10 Can public goods experiments inform policy?

Interpreting results in the presence of confused subjects

Stephen J. Cotten, Paul J. Ferraro, and Christian A. Vossler

Introduction

Public policy is frequently used to induce individuals to contribute to public goods when it may be in their private interests to free-ride off the contributions of others. To explore how individuals behave in various public goods decision settings and to gain insights into how institutions might be better designed to encourage the provision of public goods, economists employ laboratory experiments. The cornerstone of experimental investigations on the private provision of public goods is the voluntary contributions mechanism (VCM). Understanding behavior in experimental implementations of the VCM game is critical for the work of economists with institutional and policy-oriented interests.

The standard linear VCM experiment places individuals in a context-free setting where the public good, which is non-rival and non-excludable in consumption, is simply money. Participants are given an endowment of "tokens" to be divided between a private account and a public account. Contributions to the private account are converted to cash and given to the individual. Contributions to the public account yield a cash return to all group members, including the contributor. If the marginal return from contributing a token to the public account is less than the value of a token kept in the private account, but the sum of the marginal returns to the group is greater than the value of a token kept, the individually rational contribution is zero (i.e. the individual free rides) while the social optimum is realized when everyone contributes their entire endowment to the public account.

In single-round VCM experiments where a public good contribution rate of zero is the unique Nash equilibrium, subjects contribute at levels far above this: on average, 40–60 percent of endowments. In repeated-round VCM experiments, contributions start in the range of 40–60 percent but then decay towards zero (ending around 10 percent of endowments on average). Thus, there seem to be motives for contributing that outweigh the incentive to free ride.

Possible motives underlying contributions include: (1) "pure altruism" (sometimes called "inter-dependent utility"), which describes a situation in which an individual's utility function is a function of his own payoff and the payoffs of her group members; (2) "warm-glow" (often called "impure altruism"; Andreoni, 1990), which describes a situation in which an individual gains utility from the simple act of contributing to a publicly spirited cause; and (3) "conditional cooperation" (Andreoni, 1988; Fischbacher *et al.*, 2001), which is a predisposition to contribute in social dilemmas but punish by revoking contributions when significant free riding behavior is observed. A fourth motive, which the VCM literature often ignores but we are particularly interested in, is "confusion." We define confusion as behavior that stems from the failure of an individual to identify the dominant strategy of zero contributions. More broadly, confusion behavior results from a failure for individuals to discern the nature of the game, and individuals do not understand how to utility-maximize in the context of the game.

Investigations into the identification and relative importance of various motives for contributions in the VCM game and closely related games have led to conflicting conclusions. For example, Palfrey and Prisbrey (1997) find statistical evidence of warm-glow but no evidence of pure altruism; Goeree *et al.* (2002) find the opposite. Fischbacher *et al.* (2001) and Fischbacher and Gächter (2004) find no evidence of pure altruism or warm-glow, but find significant conditional cooperation.

Efforts to compare public goods contributions across different subpopulations have likewise led to mixed results. A particularly well-studied issue is whether contributions behavior differs between men and women (Eckel and Grossman, 2005). Brown-Kruse and Hummels (1993) find that men contribute more than women. Nowell and Tinkler (1994) find females are more cooperative. Cadsby and Maynes (1998) find no significant differences between men and women.

Particularly troubling is the apparent lack of correspondence between contributions behavior in experimental and naturally occurring settings. Whereas many studies find that economics students or economists are less likely to contribute to public goods in experiments (e.g. Marwell and Ames, 1981; Cadsby and Maynes, 1998), attempts at externally validating this claim yield contradictory results (Yezer *et al.*, 1996; Laband and Beil, 1999; Frey and Meier, 2004). Laury and Taylor (forthcoming) use behavior in a one-shot VCM experiment to predict behavior in a situation in which individuals can contribute to an urban tree planting program, a public good. Using the empirical approach of Goeree *et al.* to estimate a pure altruism parameter for each subject, the authors find little or no relationship between subjects' altruism parameters and subjects' contributions to urban tree planting.

There are several possible reasons for these puzzling results, including differences in the design, implementation, and participants used in these experiments. We focus on an alternative explanation: confusion confounds the interpretation of behavior in public goods experiments. This chapter presents results from one new experiment and two previous experiments that use the

"virtual-player method," a novel methodology for detecting confusion through a split-sample design where some participants play with non-human players (automata) that undertake predetermined strategies or choices. Each experiment involves a slightly different public goods game and a different subject pool (with presumably differing abilities). The level of confusion in all experiments is both substantial and troubling. These experiments provide evidence that confusion is a confounding factor in investigations that discriminate among motives for public goods contributions, in studies that compare behavior across subpopulations, in research that assesses the external validity of experiments, and in attempts to use experimental results to improve policy design. We conclude by proposing ways to mitigate confusion in standard public goods experiments, and present results from a pilot study that uses "context-enhanced" instructions.

Prior evidence of confusion in public goods experiments

Andreoni (1995) was the first to identify and test the hypothesis that confusion plays an important role in the contributions decisions of participants in public goods games. Specifically, Andreoni (p. 893) hypothesizes

> that the experimenters may have failed to convey adequately the incentives to the subjects, perhaps through poorly prepared instructions or inadequate monetary rewards, or simply that many subjects are incapable of deducing the dominant strategy through the course of the experiment.

