Field Experiments in Economics: Some Methodological Caveats

by

ANDREAS ORTMANN

Center for Economic Research and Graduate Education
Charles University (CERGE)
and
Economics Institute (EI)
Academy of Sciences of the Czech Republic
Prague, Czech Republic

July 31, 2003

CERGE-EI

P.O.BOX 882, Politickych veznu 7, 111 21 Prague, Czech Republic
tel.: (420-2) 242 30 117, fax: (420-2) 242 11 374, 242 27 143
e-mail:andreas.ortmann@cerge-ei.cz
ortmann@mpib-berlin.mpg.de
aortmann@yahoo.com

DRAFT; DON'T QUOTE: DRAFT; DON'T QUOTE: DRAFT; DON'T QUOTE:

* Thanks to the organizers and participants of the Middlebury Conference on Field Experiments in Economics (April 26-27, 2003). Special thanks to Glenn Harrison, Ralph Hertwig, and Angelika Weber for comments on earlier versions of this manuscript. The usual caveats apply.

Abstract

The results of standard lab experiments have long been questioned because of the convenience sample of subjects – college students – they typically employ and the abstract nature of the typical lab setting. These procedural conventions of experimental economics, it is argued, endanger the external validity of experiments. Various forms of field experiments (see Harrison and List 2003a for a taxonomy) have tried to address these issues by bringing the lab to non-traditional subjects (including participants in exotic locales), and/or to move the setting of experiments closer to reality by using real goods and settings that are not stripped of context. While field experiments might help experimental economists to broaden their subject pools, and to increase the external validity of their investigations, I argue that these potential advantages do come at costs that can be considerable. Specifically, going into the field is likely to increase the demands on careful controls dramatically, and it also confronts experimenters with the tricky issue of how to deal with the rules of thumb, or heuristics, that subjects bring to the laboratory, whether the traditional one or the nontraditional one. The recent surge in field experiments has prompted, quite rightly in my opinion, a surge of ruminations about experimental methods both in the field and in the lab. This is a healthy development because methodological discussions have been marginalized in experimental economics far too long, to the detriment of overdue reflections on the practices of experimental economists (Hertwig & Ortmann, 2001; Ortmann, Hertwig, & Harrison, in prep). When the dust has settled, the most important contribution of field experiments might well be their having opened the door for methodological discussions that experimental economists have

avoided for too long.

1. Introduction

> "Methodological discussion, like spinach and calisthenics, is good for us ... " (Paul Samuelson, 1963, p. 231)

The standard methodology of experimental economists is to "decontextualize" the laboratory environment by attaching abstract labels to participants, the set of actions that one can chose from, or the goods that they are trading. In gift exchange or principal-agent experiments, for example, "A" or "B" sell and buy some unspecified commodity whose value has been "induced" (Smith 1976). Traditionally it was argued that such decontextualization enhances experimental control.

Evidence – to be discussed below – has been mounting over the past decade that has put into question this aspect of the experimental paradigm in economics. Experimental economists have also been questioned about their use of a convenience sample of subjects – college students – which some consider "this very weird, and very small, slice of humanity" (Henrich 2001, p. 414).

Various forms of field experiments (see Harrison & List, 2003a, for a taxonomy) have tried to address these issues by bringing the lab to non-traditional subjects, and/or to move the setting of experiments closer to reality by using real goods and settings that are not stripped of context. The arguments in favor of field experiments seem persuasive on at least two grounds: First, by bringing "the lab" to subjects, experimental economists address the doubt that college students tell us all there is about human reasoning and

