Social Preferences in Experimental Economics

James Woodward†‡

This article explores some issues having to do with the use of experimental results from one-shot games to reach conclusions about the existence of social preferences that are taken to figure in the explanation of cooperation in repeated interactions in real life.

1. Introduction. In recent years, results from laboratory experiments in economics have been used to support the conclusion that human beings possess so-called social (non-self-interested or other-regarding) preferences and that such preferences play an important role in explaining cooperative behavior outside of the laboratory. This use of experiments raises a number of interesting methodological issues; in this article, my primary focus will be on just one of them—the use of results from *one-shot* or *nonrepeated* games to detect or "measure" social preferences that are claimed to be at work in real life cooperative activities outside the laboratory that involve *repeated* interactions.

My discussion is organized as follows. Section 2 briefly reviews a few of the relevant experimental results. Section 3 introduces some terminology and the distinction between using laboratory evidence to *test* the hypothesis that people are entirely self-interested and using laboratory evidence to *measure* social preferences that are also at work outside of the laboratory. Section 4 then discusses arguments for and against the use of one-shot laboratory games to measure social preferences

2. Some Experimental Results. In an ordinary *ultimatum game* (UG), a proposer (P) proposes a division of a monetary stake of $n to a responder, R.[1] P may allocate any amount $x up to $n for himself, with the remainder

1. For more detailed discussion of these and many other results, see Camerer (2003).

$n - x$ going to R. If R accepts this allocation, both players get the proposed division. If R rejects, both players get nothing. The identities of both P and R are unknown to one another. Suppose that the game is one-shot and that both players have entirely selfish preferences. Then R will prefer any positive offer to an offer of zero. If P knows this, P will offer R the smallest possible positive amount of money (e.g., one cent if the stake is divisible down to pennies) and R will accept: this is the subgame perfect equilibrium. This is not what is observed experimentally in any population. In most populations in developed countries, Ps offer an average of 0.3–0.4 of the total stake and offers under 0.2 are rejected half the time. The modal offer is 0.5.

In a *public goods* (PG) game, each of N players can contribute an amount c_i of their choosing from an initial endowment that is the same for each player. The total amount Σc_i contributed by all players is multiplied by some factor $m (m < 1/N)$ and divided equally among all of the players, regardless of how much they contribute. In other words, each player i's endowment is changed by $-c_i + m\Sigma c_i$. In this game, if players care only about their own monetary pay-offs, the dominant strategy is to contribute nothing—that is, to free ride on the contributions of the other players. In one-shot PG games in developed countries in which subjects do not know one another's identities, they contribute on average about half of their endowment, although again there is a great deal of variation, with a number of subjects contributing nothing and others contributing everything.

In repeated PG games, subjects begin with a substantial mean contribution, which then significantly declines under repetition. If a costly punishment option is introduced that allows subjects to punish noncontributors but at a cost to themselves, a number will do so, even in the final round, in which punishment cannot influence future behavior. Introduction of this option prevents the decline in contributions with repeated play.

In the experimental results described so far, subjects were drawn from "developed" countries (United States, Europe, etc.) and there was considerable uniformity of behavior. However, in a fascinating series of cross-cultural experiments (described in Henrich 2004) in which such games were played in a number of different small-scale societies in Asia, Africa, and South America, there was considerably more variation in results. For example, among the Machiguenga in Peru, one-shot UG offers had a mean of 0.26 and a mode of 0.15 (far lower than the mean and modal offers in developed societies) and almost no offers were rejected. In contrast, among the Lamerela in Indonesia, offers in excess of 0.50 (hyperfair offers) were common and there was frequent rejection, even of hyperfair offers. A similar variation was found in PG games and trust games.

Moreover, in at least some cases, these results appear to correlate with features of social life in the societies in question. For example, the Machiguenga are described as "socially disconnected" by anthropologists, with economic life centering on the individual family and little opportunity for anonymous transactions. By contrast, the Lamerala are a whaling society, in which there is a high degree of cooperative hunting and food sharing and in which it is common for people to reject gifts out of concern that this will place them under an obligation to reciprocate. The Orma in Kenya make relatively high contributions in PG games (both in comparison to many other small-scale societies and the contemporary United States: 0.58 of their endowments). Participants in this game associated it with a *harambee*, "a Swahili word for the institution of village level contributions for PG projects such as building a school" (Ensminger 2004, 376), and in fact a *harambee* collection was going on at the time the PG game was played. Ensminger suggests that the relatively high level of contribution was at least in part a reflection of the willingness of the participants to think of it as an instance of this practice.