To test his *confusion* hypothesis, Andreoni developed a VCM-like game that fixes the pool of payoffs and pays subjects according to their contributions to the public good. The person who contributes the least is paid the most from the fixed pool. Thus contributions to the "public good" in this game do not increase aggregate benefits, but merely cost the contributor and benefit the other group members. Andreoni uses behavior from the ranking games to infer that both other-regarding behavior and confusion are "equally important" motives in the VCM.

Houser and Kurzban (2002) continued Andreoni's (1995) work with a clever experimental design that includes: (1) a "human condition," which is the standard VCM game; and (2) a "computer condition," which is similar to a standard VCM game except that each group consists of one human player and three non-human computer players (which we refer to as "virtual players") and the human players are aware they are playing with computer players. In each round the aggregate computer contribution to the public good is three-quarters of the average aggregate contribution observed for that round in the human condition. By making the reasonable assumption that other-regarding preferences and confusion are present in the human condition, but only confusion is present in the computer condition, Houser and Kurzban find that confusion accounts for 54 percent of all public good contributions in the standard VCM game.

Ferraro *et al.* (2003) independently developed a similar design with virtual players, and applied it to a single-round VCM game. They find that approximately 54 percent of contributions are due to confusion. Ferraro and Vossler (2005) extend this design to the multi-round VCM, where they find that 52 percent of contributions across rounds stem from confusion. However, unlike Hauser and Kurzban, they present evidence showing that this confusion does not decline with experience. This difference stems from the atypical behavior that Hauser and Kurzban's all-human condition exhibits (little decline in contribution) and two other aspects of their design that make it difficult to directly compare the human and computer conditions.[1]

Palfrey and Prisbrey (1997) developed an experimental design that, when combined with a few behavioral assumptions, allows the authors to separate the effects of pure altruism, warm-glow, and confusion. Their design changes the standard VCM game by randomly assigning different rates of return from private consumption each round, which enables the measurement of individual contribution rates as a function of that player's investment costs. The authors conclude that (p. 842) "altruism played little or no role at all in the individual's decision and, on the other hand, warm-glow effects and random error played both important and significant roles." While no point estimate was given of the proportion of contributions stemming from confusion in their experiment, the authors use their model results to predict that "well over half" of contributions in the seminal VCM experiments by Isaac *et al.* (1984) are attributable to error.

Goeree *et al.* (2002) use a VCM design in which group size is either two or four and the "internal" return of a subject's contribution to the public good to the subject may differ from the "external" return of the same contribution to the other group members. The authors estimate a logit choice model of noisy decision making with data from a series of one-shot VCM games (no feedback) in which the internal and external returns are varied. They find that coefficients corresponding to pure altruism and decision error are both positive and significant. Similar to Palfrey and Prisbrey, no estimate of the fraction of contributions due to confusion is given.

Fischbacher and Gächter (2004) design an experiment to test specifically for the presence of conditional cooperation. In Fischbacher and Gächter's "P-experiment," they ask subjects to specify, for each average contribution level of the other group members, how much they would contribute to the public good. By comparing the responses in this experiment with those in their "C-experiment," which is a standard VCM game with four-person groups, Fischbacher and Gächter argue that most contributions come from conditional cooperators. They find no evidence of pure altruism or warm-glow (no subjects stated they would contribute if other group members contributed zero). In contrast to previous work, they claim confusion accounts for a smaller fraction of observed contributions to the public good, "at most 17.5 percent" (p. 3).

Overall, four of the five studies above that assess magnitude find that about half of all contributions stem from confusion. This conclusion is alarming. In response to the fifth study, Ferraro and Vossler point out that Fischbacher and

Gächter's characterization of conditional cooperators also describes the behavior of confused "herders" who simply use the contributions of others as a signal of the private payoff-maximizing strategy. As such, the proportion of confusion contributions (17.5 percent) found by Fischbacher and Gächter may be best characterized as a lower bound estimate. Despite this dispute, the research on confusion in public goods experiments can be succinctly summarized: *every study that looks for confusion finds that it plays a significant role in observed contributions.*

The virtual-player method

The virtual-player method discriminates between confusion and other-regarding behavior in single-round public goods experiments, and discriminates between confusion and other-regarding behavior or self-interested strategic play in multiple-round experiments (see Ferraro *et al.* (2003) for other applications). The method relies on three important features: (1) the introduction of non-human, virtual players (i.e. automata) that are preprogrammed to exercise decisions made by human players in an otherwise comparable treatment; (2) a split-sample design where each participant is randomly assigned to play with humans (the "all-human treatment") or with virtual players (the "virtual-player treatment"); and (3) a procedure that ensures that human participants understand how the non-human, virtual players behave.

The virtual-player method makes each human subject aware that he or she is grouped with virtual players that do not receive payoffs and that make decisions that are exogenous to those of the human. Thus the method neutralizes the other-regarding components of the human participant's utility function and the motives for strategic play.[2] Thus, as long as participants understand their decision environment, any contributions made by humans in virtual-player groups can be attributed to confusion in the linear VCM game.

The random assignment of participants to an all-human group or a virtual-player group allows the researcher to net out confusion contributions by subtracting contributions from (human) participants in the virtual-player treatment from contributions in the all-human treatment. In single-round experiments where the decisions of other players are not known *ex ante*, the contributions from virtual players should have no effect on human contributions nor should they confound any comparison between all-human and virtual-player treatments. Thus, one can argue that randomly selecting the profile of any previous human participant, with replacement, as the contribution profile for a virtual player suffices to ensure comparability.

