re-examines the surprising partners–strangers findings of Andreoni (1988) in the heterogeneous environment of Palfrey and Prisbrey (1992), and proposes an explanation consistent with his findings and findings in past experiments. This new explanation combines the 'uncontrolled incentives' rationalization with a statistical model of subject decision errors. The design permits a separation of the three basic effects that have been hypothesized to explain voluntary contribution in experiments, namely altruism, reputation-building, and noise. It also allows direct measurement of experience effects.

3. The independent private values environment

We consider a group of N individuals, each with X_i, a divisible endowment of a private good, and a value for increments of the private good. Each individual must choose an amount of his/her endowment to keep and an amount to give to the public good. The utility of the individual is

$$U(y, x_i) = Vy + r_i x_i ,$$

where V is the value of the public good, y is the amount of the public good produced by the entire group, r_i is the individual's value for the private good, and x_i is the amount of the endowment that is kept for private use. The technology is such that, for every unit of the private good contributed, one unit of the public good is produced.

By varying an individual's r_i over a number of decision periods, it is possible to estimate that individual's decision rule, $D_i(r_i/V)$, where r_i/V is the individual's marginal rate of substitution. Theoretically, an individual's decision rule should be of the following form:

$$D_i(r_i/V) = \begin{cases} 0, & \text{if } r_i/V < 1 + a_i + \varepsilon_i, \\ X_i, & \text{otherwise}, \end{cases}$$

where a_i is individual i's level of altruism, and ε_i is a random error term. This type of decision rule is called a *cut-point rule* and the value $c_i = 1 + a_i$ is called the *cut-point*. Without the error term ε_i and as long as the game does not have an infinite number of decision periods, the above rule is the complete-information, dominant-strategy decision rule. The inclusion of the error term accounts for the possibility of random errors or unpredictable behavior by subjects.

Depending on the assumptions made about the distributions of a_i and ε_i, it is possible to estimate the decision rules in a variety of ways. Possible assumptions about a_i are: (i) all individuals have the same level of altruism and therefore the same a_i; (ii) a_i is never negative; or (iii) q_i is drawn from some distribution. There are also many ways in which ε_i can be distributed, some that assume that all types of errors are equally likely, and others that assume that drastic errors are less likely.

We offer two methods for estimating the decision functions. The first is to use an ordered probit analysis. The ordered probit analysis implicitly assumes that all subjects use the same decision rule and that the ε_i's are distributed in a Normal distribution with mean zero. The assumption of a Normal distribution makes drastic errors (contributing when r_i/V is much larger than c_i), less likely than small errors (contributing when r_i/V is close to c_i). The second method is non-parametric and is called a classification errors analysis. This method is used to estimate individual decision rules.

4. Experimental design

All experiments were run using computers in the experimental economics laboratory at the Universitat Pompeu Fabra. We conducted four experimental sessions, with each session consisting of a sequence of four parameter treatments. There were 12 first-year, undergraduate, economics students who participated in each session, making a total of 48 different subjects.

At the beginning of each session each subject was seated in front of a computer terminal. All terminals were in the same room and were physically isolated from each other with partitions. Subjects were paid in points (1 point = 0.1 Spanish pta.). At the end of a session each subject was paid in private the total amount he/she had earned during the session. Average earnings per subject for the sessions equalled 1266 ptas. and each session lasted a little more than an hour.

Each of the four parameter treatments within a session were conducted as follows. The experimenter read aloud the instructions, which included all of

the following information. Each treatment involved a sequence of 10 decision periods. In each period every subject was endowed with 9 tokens, each of which they could either keep or spend, and each subject was privately assigned a token value (r_i, in our notation) which specified how much each kept token would be worth to that subject in that period. A new token value was drawn for each subject in each period, independently, from a uniform distribution between 1 and 20 points, in 1 point increments. Subjects were not told the other subjects' exact token values.