3. Social Preferences? What do these results show about the existence of "social preferences"? Before exploring this question, some distinctions and regimentation of language will be helpful. I will assume that to talk to *preferences* of any kind, whether selfish or not, these must be both well behaved—in the sense of satisfying the usual axioms of revealed preference theory—and at least somewhat stable or robust in the sense of holding across some nontrivial range of contexts and changes in background conditions. I will say that subjects have entirely *self-interested preferences* if and only if (iff) they always choose in such a way as to maximize their own (expected) monetary payoffs. Subjects exhibit non–self-interested or *"prosocial" behavior* iff they do not always make choices that maximize their own expected monetary payoffs. Prosocial behavior thus includes both cases in which subjects make choices that increase monetary payoffs to others at the cost of a reduction of their own monetary payoffs and cases in which subjects are willing to reduce their own monetary payoffs in order to impose costs on others (cases in which subjects exhibit negative reciprocity). Subjects have *social preferences* iff they exhibit prosocial behavior *and* what is responsible for this behavior has the structure of a preference—that is, it is well behaved and at least somewhat stable.

In thinking about the results from the experiments described above, it also will be useful to distinguish two projects or sets of issues. The first (P1) has to do with using the experiments to *test* the hypothesis S that the subject's preferences are entirely self-interested. The second issue (P2) is this: If, as many investigators believe, S can be rejected, can we then use the behavior exhibited in these games to reliably *measure* or *estimate*

the subject's social preferences and then use this information to explain their behavior in other real life contexts? One way in which P1 and P2 differ is that an affirmative answer to P2 requires that the behavior and preferences present in the laboratory games exhibit some nontrivial degree of stability or robustness: it must be possible to detect or measure preferences in laboratory contexts and then generalize or export this information to other contexts, including real life field contexts under conditions that differ from those in the laboratory. It might be reasonable to reject S even if these assumptions about robustness are not true. That is, in principle at least, the subject's *behavior* in the laboratory or elsewhere might be prosocial or not fully explainable in terms of self-interested preferences and yet it might also be true that the additional influences on their behavior (whatever we decide to call these) are highly nonrobust and context dependent, in which case the measurement project (and perhaps even talk of "social preferences") might not be well motivated. In this connection, both fans and critics of the notion of social preferences have tended to assume that the notion of a social preference requires that there be at least some consistency of type at the individual level with, say, individuals who have a taste for negative reciprocity in some contexts (e.g. an ultimatum game of a certain design) also exhibiting this taste in other contexts as well—for example, in a willingness to punish noncontributors in PG games or in a willingness to reject low offers in a real life wage negotiation. I will follow this usage below.[2]

4. Social Preferences and Repeated Games. Why might one think that the experimental results described in Section 2 *fail* to show that subjects exhibit non-self-interested behavior and/or have social preferences? One relevant set of considerations derives from the theory of repeated games. Suppose that we are dealing with subjects whose preferences are entirely self-interested. Assume also that interactions are two–person or bilateral and that, in particular, two players find themselves in a game (like an ultimatum game [UG] or a prisoner's dilemma [PD]) that has the structure of a "social dilemma": there are gains for both players (in comparison to the baseline outcome in which each chooses a noncooperative strategy)

2. Although space precludes detailed discussion, let me note that there is little positive evidence for such consistency or stability and that there is at least some evidence against it. Experimentation shows that behavior (and presumably whatever preferences underlie such behavior) in many of the games described above is influenced in highly nonsystematic ways by a variety of different conditions—informational or anonymity conditions, framing, available alternatives, etc. See Camerer 2003 and Woodward, forthcoming, for additional discussion. The issue of consistency of subject's behavior/preferences across games and between games and real life behavior remains surprisingly unexplored.