behavior. By drawing on a broader subject pool – possibly more widely distributed geographically and culturally --, we may be able to analyze a broader spectrum of human behavior and therefore better understand human reasoning (and individual differences) along various demographic dimensions (e.g., age, gender, cognitive ability, cultural influences, etc.) that experimental economists probably should have controlled for routinely all along but for reasons that are hard to pinpoint did not. Second, by implementing field experiments in "laboratory" environments that are not (completely) stripped of context (e.g., the auction experiments by List & Lucking-Reilly 2000, List 2001, and Harrison & List 2003),[1] i.e., by running experiments in "naturally occurring environments" with real goods rather than the abstract goods (and induced valuations) experimental economists typically use, we allow subjects to draw on the myriad rules of thumb, or heuristics, that subjects acquire in their daily lives and that enable them, arguably, to navigate their environments on average just fine (e.g., Cosmides & Tooby 1996). Applied to experiments this view implies (e.g., Ortmann & Gigerenzer 1997) that experimental economists' almost universal practice of presenting their tasks stripped of context is likely to impede rather than enhance experimental control. Field experiments, and lab experiments, that allow subjects access to "field referents" (as Harrison & Rutstroem, 2001, have called the rules of thumb, or heuristics, that subjects bring to the laboratory) may well give us more control than their standard context-free counterparts that currently are the norm. Tapping field referents may therefore give us more external validity.

[1] Interestingly, even in those cases where standard abstract toy games are being used, it seems that non-traditional subjects sometimes are better at understanding the structure of the underlying social dilemma (e.g., see Henrich et al 2001, p. 76, where the authors report that Orma experimental subjects recognized the public good game for what it is and identified it by its local name; see also Henrich et al 2002, p. 16).

The potential advantages of field experiments, however, do often exact a price. While I appreciate the possibilities that field experiments offer (and while I am sympathetic to the difficulties of doing field work in remote locations where implementation of even simple experiments can become a challenge), it is important to keep in mind that field experiments might increase the variability of experimental practices (which may confound the variation between groups with distinct sets of demographic characteristics). It is also important to keep in mind that the rules of thumb, or heuristics, or field referents which participants bring to field, or for that matter lab, experiments may increase or decrease experimental control. In other words, "going into the wild" poses interesting new methodological challenges and trade-offs that currently we do not well understand.

2. A famous field experiment revisited

Henrich (2000) was one of the first field experiments that caught the attention of the economics profession, for three reasons. First, it was published in the *American Economic Review*, giving it the kind of authentication that warrants attention. Second, it produced results – the "Machiguenga outlier" (Henrich et al 2002, p. 2) – that contradicted conventional wisdom, here seemingly well-established results from numerous ultimatum game experiments. Third, it triggered a major project involving a dozen researchers doing research in 12 countries, and an even larger number of small-scale societies, all over the world. The major findings of this project are claimed to be a refutation of the "canonical model" in each of the studied societies, "considerably more behavioral variability across groups than had been found in previous experiments", evidence that group-level differences in economic organization and the degree of market integration matter while

individual-level economic and demographic variables do not, and the resultant claim that "behavior in the experiments is generally consistent with economic patterns of everyday life in these societies." (Henrich et al. 2001, pp. 73/74; see also ; Henrich et al. 2002, p. 2) In other words, culture is claimed to affect economic behavior (as measured by giving behavior in ultimatum, dictator, and public good provision experiments).

The major experiment reported in (Henrich 2000) attempted to replicate with a non-traditional subject pool – the Machiguenga of the Peruvian Amazon – numerous previous studies of the ultimatum game. While previous studies (e.g., Gueth, Schmittberger, Schwarze 1982; Roth 1995) had demonstrated systematic deviations from the canonical game theoretic prediction[2], the mean and mode of the offers of the high-stakes gambles were dramatically reduced (0.26 and 0.15 vs. 0.48 and 0.5 in the control group of UCLA graduate students in anthropology which roughly replicates results with students from other disciplines as well as experiments with lower stakes).[3] These rather unfair offers

[2] I use the term "canonical game theory" to signify the deductive game theory that one finds in standard text books like Kreps (1990) and Mas-Colell et al. (1995) and that is the preferred strawman of the behavioral economics and finance crowd. Almost a decade after the path-breaking quantal response modeling papers of McKelvey and Palfrey (1995, 1998) and Goeree and Holt (2000, 2001) started circulating, one would expect that a more challenging target would be defined by this crowd. Alas, unfortunately that has not been the case. The proponents of the quantal response approach acknowledge that people make mistakes (increasingly so as making mistakes becomes less costly) and that therefore choices are probabilistic. In other words, choices are determined not just by the signs of payoff differences but also by the magnitudes of gains and/or losses. This approach has been able to rationalize a broad spectrum of experimental results even for those notoriously difficult ones resulting from corner point equilibria.