In each period subjects were assigned into groups of four. In addition to the value of his/her kept tokens, every member of a group earned an amount (V, in our notation) for each token that was spent by any member of the group. This amount, V, was the same for all members of the group and was fixed for the entire 10 periods of a parameter treatment. In the first two parameter treatments of each session V was 6 points, and in the last two of each session V was 10 points. The above information about the distribution of token values and how earnings were determined was explained in great detail, using a table displayed on the board in front of the room and by working through examples. Subjects were then prompted for questions they had about the rules according to which the earnings were determined. Two practice rounds were conducted, in which subjects were instructed to spend a number of tokens equal to the last digit of their subject ID number. During the practice rounds the experimenter carefully went over the keyboard instructions and the screen display for the subjects. At the start of each period the screen displayed for each subject V, r_i, and also displayed a payoff table. At the end of the period, after everyone in the room had made their spending decisions, subjects were told how much each of the other members of their group had spent, and the correct entry in the payoff table was highlighted. The subjects also could access a history screen which kept a record of all information they had received in earlier periods of that experiment. After the practice rounds a quiz was given to the subjects to verify that they understood the basic rules of the experiments, including how token values were assigned and how earnings were computed. A translation (from Spanish) of the instructions and procedures can be found in the appendix of Palfrey and Prisbrey (1993).

In two of the sessions, which, following Andreoni (1988) we call *Strangers*, the subjects were randomly assigned new groups after each decision period. The random assignment process was used to approximate one-time encounters. In the other two sessions, named *Partners*, the subjects were assigned to new groups only between each of the four 10-period parameter treatments; i.e. during a particular 10-period treatment, subjects were repeatedly assigned to the same group. The subjects were told at the beginning of the session whether their groups would be randomly changed between periods or if groupings would remain the same between periods.

This design enables us to examine experience effects, in addition to the effects of partnership. Decisions in the first and third treatments of each session are coded as inexperienced decisions. The rationale for this division is that in the first and third treatments, the subjects see a particular V for the first time. In the second and fourth treatments of each session the subjects see a public good value for the second consecutive time, and these decisions are coded as experienced. No subject participated in more than one session.

5. Analysis of the data

The data analysis centers on the measurement of subject decision rules and is specifically organized around the measurement of *cut-point rules* and *error rates*.

5.1. Aggregate data – a simple classification analysis

As a first cut, we estimate a common cut-point, c, and common error rate, ε, which best describe the aggregate data. The analysis proceeds by determining the rate of classification errors in the data for each possible cut-point. For each token and for each subject, the subject's decision (spend or keep the token) is classified as an error, if, under the hypothetical cut-point rule, the subject should have contributed the token (i.e. the subject had a value r_i/V which was strictly less than the hypothetical cut-point), but the subject did not contribute the token, or if, under the hypothetical cut-point, the subject should not have contributed the token, but did contribute the token. Since each subject is endowed with nine tokens in each period, for every hypothetical cut-point the number of errors we measure for any given subject in any given round can be any non-negative integer less than 10. The estimated common cut-point, c^*, is the hypothetical cut-point with the fewest classification errors, and the estimated common error rate, ε^*, equals the rate of classification errors if c^* is the cutpoint.

Fig. 1 shows the number of classification errors as a function of the hypothetical cut-point, and illustrates the effects of reputation. For both the Partners and the Strangers treatment, the theoretical cut-point with the lowest rate of classification errors is $c^* = 1$, which is consistent with the joint hypotheses of (a) homogeneity of subject decision rules and (b) no altruism in the subject pool.