However, in the typical multiple-round public goods game where group contributions levels are announced after each period, it is important to exercise additional control as the history of play may affect contributions in the virtual-player treatment. Indeed, Ferraro and Vossler find that confused individuals use past contributions of virtual players as signals of how much to contribute. The additional control comes by establishing that each human in the all-human treatment has a human "twin" in the virtual-player treatment: each twin sees exactly the same contributions by the other members of her group in each round. Thus, the only difference between the two treatments is that the player in the virtual-player treatment knows she is playing with preprogrammed virtual players, not humans.

To illustrate, consider a game that involves repeated interactions with groups consisting of three players. Participants in a group in the all-human treatment are labeled as H1, H2, and H3. Subject V1 in the virtual-player treatment plays with two virtual players: one that makes the same choices H2 made in the all-human treatment, and the other that makes the same choices H3 made. Likewise, player V2 plays with two virtual players, one playing exactly like human subject H1 and the other playing exactly like H3. And so on. Note that having an imbalance between all-human and virtual-player treatments, which would occur if some participants do not have a "twin" or if a player in one treatment has multiple twins in other, confounds comparisons.

To ensure that participants in the virtual-player treatment believe the virtual-player contributions are truly preprogrammed and exogenous, each subject has a sealed envelope in front of her. The participants are told that inside the envelope are the choices for each round from the virtual players in their groups. At the end of the experiment, they can open the envelope and verify that the history of virtual group member contributions that they observed during the experiment is indeed the same as in the envelope. The subjects are informed that the reason we provide this envelope is to prove to them that there is no deception: the virtual players behave exactly as the moderator explained they do. Post-experiment questionnaires are useful at assessing whether participants fully understand the nature of virtual players.

Application of the virtual-player method to the Goeree, Holt, and Laury experiment

The experiment of Goeree, Holt, and Laury (hereafter referred to as "GHL") is a variant of the static linear VCM game that endeavors to test the significance and magnitudes of contributions stemming from pure altruism and warm-glow. Each participant decides how to allocate 25 tokens between a private and a public account in each of ten "one-shot" decision tasks (referred to as "choices" in instructions), without feedback, where the internal (m_i) and external rates (m_e) of return, and group size (n), vary across tasks. For each decision task, a token kept in the private account yielded 5 cents. The internal rate of return refers to the marginal return to oneself from a token contributed to the public account, and ranged from 2–12 cents. The external rate of return refers to the marginal return to other players from one's contribution to the public account, and was either 2 or 4 cents. Group size was either two or four players. Formally, the profit function of the individual i (in cents) for a particular decision task is given by

$$\pi_i = 5(25 - x_i) + m_i x_i + m_e \sum_{i \neq j}^{n} x_j$$

where $x_i \in [0, 25]$ denotes public account contributions from player i. Since the internal rate of return in GHL is always lower than the value of a token kept, it is still the individual's dominant strategy to contribute nothing. The sum of the external and internal rates of return is always greater than 5 cents, so that full endowment contribution maximizes group earnings.

In the typical one-shot VCM, the external and internal rates of return are equal ($m_i = m_e$), i.e. all players receive the same return from the public good. The rates are varied in the GHL design because participants exhibiting pure altruism should increase their contributions when the external return or the group size increases. Such systematic correlations should be identifiable by observing patterns in individual contributions across the various decision settings. If considerable contributions are observed, but they show little correlation with external return and group size, the conjecture is that contributions are largely attributable to warm-glow.

We replicate the GHL experiment using the virtual-player method to explore whether conclusions drawn from the original study are robust after quantifying and netting out confusion contributions. We made two small changes in the way subjects were grouped and paid. GHL assign subjects to two- and four-member groups by selecting marked ping-pong balls after all decisions are made. We pre-assign participants to two- and four-member groups based on their subject ID number. This is important for virtual-player sessions as it allows us to give each participant an envelope with the aggregate contributions of other players as well as earnings from virtual-player contributions for each possible decision selected. The pre-assignment into groups shortens the length of both all-human and virtual-player treatments. GHL randomly choose only one of the ten decisions to be binding using the roll of a ten-sided die and use a second, unrelated experiment to supplement earnings. Rather than engage our participants in a second experiment, we pay participants based on three randomly chosen decisions instead of one. This change increases the saliency of each decision.

Experiment instructions are presented both orally and in writing. The instructions for all-human and virtual-player treatments are available from the authors on request. The all-human instructions are from GHL, with minor revisions. The virtual-player instructions are similar with the exception of emphasizing that participants are matched with virtual players, whose contributions are predetermined. As in GHL, participants make decisions via paper and pencil. Decision sheets are identical to GHL. Following the decisions, a post-experiment questionnaire is given to collect basic demographic information as well as to assess understanding of the experimental design and decision tasks.

A total of 53 participants were recruited from a pool of undergraduate student volunteers at the University of Tennessee in the Spring of 2005. Of these, 23 students participated in the all-human treatment, which serves as a replication of the GHL design, whereas 30 students participated in the virtual-player treatment.[3] Experiment sessions consisted of groups ranging from four to 12 people, and participants were visually isolated through the use of dividers. Matching was anonymous; subjects were not aware of the identity of the other members of their group(s). All sessions took place in a designated experimental economics laboratory. Earnings ranged from $8 to $15 and the experiment lasted no more than 1 hour.

Results

Goeree, Holt, and Laury application

Table 10.1 presents mean and median contributions from the all-human treatment, which serves as a replication of the GHL study. The pattern of contributions in relation to design factors is quite similar between this study and the GHL study, with contributions generally increasing with respect to external return and group size. This suggests that pure altruism is an important motive.