[3] Anthropology students, as Henrich points out (footnote 2 of Henrich 2000a), are likely to be an even weirder and smaller slice of humanity than students *per se*, so we should expect some variation in the results from what we see typically.

notwithstanding, the rejection rate was extraordinarily low among the Machiguenga (less than 5 percent).

These intriguing results prompt a number of interesting questions such as whether these results are generated by the atomistic ways in which the Machiguenga live, and to what extent these ways are responsible for the rules of thumb, or heuristics, that seem reflected in the behavior of those subjects.

Every test of such conjectures, however, is always a joint test of the theory and the way the test is implemented; this is the essence of the Duhem-Quine thesis (Smith 2002). Let us therefore take a quick look at the procedural aspects of this particular experiment and ponder the author's claim that "such things as procedural differences seem unlikely to explain the substantial differences observed between the Machiguenga and the typical robust results" (Henrich 2000, 975; for similar statements see Henrich et al 2001, p. 77 and Henrich et al 2002., pp. 18 - 20). Says he:

> First, I gathered 12 men together between 18 and 30 under the auspices of "playing a fun game with money." I explained the game to the group in Spanish using a set script written in simple terminology like "first person" to refer to the proposer and "second person" for the responder (Spanish is a second language for the Machiguenga). After this I had a bilingual school teacher (an educated Machiguenga) re-explain the game in the Machiguenga language (translating from my script), and display the money that we would be using to make the payments. After this, each participant entered my house (the guest hut) individually. We explained the game a third time, and I asked a number of hypothetical, practice questions intended to test the participants' comprehension of the game. We reexplained parts of the game as necessary. Often numerous examples were necessary to make the game fully understood. ... The following day, after having successfully gotten 12 responses and paid out some money, I began seeking randomly selected individuals to play the game. Most people had already heard of the game and were eager to play. I privately explained the game to each individual (usually in his or her house) and ran through the same testing procedure as the previous day. During this process several people were rejected because they, after 30+ minutes of explanation, could not

> understand the game - at least they could not answer the hypothetical questions.
>
> The initial 12 players were volunteers, but the next 30 players were selected at random from my demographic survey. ... Machiguenga players were told that their anonymous partner was another member of the community (Camisea), but nothing more was said about how this indvidual would be chosen, their age, sex or family.
>
> Demographically, Camisea contains 260 people from 36 households, with about 70 adults, These 36 households can be roughly divided into 12 extended families. The player pool contains 14 females and 28 males. ... " (Henrich 2000, 975; emphasis added by present author)

3. Experiments as joint tests: the issue of procedural regularity

I doubt that there is an experimental economist who would not, compared to standard implementations of ultimatum games, be concerned about the procedural differences reflected in the description of the experiment reported in Henrich (2000). There is, for example, no telling what the announcement of a "fun game" did to subjects' perception (frame) of what this experiment was about. We do know that even seemingly minute changes in instructions can have statistically significant and dramatic effects on outcomes (e.g., Hoffman, McCabe, Smith 2000, or Burnham, McCabe, Smith 2000)[4]. Second, I am not aware of instances where an ultimatum game experiment was conducted one-on-one (or, better, one-on-two), allowing subjects to communicate their experiences with the experimental situation to subjects that follow them. Third, it is highly unusual that

[4] Hoffman et al. report the results of exchange ultimatum game experiments. They compare two treatments – an impersonal exchange situation and a personal exchange situation -- whose instructions differ by two short sentences that remind sellers of the strategic nature of their interaction with buyers and the possibility of rejection; these treatments are significantly different. Burnham et al. report the dramatic effects of changing in the instructions for an extensive form bargaining game the word "opponent" to "partner".

instructions for such a simple game had to be repeated (twice). Fourth, it is highly unusual that even then numerous examples – apparently unscripted – had to be used to communicate to subjects the task at hand. Fifth, it is noteworthy indeed that even after 30+ minutes of explanation “several people” did not understand the game and had to be dismissed from the experiment. Sixth, it is important to know how the dismissals were explained to other potential participants and how these dismissals were perceived by other potential participants. Seventh, it would be interesting to know what the subjects knew about their selection (recall that the first 12 subjects were volunteers, the others were selected based on criteria that are not explicated; gender was definitely not it, as the gender composition of the sample was imbalanced). Eighth, it is imperative to know to what extent the participants conceptualized this game as part of some meta-game. The experiment was conducted by a researcher who apparently had visited that community before, and was very likely to do so again, making the game experiment into a stage game of an indefinitely repeated game between subjects and experimenter(s).