We next consider the hypothesis suggested by the reputational model, namely that subjects in one-time encounters have a lower cut-point than subjects in repeated encounters. Fig. 1 shows that the c^* in the Strangers

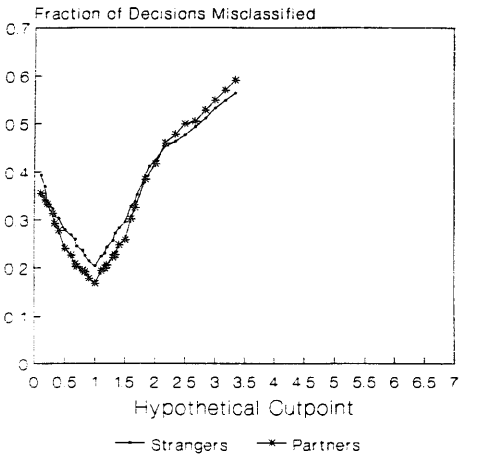

Fig. 1. Cut-point analysis UPF data. Partners vs. Strangers. *Key*: —●—, Strangers: --*--, Partners.

condition is equal to the c^* in the Partners condition, so using this method of decision rule estimation, there is no evidence of a reputation effect.

However, the data show support for an alternative 'noise' hypothesis to account for the differences between the Strangers and Partners data: as seen in Fig. 1, subjects in one-time encounters have a higher error rate than subjects in repeated encounters. Experience reduces error rates in much the same way as partnership (Palfrey and Prisbrey, 1993, fig. 2).

The graphical presentation is further reinforced by a least squares regression with the average group error rate per round, assuming that all subjects use a cut-point of 1 as the dependent variable. The regression contains four independent variables: a constant; *PART*, which is 1 for the Partners data and 0 for the Strangers data; *EXPER*, which is 1 for data from experienced subjects and 0 otherwise; and *PER*, which runs from 1 to 10 and is the number of the period. The results of the regression are shown in Table 1.

The variable *PART* is negative and significant, reflecting the lower average error rates in repeated encounters. The variable *EXPER* is also negative and significant, reflecting the lower error rates in experiments with experienced subjects. The regression also shows that error rates fall over a 10-round session since the coefficient on *PER* is negative and, for a one-tailed test, significant.

Table 1
A least squares regression with the average error rate per round as the dependent variable

Independent variable	Estimated coefficient	t-statistic
Constant	0.241	15.24
PART	−0.035	−2.89
EXPER	−0.032	−2.68
PER	−0.004	−1.89
No. of obs.	160	
R^2	0.11	
$\bar{R}^2$	0.09	

5.2. Aggregate data – an ordered probit analysis

An alternative approach to measuring an 'average decision rule' among the subjects is ordered probit analysis (McKelvey and Zavonia, 1975). The ordered probit analysis estimates the probability of any number of tokens being contributed as a function of the marginal rate of substitution. An advantage of this approach is that it is easy to measure the independent effects of reputation, experience and period using dummy variables, and to summarize these effects in a concise way (see Table 2).

The dependent variable in the analysis is the subject's decision, a number from 0 to 9. The independent variables are: a constant; r/V; *PART* and *PARTS* which are, respectively, constant and slope[7] dummies for the partners treatment; *EXPER* and *EXPERS* which are, respectively, constant and slope dummies for experience effects; and *LATE* and *LATES* which are, respectively, constant and slope dummies for decay effects[8] over a 10-period session.

We calculate a probit response curve equal to the predicted percentage of tokens contributed as a function of r/V and in Fig. 2 plot this curve for several of the treatments. To do this we compute a 'score' for each value of r/V, which determines the location of the mean of a Normal density function on a line divided into intervals by the probit-generated threshold values. In the present situation there are nine intervals, one interval for each of the possible decisions, 0–9. The area below the density and between the thresholds n and $n-1$ is equivalent to the estimated probability that event n occurs. A curve that gives the expected contribution as a function of r/V can then be generated. For comparison we also display in Fig. 3 the aggregate empirical contribution frequencies as a function of r/V.

[7] Slope dummies are the product of the dummy variable and r/V.

[8] Recall that past experiments have observed that contribution rates decay over a 10-period session. The dummy variable *LATE* is 0 in rounds 1–5 and 1 in rounds 6–10.