that can be obtained from mutual choice of cooperative strategies but that a noncooperative choice is individually rational for each if they interact only once. Assume, however, that their interaction is repeated—in particular, that it has the structure of a repeated two-person game of indefinite length and that certain other conditions are met—the players care about their future sufficiently (do not discount too heavily) and have accurate information about the choices of the other player. Then, for many games meeting these conditions, the mutual choice of a cooperative strategy is a Nash equilibrium of the repeated game. To take what is perhaps the best known example, mutual choice of tit-for-tat is a Nash equilibrium of a repeated PD of indefinite length, assuming that the players have sufficiently high discount rates and accurate information about one another's play. What this shows (it is claimed) is that one can explain cooperative behavior in repeated games (to the extent that it occurs) simply by assuming that the players have self-interested preferences; at least in these contexts, the assumption that players have social preferences is unnecessary and hence unwarranted.

In my view there are a number of reasons for doubting that this analysis provides a full explanation for cooperative behavior even in repeated games, but in what follows I want to put this issue aside and assume, as many researchers do, that the analysis does show that cooperative behavior in repeated games can be accounted for just in terms of selfish preferences. In particular, I want to explore a line of argument, due to Ken Binmore (see, e.g., Binmore 2005, 2007) and Larry Samuelson (2005), among others, that uses this analysis of repeated games to criticize the use of evidence from one-shot games to measure social preferences.

Very roughly, the Binmore/Samuelson critique runs as follows. They begin with the claim that all or most real life social interaction is best modeled as a repeated game and that most subjects have very little experience with the sorts of anonymous one-shot games that are played in laboratory experiments. This is particularly likely to be true in the small groups or small-scale societies in which human beings lived for most of their history—experience of which presumably played an important role in shaping whatever dispositions to cooperate we presently possess. In such societies, and even in contemporary industrialized societies, most people find themselves repeatedly interacting with the same individuals in contexts in which questions of cooperation arise. Call this the *frequency of repetition* claim. Binmore and Samuelson suggest that, in consequence, when subjects play a one-shot game in the laboratory, one of two (not mutually exclusive) things is likely to happen. On the one hand, subjects may simply be confused, not really understanding the structure of the game being played and the real consequences of their choices. Alternatively, subjects may play the laboratory game by importing (perhaps con-

sciously, perhaps tacitly[3]) patterns of behavior, heuristics, or norms that derive from their experience in playing some repeated game in real life that strikes them as similar to the laboratory game, as the participants in Ensminger's experiments apparently did. Call this the *importation* thesis. (If the importation thesis seems far-fetched, note that it seems to provide a fairly accurate description of the relationship between many of the experiments reported in the Henrich et al. [2004] anthropological study and real life behavior.)

In either case, Binmore and Samuelson claim that problems result for the whole project of using one-shot games to measure social preferences. If subjects are confused, they may not have well-defined preferences (or be maximizing a well-defined utility function) at all. If subjects import patterns of behavior from a repeated game, then they are not playing the one-shot game that the experimentalist assumes that they are playing and inferences about their preferences based on that assumption will be invalid. As a simple illustration of this second possibility, suppose that subjects are presented with a one-shot game in which monetary payoffs to each player (not necessarily their utilities) have the structure of a PD. Suppose that the subjects cooperate. One explanation is that the subjects have social preferences that make cooperation an optimal strategy for each— they care about the payoffs to the other player as well as their own and act so as to maximize their expected utility. However, another possibility is that the subjects tacitly model the one-shot PD as a repeated game and import strategies like tit-for-tat that are equilibrium strategies for subjects with selfish preferences in the repeated game into the one-shot laboratory game. If subjects do this, we apparently can explain their behavior without any appeal to other-regarding social preferences. A similar observation applies to other one-shot games that are often taken to show the existence of social preferences.