The above set of observations and questions strongly suggest that the procedural differences were substantial. Granted, certain procedural differences (e.g., having to repeat the instructions) may have been unavoidable given the exotic nature of the participants. The issue is not so much to what extent these procedural differences were necessary but to what extent they undermined the procedural regularity that is the hallmark of experimental control (Hertwig & Ortmann 2001). Specifically, the reduced social distance, the twice repeated instructions, and the numerous examples all are very likely to induce demand effects - a well documented phenomenon in experimental research that

has been illustrated in a variety of settings (e.g., Pfungst 1911; see also Rosenthal & Fode 1963; Rosenthal & Rubin 1976; Rosenthal & Rosnow 1991).

Pfungst's investigation of "Clever Hans (The Horse of Mr. Von Osten)" is arguably the most famous example of systematic research on the effects of subtle and unintentional cueing on the part of a questioner, an issue that has to be a major concern in the context of the experiment under consideration. Clever Hans, as you may recall, astounded scientists, and public audiences, with his seeming abilities to solve, for example, mathematical problems involving addition, subtraction, and even more complicated operations. For example, when asked for the square root of 16, the horse responded with four taps of a hoof. Hans, it semed, was also able to read German (by pointing to cards with his nose), among many other tricks. Fraud, of course, was suspected and an investigative panel, consisting of influential philosopher and psychologist Carl Stumpf (who headed the commission) and other experts, was formed in 1904 to get to the bottom of the matter. It didn't.

The ultimately successful detective work was undertaken by Stumpf's assistant Oskar Pfungst who designed and implemented an exemplary and ingenious set of experimental controls. For starters he controlled whether the questioner and the questioner's knowledge of the answer affected the horse's performance. The questioner didn't, the knowledge did. Pfungst then tried to understand whether the subtle cues that questioners apparently gave the horse were visual or vocal. Fitting Hans with blinders, and having Von Osten stand by its flanks, the horse was unable to replicate its remarkable performance. This clearly suggested that visual, rather than vocal, cues were the culprit. Pfungst went on to try to understand the nature of those cues which turned out to be subtle

indeed: For example, slight lowering and raising of the head triggered tapping and stopped it, respectively. In fact, as Pfungst himself demonstrated by taking Von Osten's place, it was not even necessary to trigger tapping with a question.

Pfungst even managed to produce similar Clever Hans effects with human beings in the laboratory: "Taking the role of Clever Hans himself, Pfungst invited subjects into the laboratory, connected them to apparatus measuring both head movement and respiration, and instructed them to ask questions. Like Clever Hans, Pfungst responded by tapping. The results were overwhelming: over 90 % of the subjects tested provided Pfungst with unintentional cues for the cessation of response, cues of which they were wholly unaware." (Wozniak, 1999)

Did the reduced social distance, the twice repeated instructions, the numerous ad hoc (?) examples induce a Clever Machiguenga effect? We don't know but surely there was plenty of room for what psychologists call experimenter expectancy effects (Rosenthal & Rosnow 1991, pp. 119 - 125, 128 - 133; see also pp. 111 - 112 for "Clever Hans"; see also Rosenthal & Fode 1963 on expectancy effects for albino rats and Rosenthal & Rubin 1976 for a summary of the first 345 studies of interpersonal expectancy effects.[5])

These demand effects might have been reinforced by the dismissal of participants who did not get the control questions right - a fact that surely spread like wildfire through the community (as anyone who has ever done experiments in as large a microcosm as a typical liberal arts college will be able to confirm) and quite possibly suggested to those selected that there was something like a normative solution that had to be found. Also,

[5] For a highly readable brief summary of Rosenthal's oeuvre on expectancy effects, see http://www.psichi.org/pubs/articles/article_121.asp

those that got selected as proposers might – especially in light of the dismissals of other potential candidates – have interpreted their role as resulting from somehow earned entitlements (e.g., Hoffman & Spitzer 1985)..