Table 2
Ordered probit analysis. The dependent variable is the number of tokens contributed. The log-likelihood and sample size are also given

Independent variable	Estimated coefficient	Asymptotic t-statistic
1	1.57	16.27
r/V	−0.66	−11.00
PARTS	−0.18	−2.81
PART	0.17	1.73
EXPERS	−0.15	−2.37
EXPER	0.15	1.48
LATES	−0.18	−2.86
LATE	0.18	1.76
λ_1	0.29	19.61
λ_2	0.55	38.17
λ_3	0.77	59.52
λ_4	0.93	75.23
λ_5	1.09	98.05
λ_6	1.19	105.62
λ_7	1.35	95.63
λ_8	1.57	74.11
Log-likelihood	−3303.2	
N	1920	

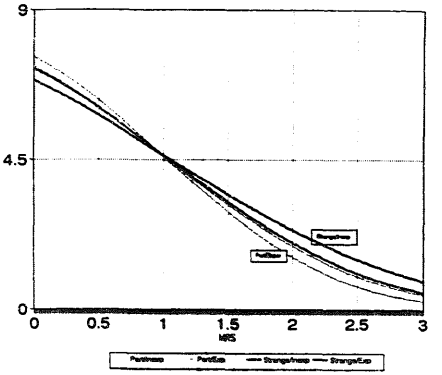

Fig. 2. The expected contribution as a function of r/V as estimated by the ordered probit model.

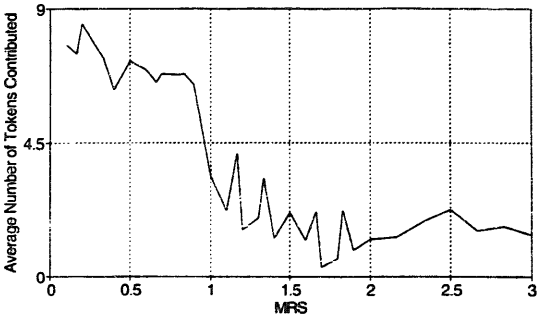

Fig. 3. Empirical contribution rates, aggregated over all treatments and sessions.

If subjects behaved in a way that is perfectly consistent with Nash equilibrium and made no errors, the response function would be graphed as a step function which dropped from 9 to 0 at $r/V = 1$. If the subjects' decision rules are prone to errors, the estimated curve in Fig. 2 would not be a step function, but would be S-shaped.[9] The more errors that are made relative to the average cut-point, the flatter the curve would become. The estimated cut-point based on the probit analysis is equal to the value of r/V at which the predicted contribution is half of the endowment, or 4.5 tokens.

Fig. 2 shows the close proximity of the estimated cut-points for the various treatments (inexperienced vs. experienced and Partners vs. Strangers). All four estimated cut-points are very close (within 0.05) to one. The only difference between the curves is in their slopes. The steepest curve comes from the Partners with experience treatment, the next from the Partners with no experience, the next from the Strangers with experience, and the flattest curve is from the Strangers with no experience treatment. These observations are consistent with the results of the previous section.

The fact that *PARTS* is significant indicates that there is more noise in the one-shot treatments than in repeated encounters. The subjects in one-time encounters have flatter expected contribution curves and therefore have a higher error rate.[10] The variable *EXPERS* is significant, supporting the hypothesis that experience reduces noise: inexperienced subjects have flatter response curves than experienced subjects. The coefficients on *LATE* and

[9] Heterogeneity of subject decision rules can also be a source of flattening of the response curves. The explicit measurement of heterogeneity of cut-points and error rates is conducted in the next subsection.

[10] This could be due either to more individual error, more variance across subjects, or a combination of both. See Subsection 5.3.

LATES mirror these results, indicating that the effect of the 10-period repetition is similar to experience effects. The response curves are steeper in the last half of a 10-period session than in the first half, but the average contribution rate is unchanged.[11] At first glance this would seem to contradict past findings of significant decay. But in fact there is no contradiction at all. It simply means that the observed decay in past experiments was due to learning, not reputation.[12]

This lack of reputation effects is further documented in Table 3, where the effect of *PART* and *LATE* on average contributions is cross-tabulated. Reputation effects would predict *more* decay in the Partners treatment than in the Strangers treatment. In fact, the opposite is observed (although the difference is not statistically significant at the 5% level).