To quantify formally the magnitude of altruism and warm-glow, GHL consider different theoretical specifications for individual utility and estimate utility function parameters using a logit equilibrium model. For the sake of parsimony, we refer the interested reader to the GHL study for details. We estimate logit equilibrium models with our data and concentrate on interpretations of estimated parameters and comparisons of parameters across treatments.

Estimated logit equilibrium models are presented in Table 10.3 for all-human and virtual-player treatments. The "altruism" model considers the altruism motive but not warm-glow, the "warm-glow" model considers warm-glow but not altruism, and the "combined" model considers both motives. Consistent with the contributions pattern observed in Table 10.1, the logit equilibrium model results for the all-human treatment suggest that pure altruism is an important motive. In particular, the parameter α is a measure of pure altruism, and we find this parameter to be statistically different from zero using a 5 percent significance level. Our estimates suggest that a participant is willing to give up between 5 cents ("altruism" model) and 15 cents ("combined" model) in order to increase another person's earnings by $1. The parameter g measures warm-glow, which we find to be insignificant. The parameter μ is an error parameter. While μ measures dispersion and does not indicate the magnitude of confusion contributions, statistical significance of this parameter does indicate decision error is

Table 10.1 GHL application, all-human treatment results

	Decision task									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Group size	4	2	4	4	2	4	2	2	4	2
Internal return	4	4	4	2	4	4	2	4	2	4
External return	2	4	6	2	6	4	6	2	6	12
Mean	9.2	10.1	10.8	5.2	9.7	9.9	6.5	5.2	8.7	12.3
Median	5	10	11	4	9	9	5	3	6	12

present (Goeree *et al.*). Estimates of μ are indeed statistically different from zero at the 5 percent level for each specification.

Overall, the main conclusions drawn from GHL carry over in our all-human treatment model: pure altruism and confusion are important motives behind contributions whereas warm-glow is not. We now discuss the outcome from the virtual-player treatment and present two main results about the role of confusion.

Result 1: Positive contributions stem largely from confusion and subjects use experimental parameters as cues to guide payoff-maximizing contributions, leading to behavior that mimics behavior motivated by pure altruism.

Contributions in the virtual-player treatment, presented as Table 10.2, are generally smaller than in the all-human treatment but not strikingly so. Specifically, mean contributions across all decision tasks are 6.7 tokens or 27 percent of endowment in the virtual-player treatment as compared to 8.8 tokens or 35 percent in the all-human treatment. Put another way, virtual-player contributions are approximately 75 percent of all-human contributions. Assuming that other-regarding preferences and confusion are present in the all-human treatment, but that only confusion exists in the virtual-player treatment, this suggests that an alarming 75 percent of all-human treatment contributions stem from confusion.

Perhaps more startling is the observed correspondence between all-human and virtual-player treatment contributions across decision tasks, as illustrated in Figure 10.1. From Figure 10.1, one observes that subjects in the virtual treatment alter their contributions based on the same stimuli as subjects in the all-human treatment; the two response patterns are parallel such that the difference between sets of contributions across decision tasks are approximately equal.

Turning to the logit equilibrium models estimated from virtual-player treatment data, we find that estimated pure altruism parameters are statistically different from zero. In particular, we find that a participant is willing to give up between 4 cents ("altruism" model) and 16 cents ("combined" model) in order to increase a virtual player's earnings by $1. Using the estimated "altruism" and "combined" models from the two treatments we test for equality of altruism parameters between the two (leaving other parameters unconstrained) using Wald Tests. For both specifications we fail to reject the hypothesis of equal

Table 10.2 GHL application, virtual-player treatment results

	Decision task									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Group size	4	2	4	4	2	4	2	2	4	2
Internal return	4	4	4	2	4	4	2	4	2	4
External return	2	4	6	2	6	4	6	2	6	12
Mean	6.1	6.9	9.1	2.7	7.7	7.7	4.4	4.1	7.2	10.8
Median	5	6	7	0	7.5	5.5	2	3.5	5	10.5

Table 10.3 GHL application, estimated logit equilibrium models

	All-human treatment			*Virtual-player treatment*		
	Altruism	*Warm-glow*	*Combined*	*Altruism*	*Warm-glow*	*Combined*
α	0.054*	–	0.148*	0.034*	–	0.163*
	(0.021)		(0.064)	(0.014)		(0.050)
g	–	–0.470	–1.583	–	–1.231	–2.383*
		(0.769)	(1.059)		(0.796)	(0.987)
μ	19.310*	32.382*	28.269*	11.914*	24.801*	21.132*
	(3.447)	(11.628)	(9.054)	(1.460)	(7.150)	(5.311)
Log-L	–671.497	–673.071	–668.718	–824.510	–823.148	–813.308
N	230	230	230	300	300	300

Notes
Standard errors in parentheses.
* indicates parameter is statistically different from zero at the 5 percent level.