If indeed participants conceptualized this game as part of some meta-game then a whole bag of new questions (and necessary controls) is opened: What was subjects' experience with the experimenter? How often did they interact with him? Did he use deception in earlier experiments? (Ortmann & Hertwig 2002)

Henrich and his collaborators are aware of the possibility of experimenter biases. Similar to Henrich (2000, 975), Henrich et al explicitly acknowledge that "some of the variability among groups may be due to variations in implementation." (Henrich et al 2001, 77; see also Henrich et al 2002, pp. 18 - 20). They argue, however, that "the experiments were run from identical protocols across groups and were thus as similar in procedures and stake size as we could achieve." (Henrich et al 2001, p. 77) They also point at the control sessions that Henrich did with UCLA graduate students in anthropology. While these sessions controlled indeed for stake size and experimenter, they did not control for most of the other procedural idiosyncracies enumerated above (e.g., whether the experiment was conducted one-on-one or in a group setting, etc.), creating exactly the kind of procedural irregularity that undermines the hallmark of good experimental work: replicability (Hertwig & Ortmann 2001).

Interestingly, the authors tested for a systematic relationship "between the time each experimenter had spent in the field prior to administering the games and the mean UG of each group" (Henrich et al 2002, p. 19), without finding a consistent pattern. But if one controls for such a vague relationship, should one not also control for subjects' earlier

experience with the experimenter including important events such as the use of deception in earlier experiments? Or, the number of examples it took to get the subject to understand the task at hand? Or, what the subject had heard about the experiment from those that got dismissed because they answered questions incorrectly?

None of the concerns voiced above may ultimately matter. (And in fact, I am persuaded by the theoretical argument that culture matters although I draw very different – modelling – implications from it; see below.) The point is that we don't know with a reasonable degree of confidence how the procedural differences affected the results. What we do know is that the procedural differences were substantial, and that what we know about framing, demand effects, and meta-games cast doubts on the authors' claim that the substantial procedural differences did not matter. We simply would have more confidence in these results if they were not confounded by substantial procedural differences (which are very different from the procedural variations such as stake size and degree of anonymity) that experimental economists have discussed and that the author mentions (Henrich 2000, p. 974).)

The issue is to what extent Henrich gave up, or had to give up, in exchange for accessing a very different subject pool, the kind of experimental control that the experimental lab (and our use of that weird slice of humanity that we often use as subjects) guarantees us. I submit that the trade-off was substantial and that therefore the value of these results is limited at best. Similar caveats apply to the larger study (Henrich et al. 2001, 2002) that reports ultimatum game results from 15 small-scale societies around the world. I submit that there is a reasonable possibility that one key result of the larger study – that group-level differences in economic organization and the degree of market integration

matter while individual-level economic and demographic variables do not (Henrich et al. 2001, 2002) – is the result of experimenter biases. There is simply no persuasive evidence that the likely differences in implementation did not matter. In other words, we do not know whether what produced the effects was culture or procedural differences.

This then is the interesting, and bigger, issue: By tapping other subject pools, we may be forced to give up the kind of procedural regularity that has contributed to our confidence in our results (Hertwig & Ortmann 2001). This issue is, of course, not specific to the studies that I have discussed above in some detail but of a very general nature.

4. Simple heuristics make us smart, or dumb: The issue of field referents

The standard methodology of experimental economists is to "decontextualize" the laboratory environment by attaching abstract labels to participants, the set of actions that one can chose from, or the goods that they are trading. In gift exchange or principal-agent experiments, for example, "A" or "B" sell and buy some unspecified commodity whose value has been "induced' (Smith 1976). Traditionally it was argued that such decontextualization enhances experimental control.