Defenders of the use of one-shot games to argue for the existence of social preferences, such as Camerer and Fehr (2004) and Gintis (2006), have responded to this critique by criticizing both the frequency thesis and the importation thesis. In connection with the former, it is suggested (I believe plausibly) that, at least in societies like ours and perhaps more generally, real life one-shot interactions are actually fairly common and

---

3. Experimenters go to considerable lengths to make sure that subjects understand (at a conscious level) the rules of the game that they are playing, including whether it is one-shot or repeated. What Binmore and Samuelson are suggesting is that, despite their intellectual recognition that the game is one-shot, subjects may nonetheless import (perhaps as a result of unconscious processing or implicit learning) patterns of play from repeated games with which they have experience.

frequently involve cooperation.[4] For example, situations in which people face decisions about whether to tip in restaurants to which they will not return seem fairly common in societies like ours. Moreover, many people seem to act in a way that does not maximize their material pay-offs in such situations. If so, such behavior cannot be explained in terms of the importation of patterns of play that are sustained by self-interested preferences in repeated games. However, as we shall see, despite the rhetorical role played by the frequency claim in Binmore's and Samuelson's discussion, many of their methodological worries about the use of one-shot games to explain real life cooperation seem largely independent of whether the frequency claim is correct.

A second line of response to the Binmore-Samuelson argument criticizes the importation thesis, or at least the strong form of it that seems to be required by their argument. According to this strong form of the thesis, subjects always and automatically transfer strategies appropriate for some repeated game they have experienced in "real life" into one-shot games in a behaviorally inflexible way—that is, they always play one-shot games as though they are such repeated games or fail to distinguish between one-shot and repeated versions of the same game. Defenders of the use of one-shot games to measure social preferences contend that this version of the importation thesis is inconsistent with a substantial amount of behavioral evidence suggesting that even naive subjects play differently in repeated and one-shot games of the same type and that they modify their behavior in the latter depending on the play of their opponents and on considerations having to do with reputation formation. For example, Fehr and Fischbacher (2003; also discussed in Camerer and Fehr 2004) conducted a series of UGs in two different conditions, each involving subjects playing a different opponent in each of 10 different iterations.

---

4. At the risk of belaboring the obvious, it is worth observing that there is an ambiguity in the notion of a one-shot or nonrepeated game that has been a potential source of confusion in this discussion. On the one hand, this may be a game that a subject plays only once (ever, with any one). Such games may well be uncommon in our or any other society, and since subjects have no prior experience with them, it is not implausible that they often may play in a "confused" way—certainly play will not be governed by preexisting norms or patterns of behavior with which the subjects are familiar. On the other hand, a "one-shot" game may be a game that is played repeatedly by a subject but always with different people—as when the same person makes a series of decisions on different evenings about whether to tip in different restaurants or when a merchant makes a series of decisions about whether to cheat different customers. It seems to me that, in societies like ours, people often will have repeated experience with interactions that are one-shot in this second sense, and, of course, there often will be norms, expectations, or repeated patterns governing such interactions—e.g., norms about tipping. To the extent that this is so, the claim that behavior in such interactions reflects confusion or unfamiliarity will be less plausible.

In one condition, proposers knew nothing about the past play of responders so that responders had no opportunity to build a reputation by rejecting low offers. In the reputation condition, proposers knew about the past play of responders. It was found that the great majority of responders increase their threshold for the lowest offers they will accept in the reputation condition as opposed to the no-reputation condition. This suggests that they understood the value of reputation formation and, along with it, the contrast between one-shot and repeated play (Camerer and Fehr 2004). There are many other experiments pointing to a similar conclusion. For example, subjects play repeated PDs with the same opponent differently from one-shot PDs—they cooperate more when they play against the same opponent, and they play repeated one-shot PDs (i.e., PDs played against a different opponent each time) differently than PDs that they play only once—cooperation declines over time in repeated one-shot play. Again, this seems to suggest that subjects are sensitive to the strategic and reputational considerations that distinguish repeated from one-shot play and are influenced by learning from previous play.

Although this response to the importation thesis has been advanced in *support* of the claim that one-shot games can be used to detect or measure social preferences, it seems to me that, to the extent that the response is correct, it undermines or at least casts doubt on that claim. The essential problem can be put in the form of a dilemma. Suppose that our interest is in using a one-shot laboratory game in which subjects cooperate (e.g., a one-shot PG game) to understand why subjects also cooperate in a real life repeated interaction, such as the *harambee* institution. Logically, there seem to be just two possibilities. One is that, as Binmore and Samuelson suggest, subjects import patterns of behavior that are appropriate in the real life repeated game into the lab game. Then, to the extent that purely selfish preferences are sufficient to explain cooperation in the real life repeated game, it would appear, for reasons we have discussed, that no inference to the existence of social (nonselfish) preferences on the basis of performance in the laboratory game is warranted. That is, if subjects play one-shot games by importing, as Binmore and Samuelson suggest, this observation plus the assumption of selfish preferences seems sufficient to explain their one shot laboratory play.