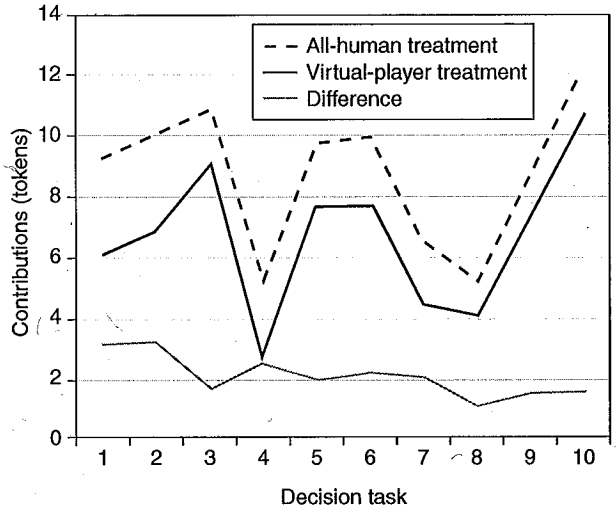

Figure 10.1 GHL application, comparison of all-human and virtual-player contributions.

altruism parameters ("altruism" model: $\chi^2 = 0.647$, $p = 0.421$; "combined" model: $\chi^2 = 0.033$, $p = 0.855$). Of course, by design, participants in the virtual-player treatment are not exhibiting pure altruism, unless one believes pure altruism includes preferences over the utility of fictional automata.

Why would virtual-player participants respond to the same stimuli as all-human treatment participants? Confused subjects are using the changes in the parameters across decision tasks as a cue of how to behave. The altruism parameter is picking up confusion about the role of the external return in the subject's private payoff

function. In a confusing situation, most people look for cues to direct them towards the optimal behavior. In the GHL experiment, subjects have to make ten contributions decisions for which the internal and external rates of return, and group size, are all changing. It should not be surprising that a confused subject will infer meaning from the changes in these parameters and decide that her behavior ought to change in response to them. This behavior is similar in spirit to the "herding" behavior found in the Ferraro and Vossler dynamic VCM experiment. In this experiment, subjects use past group member contributions (human or non-human) as a cue of how to choose their own optimal responses.

Recall that Laury and Taylor (forthcoming) run subjects through a GHL experiment and then ask participants to contribute to an urban tree-planting program. Subjects with positive altruism parameters are found to be *less* likely to contribute to the naturally occurring public good, even after controlling for experimental earnings and subject demographics and attitudes. Based on our logit equilibrium model results and observed correspondence between contributions both virtual-player and all-human treatment contributions to changes in the marginal per capita return (MPCR), it is quite likely confusion confounds their comparison.

Result 2: The common observation in public goods experiments that contributions increase with increases in the marginal per capita return likely results from subject confusion rather than altruism or expectations about the minimum profitable coalition.

Decision Task 4 and Decision Task 6 involve a group size of four. The internal return is equal to the external return, but these returns increase from two to four across the two tasks. Thus, the lone design difference is analogous to a change in the MPCR in standard VCM experiments. In particular, the MPCR doubles from 0.4 to 0.8 from Decision Task 4 to Decision Task 6. A "stylized fact" from the experimental public goods literature is that an increase in the MPCR increases contributions, which has been attributed to altruism and "minimum profitable coalitions" (Cox and Sadiraj, 2005).[4] In the all-human treatment, mean contributions are 5.2 in Decision Task 4 and 9.9 in Decision Task 6 – an increase of 4.7 tokens – which is consistent with the results on MPCR changes in the literature. In the virtual-player treatment, contributions go from 2.7 to 7.7, which is a nearly identical change of 5.0 tokens. Thus our results are consistent with the "MPCR effect" being related to confusion.

Such pervasive evidence of confusion may cause readers to doubt the validity of the virtual-player method. In addition to the emphases placed in the instructions and the use of the sealed envelope, we also used a post-experiment questionnaire. We asked all subjects to answer the following question:

> If all you cared about was making as much money as possible for *yourself*, how many tokens should you have invested in each decision? (you may not have cared about making as much money as possible for yourself, but if you did, what is the correct answer?).

Subjects were aware that they would be paid $1.50 for a correct answer. A total of 13 out of 53 answered this question incorrectly, suggesting that 25 percent of respondents were unable to discern the dominant strategy of zero contributions after participating in the experiment (note that this is a lower bound given that some subjects may only realize the correct answer after being asked the question and, as noted in Ferraro and Vossler, other subjects who erroneously believe they are playing an assurance game will often answer "zero" to this question).

For those in the all-human treatment we asked respondents to state the contributions level that would have maximized group earnings. All participants correctly stated 25 tokens or full endowment. Thus, it appears that an important issue with the public goods game is that some self-interested individuals are simply not able to deduce the dominant strategy. Since decision errors can only be made in one direction (contributions are non-negative), this confusion necessarily leads to what looks like other-regarding behavior.

Ferraro, Rondeau, and Poe

We draw from previous experiments to strengthen our arguments about the confusion problem in public goods experiments. The first experiment is from Ferraro, Rondeau, and Poe (2003), who use the virtual-player method to study behavior in a single-round VCM-like game. Group size is 21, individual endowment is $12, MPCR is $0.07, and there is a cap on returns from the public good of $7 each. Thus, while the social optimum is for the group to contribute $100 (divided equally this is $4.76 each) – rather than full endowment – the dominant strategy is still for the individual to contribute nothing.

This study uses "Ivy League," Cornell University undergraduates from an introductory economics class, whom all have prior experience in experiments. Total sample size is 85. As stated by Ferraro *et al.* (p. 103), "our subject pool can be considered an 'extreme' environment in which to search for altruistic preferences: subjects were 'economists in training,' operating in an environment in which self-interest was being reinforced." Results from this experiment are presented in Table 10.4. Using the entire sample, all-human treatment contributions are $2.14 and virtual-player treatment contributions are $1.16, such that we estimate confusion accounts for 54 percent of contributions. Thus, even with some of the world's brightest young individuals as subjects, it appears as though confusion contributions are quite substantial.