Evidence from psychology and increasingly also from economics strongly suggests that things are not quite so easy (Ortmann & Gigerenzer 1997; Gigerenzer. Todd, & the ABC Research Group 1999; Todd & Gigerenzer 2000; Harrison & List 2003a) . There is convincing evidence from psychology, for example, that the decontextualization of laboratory situations has a dramatic and negative effect on people's ability to perform simple logic reasoning tasks.

Ortmann & Gigerenzer (1997) summarize the evidence related to the Wason Selection Task, “the most intensively researched single problem in the history of the psychology of reasoning” (Evans, Newstead and Byrne 1993, p. 99). This task presents subjects with a simple conditional statement such as “If there is an ‘A’ on one side, then there is ‘2' on the other side”and four cards which have a letter on one side and a number on the other side. Subjects are asked to indicate the card(s) that definitely need to be turned over to see if the conditional statement has been violated. The results of literally hundreds of studies of this kind showed that only about 10 % of the subjects made the normatively (as described by propositional logic) correct selection in this problem. When, in contrast, the conditional statement is present in contextualized form such as “If an employee works on the weekend, then that person gets a day off during the week” subjects’ performance improved dramatically. Gigerenzer and Hug (1992), for example, reported that typically 70 - 90 % of subjects made the normatively correct selection in the day-off version of the four-card problem. Gigerenzer and Hug (1992) also detected significant “perspective effects”, indicating that performance in the various versions of the Wason selection task is a systematic function of the perspective that subjects were given (e.g., that of the employer or that of the employee in the day-off version of the four-card problem.) It is just not context alone that was responsible for these dramatic improvements but perspective too.

While this insight is a ringing endorsement of experimental economists’ practice of letting subjects enact the roles of employers and employees, agents and principals, the overall results of the impact of contextualization strongly suggest that experimental

economists' practice of decontextualization may not have the beneficial effect that was used to rationalize it.

Over the past decade a few experimental economists have picked up on this theme. Dyer & Kagel (1996), for example, addressed the issue why "sophisticated bidders" (here executives from the commercial construction industry) suffer, like other subjects, from winner's curse in common value laboratory auctions. Simple survivorship arguments suggest that their failure to shade bids cannot be an equilibrium. Indeed, Dyer and Kagel identified a number of differences between theoretical and experimental treatments of one-shot common value auctions and practices in the commercial construction industry, among them ways for the low bidder to withdraw bids without penalty because of "arithmetic errors." Hannan, Kagel, & Moser (forthcoming) find that US undergraduate students provide, in principal-agent games, substantially less effort than do MBAs (who tend to have years of work experience). Like the experimental participants in the small-scale societies in which Henrich et al experimented, or the executives from the commercial construction industry, these MBAs brought field referents, i.e. rules of thumb, or heuristics, that they acquired in their daily lives, into the laboratory situation. Harrison & List (2003a) provide additional examples.

The idea that experimental participants bring to the laboratory rules of thumb, or heuristics, or "repeated game expectations that characterize their daily life" (Hoffman et al. 1996, p. 300) is not a new one. But it has gained some currency lately. Let us hence briefly speculate where people get the rules of thumb, or heuristics, or expectations with which they enter both the lab and field experiments, and why they often have trouble leaving them at the front door.

One typical line of argumentation from evolutionary psychology suggests that humans typically live in repeated-game contexts and are "intuitive statisticians" (Cosmides & Tooby 1996): they construct population frequencies of the typical actions (or characteristics) of members of a particular category in a particular context through "natural sampling" (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby 1996). These "base rates" of actions (or characteristics) will be typically used to predict the behavior (nature) of the members of a particular category in a particular context. In this sense, repeated-game contexts generate base rates that enter decision making under uncertainty.

Many laboratory situations, and field situations, are far removed from everyday experiences.[6] Instructions are instances of "individuating information" about situations that are atypical in many respects (e.g., a one-shot game or finitely repeated game, abstractness of the laboratory scenario, existence of an experimenter whose agenda may not be obvious, etc.). Most importantly, instructions provide information about the incentives that members of a particular category have in a very particular and, in any case, highly unusual context. This individuating information, if it is understood at all (e.g., Binmore 1999), often leads experimental participants to inferences about behavior which contradict what base rates suggest. Elsewhere, Ortmann & Hertwig (2000) have conjectured that this seems to be the case for all those games where the intuitive game-theoretic predictions for one-off or finitely repeated games on the one hand and indefinitely repeated games on the other, differ (e.g., public good provision/common resource problems, gift exchange/principal-agent, ultimatum, trust, and moon-lighting games, etc.).