5.3. *Individual data – classification analysis*

The probit analysis reported above is carried out under a maintained hypothesis of homogeneity of subject decision rules. While that approach has the virtue of providing a concise summary of the aggregate features of the data, we cannot use that approach to identify the relative contribution of *hetereogenity* and *subject error* to the flatness of the response curves (i.e. the 'noise' in the data). To identify those two sources of noise, it is necessary to analyze the data at the individual level and explicitly allow for heterogeneity of decision rules across subjects.

In this subsection we apply the simple classification analysis of Subsection 5.1 at the individual level. By doing so we are able to estimate a *distribution*[13] of cut-points across the entire subject pool. From these

Table 3
Mean contribution (out of 9 tokens) as a function of *LATE* and *PART*. $N = 480$ in each cell

	Partners	Strangers
Early ($t = 1 - 5$)	3.39	3.78
Late ($t = 6 - 10$)	3.53	3.64

[11] The average contribution rate in periods 1–5 is 3.583 and the average contribution rate in periods 6–10 is 3.585.

[12] If we censor all our observations with $MRS < 1$, then indeed we also measure significant decay that is large in magnitude.

[13] Rapoport (1987) has argued that heterogeneity may be an important ingredient of a complete explanation for behavior in other (step-level) public goods environments. Isaac et al. (1984), Ledyard (1993), and Palfrey and Rosenthal (1994) make similar points.

estimated cut-points we can compute error rates for each individual as the percentage of decisions that violate their estimated cut-point rule. The distributions of error rates and the distribution of cut-points are then compared across treatments.

5.3.1. The distribution of individual cut-points

Fig. 4 shows the distribution of estimated individual cut-points across the 192 observations.[14] The distribution[15] is centered at 0 (i.e. Nash cut-points) and is nearly symmetric. The median cut-point is 0 and accounts for approximately 30% of the observations. Two-thirds of the observations range from −3 to +3, with the remaining one-third evenly divided below −3 and above +3. Three-quarters of the observations range from −4 to +4, again with the remainder being evenly divided between large negative and large positive cut-points. Consistent with the probit analysis, we find that *on average* subjects are neither altruistic nor spiteful. By this we do not mean that we find no subject behaving altruistically. There are, in roughly equal

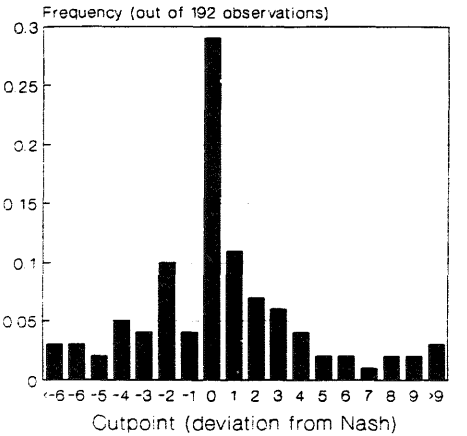

Fig. 4. Classification of error minimizing cut-points, UPF data for series 1.

[14] For each of our 48 subjects we report four separate 'observations' corresponding to the four treatments that a subject participated in: low-V – inexperienced; low-V – experienced; high-V – inexperienced; and high-V – experienced.

[15] Deviations from Nash cut-points are measured in token value units. A cutpoint of 0 corresponds to $MRS = 1$ in earlier figures. In a few of the observations more than one hypothetical cut-point minimized classification errors. Such ties were broken by choosing the one closest to 0.

numbers, both altruists (subjects with positive cut-points) and spiteful subjects (with negative cut-points) in this subject pool. However, to the extent that we observe these deviations from 0 cut-points, those deviations are typically small in magnitude.