As discussed in the Introduction, public goods experiments are often used to make inferences about the behaviors of subgroups in the population (by gender, race, culture, etc.). We use raw data from this experiment to analyze further behavior according to gender (not reported in the original article). Table 10.4 presents mean contributions by gender and treatment. Based on the all-human treatment results, contributions from females are $0.92 higher than males and this difference is statistically significant using a Mann-Whitney Test ($p = 0.07$). However, virtual-player treatment contributions are also larger for females by $1.24 ($p < 0.01$). Thus, most of the purported difference between genders

Table 10.4 Ferraro *et al.* (2003) VCM experiment, mean contributions

	All	*Males only*	*Females only*
All-human treatment	2.14	1.77	2.69
Virtual-player treatment	1.16	0.84	2.08
Difference	0.98	0.93	0.61
% confusion contributions	54%	47%	78%

disappears when confusion contributions are removed ($0.93 for males versus $0.61 for females). What appears like a gender-effect is likely a gender-based difference in confusion for this specific sample.

Ferraro and Vossler

The other prior experiment we draw upon is from Ferraro and Vossler (2005), who apply the virtual-player method to the dynamic VCM game. They use an archetype multiple-round VCM game with group size of four, an MPCR of 0.5, and feedback on group contributions after each round. Subjects are undergraduate students from Georgia State University. The sample consists of 160 subjects: 80 in an all-human treatment and 80 in a virtual-player treatment.[5]

Figure 10.2 presents mean contributions (measured as a percentage of endowment) by round for the all-human and virtual-player treatments. The first observation is that confusion contributions are considerable. With

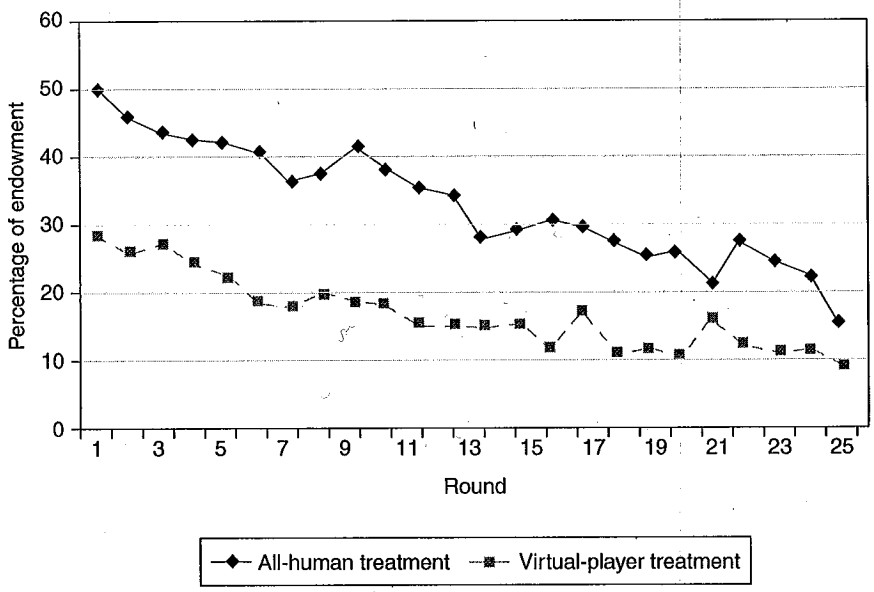

Figure 10.2 Ferraro and Vossler (2005) experiment, mean contributions.

participants contributing 32.5 percent and 16.8 percent of endowment in the all-human and virtual-player treatments, respectively, this suggests 52 percent of total contributions in the standard VCM game stem from confusion. Second, while the virtual-player treatment contributions do decrease over rounds, average contributions still amount to 10 percent of endowment in round 25.

Ferraro and Vossler carefully analyze the data using a dynamic pooled time-series model and find that the reduction in contributions in the virtual-player treatment is largely driven by the decline in observed contributions from virtual players in previous rounds. Thus, the standard decay in VCM experiments over rounds is not due to learning the dominant strategy or a reduction in confusion. Instead, similar to the correlation between MPCR and confusion contributions in our GHL application, confused participants in the virtual-player treatment simply use any available cue to help determine contributions. As additional validation of this result, Ferraro and Vossler report responses to a question similar to the one in our GHL experiment concerning what the purely self-interested, dominant strategy is. They find that 30 percent of respondents are unable to determine the dominant strategy of zero contributions, and additional evidence from the post-experiment questionnaire and focus groups suggests that this proportion is a lower bound.

Discussion

Through the course of three different applications of the virtual-player method, we find that at least half of contributions in public goods games stem from confusion. This finding in itself may not seem alarming, given that decision errors are likely rather commonplace in many economics experiments. Unfortunately, these confusion contributions do not simply amount to harmless statistical noise. In particular, we have shown that confusion contributions are sensitive to changes in design parameters, distort inferences about the role of other-regarding preferences, and confound comparisons between subpopulations. Furthermore, confusion just does not simply go away over the course of many repeated rounds. Overall, these results call into question the internal and external validity of this line of experimentation.

Do our results suggest we should just stop drawing inferences from public goods experiments? Certainly not, but they do suggest we need to rethink how these experiments are implemented. As a starting point for discussion, recall that Andreoni (1995) cites three potential causes of confusion contributions: (1) inadequate monetary rewards; (2) poorly prepared instructions; and (3) the inability of participants to decipher the dominant strategy.