The other possibility is that the subjects do *not* import strategies or patterns of behavior from the target repeated real life game into the one-shot laboratory game. Assume, for the sake of argument, that this shows that purely selfish preferences are not sufficient to explain their cooperative behavior in the laboratory and that "something else" is involved. As we have seen, it does not follow that this "something else" necessarily has the kind of consistency and stability that warrants talk of social preferences, but let us put this consideration aside and agree that social pref-

erences of some form figure in the explanation of the laboratory behavior. The following question then arises: What is the relationship between these social preferences that are present in the laboratory and whatever preferences are operative in the real life repeated game? Can (hypothesis 1) the same social preferences that are measured in the laboratory also be the preferences that explain cooperative behavior in the real life repeated game, even though it is also true (hypothesis 2) that it is unwarranted to think that subjects import norms, strategies, and patterns of behavior from the real life repeated game into the lab? It is not obvious that this combination of hypotheses is plausible or even coherent. Prima facie, one would think that if the laboratory experiment successfully measures or detects preferences that are at work in explaining cooperation in real life repeated interaction, this can only be because subjects in the laboratory do import those preferences into the laboratory game, where it also guides their behavior. That is, the cost of accepting hypothesis 2 seems to be that the external validity of the inference from behavior in the laboratory and the preferences detected on the basis of that behavior is undermined.

Put somewhat differently, the problem is that the proponent of using one-shot games to measure and detect social preferences seems to have two different methodological goals that are difficult to satisfy simultaneously. On the one hand, the proponent wants to use cooperative behavior in one-shot games to show that subjects have nonselfish preferences; on the assumption that play in repeated games can be explained just by appeal to selfish preferences, this requires that play in one-shot games not simply reflect what happens in some associated repeated game. On the other hand, the object of the laboratory investigation is to uncover preferences that are at work in explaining behavior in real life repeated interactions, and this seems to require some nontrivial similarity between the two contexts.

Another way of putting the dilemma is in terms of the robustness of the behavior and preferences that are exhibited in the one-shot laboratory game. A crucial question to ask is whether these are robust under repetition of the game. Evidence of the sort provided by Fehr and Fischbacher seems to suggest that, at least in some cases, behavior exhibited in one-shot games is not robust under repetition and that subjects are influenced by considerations having to do with reputation and learning in repeated games that are not at work in one-shot games. But to the extent that this is true, behavior in the one-shot laboratory game (and whatever we are able to infer about the preferences or motivations that sustain that behavior) seems likely to tell us less than we would like about behavior in the repeated game or the preferences that underlie such behavior. However, if the behavior detected in the one-shot game is robust under repetition, and if it is also true that the behavior found in the repeated game

can be explained just in terms of selfish preferences,then it is not clear how results from the one-shot game by themselves could be evidence for nonselfish preferences.

I said above that the cogency of the Binmore-Samuelson critique is largely independent of the truth of the frequency claim and that the critique stands even if that claim is rejected. We now should be able to see why this is true. The dilemma described above will be applicable whenever a researcher claims to use a one-shot laboratory experiment to detect social preferences that are also claimed to explain cooperation in a real life repeated game. Whether or not one-shot cooperative interactions are common either in our society or others, it is uncontroversial that many real life interactions in which cooperation occurs are repeated interactions. It would be an enormous and unsatisfactory restriction on the scope of a theory of social preferences if it did not apply to real life cases of cooperation involving repeated interactions. For example, the social preferences (or whatever) that are detected in the laboratory PG game should also be such that they or similar preferences figure in real life patterns of repeated cooperation such as the *harambee*. Moreover, issues about robustness remain even if we restrict attention to inferences from one-shot laboratory games to real life interactions that are one-shot in the sense of not being repeated with the same partner. As we have noted, a fairly generic result from laboratory experimentation is that prosocial behavior declines under repetition of one-shot games. To the extent that many one-shot interactions are repeated with different partners in real life (the same person tips one time in each of many different restaurants on a trip), why does not prosocial behavior in such interactions erode under repetition? It looks as though whatever it is that is responsible for patterns of continued cooperation in real life one-shot interactions (never mind cooperation in repeated interactions) cannot be determined just from laboratory games in which play occurs only once.