If we break down the distribution of cut-points by the Partners/Strangers treatment, we find a systematic effect, but not what we would expect from the hypothesis that repeated groups have 'reputation effects' that lead to more contribution. The reputation hypothesis predicts that repeated groups will have cut-points that are typically higher than the cut-points in the one-shot treatment. *We do not find this.* The average or median cut-point in both treatments equals 0. The difference between the two distributions is that the distribution for Strangers is more dispersed than the distribution for Partners. This is illustrated in Fig. 5, which displays the empirical cumulative frequencies separately for the Strangers data and the Partners data.

5.3.2. The distribution of classification errors

From the above classification analysis we can also obtain estimates for the distribution of classification errors across individuals. The error rate we compute is the fraction of an individual's decisions (within one treatment) that are misclassified according to that individual's estimated cut-point. Over 20% of the time subjects can be perfectly classified; 60% of the time we measure error rates below 10%; and 25% of the error rates fall between 10% and 20%. Fig. 6 shows the effect of experience on error rates. There is a leftward shift in the error rate distribution, indicating fewer errors with experience. Error rates are also systematically lower in the Partners treatment than in the Strangers treatment (see Palfrey and Prisbrey, 1993, fig. 8).

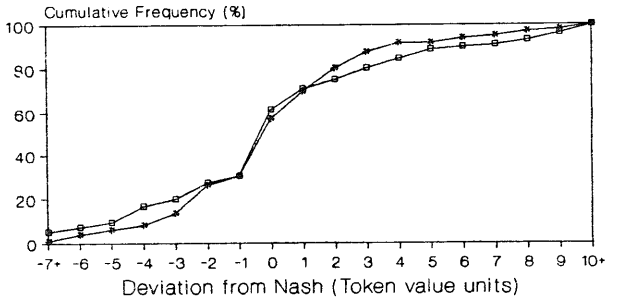

Fig. 5. Individual cut-points, UPF data. Partners vs. Strangers. *Key*: —*—, Partners; —□—, Strangers.

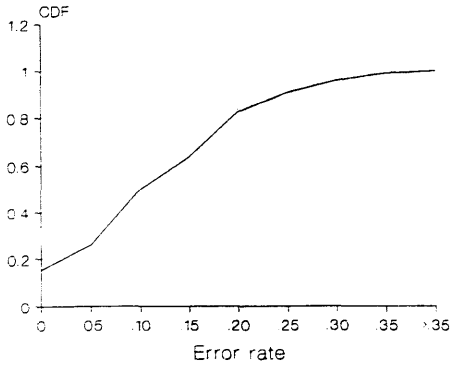

Fig. 6. Fraction of decisions misclassified relative to estimated cut-point (cumulative frequency distribution). *Key*: --------, experienced ($N = 96$); ——, inexperienced ($N = 96$).

6. Conclusions

The results in this paper point to a new interpretation of observed violations of dominant strategies to free ride in voluntary contributions experiments. The explanation we suggest is not that subjects are on average either particularly altruistic or particularly spiteful. Consistent with Andreoni (1988) we find no evidence of reputational effects of the sort proposed in Kreps and Wilson (1982) and others. Rather, subjects exhibit statistical fluctuations in their decision-making, manifested as random noise[16] in the data. This noise has both a heterogeneity component and an individual subject error component. This explanation is consistent with our data, both at the aggregate level and at the individual level.

How does such an explanation account for the apparently altruistic behavior in past experiments where subjects have a dominant strategy to free ride? The answer we propose is that in those experiments the design automatically censors all observations of subjects who have a dominant strategy to give, but end up free riding. In other words, in past experiments the only kind of observable 'error' relative to Nash theory was seemingly altruistic behavior. If we re-examine our data *censoring all observations of* $MRS < 1$ (dominant strategy to give), then we find aggregate contribution rates that are statistically significant, and of a magnitude comparable with

[16] Presumably these statistical fluctuations are not purely random from the point of view of a subject making the decision.

what has been found in these other studies. Moreover, as in Andreoni (1988) we find more contribution in the Strangers treatment than in the Partners treatment. We are able to show that this difference is due to factors that affect the *variance* in subjects' decisions and decision rules, not a systematic tendency of *mean* behavior away from the Nash equilibrium. A similar explanation applies to the Saijo and Nakamura (1993) experiments where subjects have a dominant strategy to give, but substantial free riding is observed. The observation that experience reduces violations is just a reflection that experience produces lower error rates and lower subject variation.