Are inadequate monetary rewards a problem? We think not. Experiments discussed here involve payoffs that are on average much higher than student wages for this time commitment. Further, in an investigation of "house money" effects, Clark (2002) found that having subjects play the VCM game with their own money had no discernible effect on their behavior. The results of Clark are

consistent with the presence of a substantial number of individuals who are not clear about the appropriate strategy conditional on their preferences.

Are instructions "poorly prepared"? For the experiments discussed that use the virtual-player method, the instructions are standard in experimental economics. The decision settings are presented using neutral language, effort is made to avoid context, and subjects go through simple exercises to assess their understanding of payoff computations. From our experience, the vast majority is quite capable at performing the necessary payoff calculations. Thus, while our instructions – and instructions for public goods experiments in general – are not necessarily poorly prepared, the inability of individuals to decipher the dominant strategy does suggest the need for modifying how the game is explained.

Responses from post-experiment questionnaires we used, as well as behavior, suggest that at least 30 percent of respondents simply are not able to figure out the dominant strategy of zero contributions. Ferraro and Vossler report in a post-experiment focus group that just one-quarter of participants were able to figure out the dominant strategy by reading the instructions. This has important consequences for the external validity of the experiment unless one can show confusion has similar effects and magnitudes in "real world" contributions. We believe, however, that when faced with a naturally occurring contributions decision, people recognize the tension between privately beneficial free riding and socially beneficial contributions.

Our results thus call into question the standard, "context-free" instructions used in public goods games. Standard instructions for this type of experiment use neutral language and do not reveal that the experiment is about public goods or that participants are being asked to make a contributions-like decision. Indeed, the focus group of Ferraro and Vossler reveals that many participants thought they were playing some sort of assurance game.[6] We share the sentiment of Loewenstein (Loewenstein, 1999, p. F30), who suggests "Subjects may seem like zero intelligence agents when they are placed in the unfamiliar and abstract context of an experiment, even if they function quite adequately in familiar settings."

Our experimental evidence suggests that a bit of context could go a long way. In particular, since many subjects cannot figure out the dominant strategy (but all our GHL experiment participants figured out the social optimum) perhaps we can clue them in without altering their preferences for the public good. For instance, we could explain to participants that we are asking them for voluntary contributions for a public good and that the public good is simply an amount of money that gets distributed throughout the group. Subjects can be informed that it is perfectly reasonable to give nothing.

As a pilot study, one of the authors used such context-enhanced instructions in a standard, ten-round VCM experiment run in two sections of an undergraduate environmental economics course at the University of Tennessee in September 2005. These instructions are available from the authors on request. The experiment was being used to illustrate the free riding phenomenon (before the concept was formally introduced). After students read the instructions, but

before contribution decisions were made, the students were asked to write down the dominant strategy. Only three of 25 students (12 percent) failed to identify the dominant strategy of zero contributions (mean response was 0.4 tokens). This figure is considerably below those from comparable, context-free experiments: 30 percent from the *post*-experiment questionnaire in our GHL experiment, and the estimate from Vossler and Ferraro that three-quarters could not deduce the dominant strategy *prior* to the experiment.

The pattern of contributions is quite similar in both class sections: contributions start at about 50 percent and fall to 40 percent by round 10. This rate of decay is quite low for a VCM experiment, but results are consistent with expectations based on our virtual-player treatment results. Confused, herding individuals are going to follow the group trend and so any reduction in contributions in early rounds really causes a downward spiral: conditional cooperators get an exacerbated signal of free riding and revoke contributions, herders then further reduce, and so on. Without the herders, the decay in average contributions over time should be relatively less steep.

While there are likely tradeoffs associated with adding even generic context, namely that it could systematically alter participant preferences for the welfare of others, it appears that investigation into instruction based modifications is warranted. Consistent with our conjecture, the findings of Oxoby and Spraggon (this volume) suggests that confusion also may be reduced by providing a payoff table showing the subjects' payoffs given their decisions and the decisions of others. Note that the standard VCM instructions provide information only on the payoffs associated with each level of group contributions. The value of instruction enhancements could be tested using the virtual-player method, through survey questions with monetary rewards for correct answers, through debriefing sessions, and through external validity tests.

In conclusion, we believe that public good experiments will continue to play an important role in testing economic theory and designing public policies. However, they cannot achieve their full potential as long as they are implemented in a way that leaves many subjects oblivious to the social dilemma that experimentalists are trying to induce. Without innovation in the design of these experiments, our ability to draw inferences about behavior in collective action situations, and about the effects of alternative institutional arrangements that induce private contributions to the public goods, will continue to be impaired.

Notes

1 The two potential design flaws are: (1) human subjects in the computer condition observe their group members aggregate contribution *before* they make their decision in a round, as opposed to after they make their decision, as in the human condition; and (2) the automata contribute the *average* of what human condition members contributed. If the history of contributions affects both confused and other-regarding subjects, and if participants behave differently when the contributions of other players are known *ex ante*, then such changes in design affects the comparability of the two

conditions. Indeed, the intent of our virtual-player method is simply to have participants play with virtual players and not change any other aspect of the game.

2 A similar use of virtual players was employed by Johnson *et al.* (2002) in a sequential bargaining game.

3 In one session, a graduate student was asked to participate as a last-minute measure to make the total number of participants divisible by four. This individual was subsequently dropped from the data set. Due to the nature of the game, this inclusion of this person should have no impact on the contribution level of the undergraduate participants.