I emphasize that the argument that I have given does *not* show either that people do not have social preferences or that they do not exhibit genuinely prosocial behavior. Nor would it be correct to conclude that behavior in one-shot games does not provide information of any kind about behavior in real life repeated interactions. Rather, what my argument (if successful) shows is twofold.

First, it suggests that a certain class of arguments for the existence of social preferences that are operative in real life repeated interactions is problematic, namely, arguments based on behavior in one-shot laboratory games that are thought to be analogues or instances of the repeated real life interaction. This is entirely consistent with there being other more convincing evidence for the existence of some form of social preferences or at least prosocial behavior. In fact, I think that there is such evidence—

it is based in part on experimental results from *repeated* games, such as restart and endogenous grouping effects (in which allowing cooperators to associate together boosts contributions) in repeated PG games. These avoid the problematic inference from behavior in one-shot games to behavior in repeated games that I have been criticizing .

As I read this literature, it supports a picture according to which typical subject populations are mixtures of different types, some substantial portion of whom are purely selfish and another substantial portion of whom are conditional cooperators, who will cooperate with others who have cooperated (or whom they expect will cooperate) even when this is not the pay-off maximizing choice. Perhaps unsurprisingly, given that they are conditional cooperators, the behavior of this group appears to be very sensitive to their beliefs/expectations about how others with whom they are interacting will (or have behave[d]), to their beliefs about others' beliefs and expectations regarding their behavior, to the intentions they believe others to have, to the costs of cooperation, to perceived entitlements, and to other factors as well. This context dependence perhaps explains why it has proved so hard to write down simple utility functions characterizing subject's social preferences that allow for precise quantitative predictions of behavior across contexts, including from the laboratory to the field. It also fits with the observation that social norms play a large role in coordinating cooperative activity and that experiments of the sort discussed above often seem to be more plausibly construed as measurements of such norms rather than individual-level social preferences. (cf. Bicchieri 2006).

Second, the argument I have given certainly does not undercut the reality of the differences in patterns of play in UG, PG, and other games across different cultures reported in Henrich et al. 2006. It remains a sensible project to ask whether those differences correlate with differences in cooperative behavior in real life interactions in those societies—for example, are societies whose members behave more cooperatively in PG games also more likely to contain successful institutions that provide public goods? However, if the argument I have given is correct, finding such a correlation does not by itself tell us *why* such subjects cooperate, either in the lab or in real life. That is, for all that the existence of the correlation shows, both instances of cooperation may be sustained by purely selfish preferences or they may be sustained in some other way.

REFERENCES

Bicchieri, Cristina (2006), *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.
Binmore, Ken (2005), "Economic Man—or Straw Man: A Commentary on Henrich et al.", *Behavioral and Brain Sciences* 28: 23–24.

——— (2007), *Does Game Theory Work? The Bargaining Challenge*. Cambridge, MA: MIT Press.

Camerer, Colin (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Camerer, Colin, and Ernst Fehr (2004), "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists", in Henrich et al. 2004, 55–95.

Ensminger, Jean (2004), "Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa", in Henrich et al. 2004, 356–382.

Fehr, Ernst, and Urs Fischbacher (2003), "The Nature of Human Altruism", *Nature*, October 23, 785–791.

Gintis, Herbert (2006), "Behavioral Ethics Meets Natural Justice", *Politics, Philosophy and Economics* 5: 5–32

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, eds. (2004), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Oxford: Oxford University Press.

Samuelson, Larry (2005), "Foundations of Human Sociality: A Review Essay", *Journal of Economic Literature* 43: 488–497.

Woodward, James (forthcoming), "Experimental Investigations of Social Preferences", in Harold Kincaid and Don Ross (eds.), *The Oxford Handbook of Philosophy of Economics.* Oxford: Oxford University Press.