[6] What did the ultimatum or dictator game mean to the experimental participants in those small-scale societies in Henrich et al 2001, 2002? And for that matter for your average Western student, whether in anthropology or in economics?

There is evidence from psychology that speaks to the issue of the circumstances that allow for individuating information to overcome base rate beliefs. Contra the proposition – commonly known as base rate fallacy – that "judgements are dominated by the information extracted from the evidence while prior beliefs are ignored or largely underutilized" (Ofir, 1988, p. 343), there is strong evidence that base rate beliefs are firmly entrenched and not easily overcome. In fact, in some quarters the base rate fallacy is now considered "a myth" (Koehler 1996, p. 5) and a consensus seems to be emerging that everyday experiences typically induce potent and resilient base rates that are not easily unsettled by individuating information and that the base rate fallacy may well be a myth for most of those situations that people typically encounter in real life (e.g. Koehler 1996; Cosmides and Tooby 1996). Underweighing of individuating information typically happens for everyday base rates (e.g., social or sex stereotypes). Based on this evidence, our working hypothesis is that experimental subjects come into the laboratory with similarly "potent and reliable" base rates about social contexts, not allowing them to pay due attention to the individuating information about the laboratory situation. We note that many anomalies in economics are connected to experimental designs that try to implement one-off or finitely repeated game or decision scenarios, i.e., situations where experimental participants have to overcome the very potent and resilient base rate experiences that serve them well in their daily lives.

Henrich and his collaborators seem to make a similar argument when they suggest that "behavior in the experiments is generally consistent with economic patterns of everyday life in these societies." (Henrich et al 2001). Indeed, an emerging neuroscience literature (e.g., Glimcher 2003; Gold & Shandlen 2001; Platt 2002; Sanfey et al. 2003;

Schall 2001; Schultz 2000; Schultz and Dickinson 2000) makes a strong case for such a view that supplements both the empirical evidence from the base-rate literature and the literature on the automaticity of our decisions (Bargh 1999, Bargh & Chatrand 1999, Chen and Bargh 1997).

We conclude that there is a good chance that many of the experimental results for the classes of games discussed here are artifacts of experimental design and implementation. Unfortunately, while anomalies are easy to construct, they are very difficult to deconstruct because such deconstruction involves, at least for many of the games considered above, the construction of a counterfactual reality.[7]

It is important to keep in mind that the rules of thumb, heuristics, and expectations that experimental participants bring to the laboratory, or field experiments, can work in both directions. As we have shown above for logic reasoning tasks, and as regards auctions, contextualization may well have positive effects in the sense of generating results that are in sync with canonical decision or game theory. As regards social dilemma or bargaining situations, however, contextualization might contribute to anomalies. Our understanding of the effects of field referents is clearly incomplete and it is very desirable that we understand their effects both in the field, and in the lab (where we should have understood them a long time ago.) For the time being, if in doubt, I propose to follow the advice of Hertwig & Ortmann (2001): Do it both ways!

[7] That said, the literature on base rates – because it demonstrates both the presence and absence of base rate effects, and seems to do so under well-defined circumstances – may guide us to a better understanding of where anomalies in experimental economics come from and how they can be made to disappear.

5. Discussion

“Methodological discussion, like spinach and calisthenics, is good for us ... “ said Paul Samuelson (1963, p. 231) famously, and tongue-in-cheek, in a comment to a debate on methodology published in the *American Economic Review*. Samuelson also said famously, and not tongue-in-cheek, that "(e)conomics cannot perform the controlled experiments of chemists or biologists because (it) cannot easily control other important factors. Like astronomers or meteorologists, (economists) generally must be content largely to observe." (Samuelson and Nordhaus, 1985, p. 8) One is tempted to point at the surge in experimental research over the past two decades to suggest to Samuelson that he surely got that one wrong. Very wrong indeed. And by and large that seems a defensible position. Experimental economics has come a long way to becoming respectable and it has established itself as a tool sine qua non in many economists’ methodological tool box.