4 Davis and Holt (1993, p. 332) define a "minimal profitable coalition as "the smallest collection of participants for whom the return from contributions to the [public account] exceed the return from investing in the private [account]."

5 We only report their "*VI*" and "*HI*" treatments.

6 An Assurance Game (also known as the Stag Hunt) is a game in which there are two pure strategy equilibria and both players prefer one equilibrium (payoff dominant) to the other. The less desirable equilibrium, however, has a lower payoff variance over the other player's strategies and thus is less risky (it is risk dominant). In the case of the VCM game, some subjects erroneously view contributing their entire endowment as the most desirable strategy when everyone else in the group contributes their endowments too. Subjects described this decision as "risky" because it leads to low payoffs if other players do not contribute their endowments. Contributing zero was viewed as a payoff inferior choice but "less risky." These subjects were unable to infer the dominant strategy in the VCM game.

References

Andreoni, J., 1988. Why Free Ride? Strategies and Learning in Public Goods Experiments. *Journal of Public Economics*, 37 (3), 291–304.

Andreoni, J., 1990. Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *Economic Journal*, 100 (401), 464–477.

Andreoni, J., 1995. Cooperation in Public-goods Experiments: Kindness or Confusion? *American Economic Review*, 85 (4), 891–904.

Brown-Kruse, J. and D. Hummels, 1993. Gender Effects in Laboratory Public Goods Contributions: Do Individuals Put their Money Where their Mouth Is? *Journal of Economic Behavior and Organization*, 22 (3), 255–268.

Cadsby, C.B. and E. Maynes, 1998. Gender and Free Riding in a Threshold Public Goods Game: Experimental Evidence. *Journal of Economic Behavior and Organization*, 34 (4), 603–620.

Clark, J., 2002. House Money Effects in Public Good Experiments. *Experimental Economics*, 5 (3), 223–237.

Cox, J. and V. Sadiraj, 2005. Social Preferences and Voluntary Contributions to Public Goods. Paper presented at the Conference on Public Experimental Economics, Georgia State University.

Davis, D.D. and C.A. Holt, 1993. *Experimental Economics*. Princeton, NJ: Princeton University Press.

Eckel, C. and P. Grossman, 2005. Differences in the Economic Decisions of Men and Women: Experimental Evidence. In *Handbook of Experimental Results, Volume 1*, edited by C. Plott and V. Smith. New York: Elsevier.

Ferraro, P.J. and C.A. Vossler, 2005. The Dynamics of Other-regarding Behavior and Confusion: What's Really Going on in Voluntary Contributions Mechanism Experiments? Experimental Laboratory Working Paper Series #2005–001, Department of Economics, Andrew Young School of Policy Studies, Georgia State University.

Ferraro, P.J., D. Rondeau, and G.L. Poe, 2003. Detecting Other-regarding Behavior with Virtual Players. *Journal of Economic Behavior and Organization*, 51, 99–109.

Fischbacher, U. and S. Gächter, 2004. Heterogeneous Motivations and the Dynamics of Free Riding in Public Goods. Working Paper, Institute for Empirical Research in Economics, University of Zurich.

Fischbacher, U., S. Gächter, and E. Fehr, 2001. Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. *Economic Letters*, 71 (3), 397–404.

Frey, Bruno and Stephan Meier, 2004. Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment. *American Economic Review*, 94 (5), 1717–1722.

Goeree, J., C. Holt, and S. Laury, 2002. Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior. *Journal of Public Economics*, 83 (2), 255–276.

Houser, D. and R. Kurzban, 2002. Revisiting Kindness and Confusion in Public Goods Experiments. *American Economic Review*, 92 (4), 1062–1069.

Isaac, R.M., J. Walker, and S. Thomas, 1984. Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. *Public Choice*, 43 (2), 113–149.

Johnson, E.J., C. Camerer, S. Sen, and T. Rymon, 2002. Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining. *Journal of Economic Theory*, 104 (1), 16–47.

Laband, D.N. and R.O. Beil, 1999. Are Economists More Selfish than Other "Social" Scientists? *Public Choice*, 100 (1–2), 85–101.

Laury, S. and L. Taylor, Forthcoming. Altruism Spillovers: Does Laboratory Behavior Predict Altruism in the Field? *Journal of Economic Behavior and Organization*.

Loewenstein, G., 1999. Experimental Economics from the Vantage Point of Behavioural Economics. *Economic Journal*, 109 (453), F25–F34.

Marwell, G. and R.E. Ames, 1981. Economists Free Ride, Does Anyone Else? *Journal of Public Economics*, 15 (3), 295–310.

Nowell, C. and S. Tinkler, 1994. The Influence of Gender on the Provision of a Public Good. *Journal of Economic Behavior and Organization*, 25 (1), 25–36.

Oxoby, R.J. and J. Spraggon, 2007. The Effects of Recommended Play on Compliance with Ambient Pollution Instruments. In *Experimental Methods in Environmental Economics*, edited by T.L. Cherry, S. Kroll and J.F. Shogren (in this volume).

Palfrey, T.P. and J.E. Prisbrey, 1997. Anomalous Behavior in Public Goods Experiments: How Much and Why? *American Economic Review*, 87 (5), 829–846.

Yezer, A.M., R.S. Goldfarb, and P.J. Poppen, 1996. Does Studying Economics Discourage Cooperation? Watch What We Do, Not What We Say or How We Play. *Journal of Economic Perspectives*, 10 (1), 177–186.