Acknowledgements

We acknowledge the financial support of the National Science Foundation (SBR-9223701) and the Ministerio de Education y Ciencia (DGICYT PB91-0810). We thank Estela Hopenhayn for assistance in preparing and conducting the experiments. Antonio Rangel helped with the translation of instructions from English. We are grateful to our colleagues at both Caltech and Pompeu Fabra for their advice, with speciai thanks to Antoni Bosch. The comments from two referees are gratefully acknowledged.

References

Andreoni, J., 1988, Why free ride? Strategies and learning in public goods experiments, Journal of Public Economics 37, 291–304.

Andreoni, J., 1992, Cooperation in public goods experiments: Kindness or confusion?, typescript.

Dawes, R.M., 1980, Social dilemmas, Annual Review of Psychology 31, 169–193.

Isaac, R.M. and J.M. Walker, 1988, Groups size effects in public goods provision: The voluntary contributions mechanism, The Quarterly Journal of Economics 103, 179–201.

Isaac, R.M., J.M. Walker and S.H. Thomas, 1984, Divergent evidence on free riding: An experimental examination of possible explanations, Public Choice 43, 113–149.

Kreps, D.M. and R. Wilson, 1982, Reputation and imperfect information, Journal of Economic Theory 27, 253–279.

Kreps, D.M., P. Milgrom, J. Roberts and R. Wilson, 1982, Rational cooperation in the finitely repeated prisoners' dilemma, Journal of Economic Theory 27, 245–252.

Ledyard, J.O., 1993, Public goods: A survey of experimental research, in: J. Kagel and R. Roth, eds., Handbook of experimental economics, Princeton University Press, Princeton, NJ.

Marwell, G. and R.E. Ames, 1979, Experiments on the provision of public goods. I. Resources, interest, group size, and the free-rider problem, American Journal of Psychology 84, 1335–1360.

Marwell, G. and R.E. Ames, 1980, Experiments on the provision of public goods. II. Provision points, stakes, experience and the free rider problem, American Journal of Psychology 85, 926–937.

Marwell, G. and R.E. Ames, 1981, Economists free ride, does anybody else? Experiments on the provision of public goods, IV, Journal of Public Economics 15, 295–310.

McKelvey, R.D. and W. Zavonia, 1975, A statistical model for the analysis of ordered level dependent variables, Journal of Mathematical Sociology 4, 103–120.

Palfrey, T.R. and J.E. Prisbrey, 1992, Anomalous behavior in linear public goods experiments: How much and why?, Social Science Working Paper no. 833, California Institute of Technology.

Palfrey, T.R. and J.E. Prisbrey, 1993, Altruism, reputation, and noise in linear public goods experiments, Social Science Working Paper no. 864, California Institute of Technology.

Palfrey, T.R. and H. Rosenthal, 1988, Private incentives and social dilemmas: The effects of incomplete information and altruism, Journal of Public Economics 28, 309–332.

Palfrey, T.R. and H. Rosenthal, 1991, Testing game-theoritic models of free riding: New evidence on probability bias and learning, in: T. Palfrey, ed. Laboratory research in political economy, (University of Michigan Press, Ann Arbor) 239–268.

Palfrey, T.R. and H. Rosenthal, 1994, Repeated play, cooperation, and coordination: An experimental study, Review of Economic Studies 61, 545–565.

Rapoport, A., 1987, Research paradigms and expected utility models for the provision of step-level public goods, Psychological Review 94, 74–83.

Saijo, T. and H. Nakamura, 1993, The Spite dilemma in voluntary contributions mechanism experiments, unpublished manuscript.