That said, as experimental economists venture” into the wild” (the field), they are bound to encounter increased demands for controlling carefully the environmental and demographic characteristics, broadly construed, of their settings and participants. While “naturally occurring settings” open up opportunities (e.g., Ortmann & Gigerenzer 1997; Harrison & List 2003a) they also provide increased opportunities for poorly controlled experiments. In other words, field experiments are not unproblematic and they bring about new demands in terms of what needs to be controlled and, therefore, important trade-offs for experimenters to ponder . They also bring to the fore a question that experimental economists have ignored far too long - the ambiguous role that field referents can play.

The two potential drawbacks I have discussed above map roughly into, and illustrate, the two central problems that are raised in Henrich et al. (2002): To what extent is the variation between human groups that has been observed in the field work a function of the variability of experimental practices (e.g., the effects of subtle and unintentional cueing on the part of a questioner), and to what extent are the unselfish behaviors and motives that we experimental economists have observed a function of the rules of thumb, or heuristics, or field referents that our subjects bring to field, or for that matter, traditional lab experiments?

6. Conclusion

The recent surge in field experiments has prompted a surge of ruminations about experimental methods both in the field and the lab. This is a healthy development because they have triggered overdue reflections on the practices of experimental economists. While field experiments might help experimental economists to broaden their subject pools, and to increase the external validity of their investigations, I argue that these possible advantages do come at costs that can be considerable. When the dust has settled, the most important contribution of field experiments might well be their having opened the door for overdue methodological discussions on issues such as field referents.

7. References

Bargh, J.A. & T.L. Chartrand (1999). The unbearable automaticity of being. American Psychologist, 54, 462-479.

Bargh, J. A. (1999). The cognitive monster: The case against controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), Dual process theories in social psychology. New York: Guilford.

Binmore, K. (1999). Why experiment in Economics? The Economic Journal, 109, 16-24.

Chen, M. & J.A. Bargh (1997). Nonconscious behavioral confirmation processes: The self-fulfilling nature of automatically-activated stereotypes. Journal of Experimental Social Psychology, 33, 541-560.

Cosmides, L. & J. Tooby (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. Cognition, 58, 1-73.

Dyer, D. & J. Kagel (1996). Bidding in Common Value Auctions: How the Construction Industry Corrects for the Winner's Curse. Management Science 42(10), 1463-75.

Evans, J.St.B.T., S.E. Newstead, & R.M.J. Byrne (1993). Human reasoning: The psychology of deduction. Erlbaum: Hillsdale, NJ.

Gigerenzer, G. & U. Hoffrage (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.

Gigerenzer, G. & K. Hug (1992). Reasoning about social contracts: Cheating and perspective change. Cognition, 43 (2), 127-171.

Gigerenzer, G., P. Todd, & the ABC Research Group (1999). Simple Heuristics That Make Us Smart. Oxford University Press: New York.

Glimcher, P.W. (2003). Decisions, uncertainty, and the brain: The science of Neuroeconomics. MIT Press: Cambridge, MA.

Goeree, J.K. & C.A. Holt (1999), Stochastic game theory: for playing games. not just for doing theory. Proceedings of the National Academy of Sciences 96, 10,564-67.

Goeree, J.K. & C.A. Holt (2001), Ten Little Treasures and Ten Intuitive Contradictions. American Economic Review 91.5., 1402-22.

Gold, J.I. & M.N. Shandlen (2001). Neural computations that underlie decisions about sensory stimuli. Trends in Cognitive Science, 5, 10-16.

Gueth, W., R. Schmittberger, & B. Schwarze (1982). An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization, 27, 367-388.

Hannan, R.L., J. Kagel, & D.V.Moser (forthcoming). Partial Gift Exchange in an Experimental Labor Market: Impact of Subject Population Differences, Productivity Differences and Effort Requests on Behavior. Journal of Labor Economics.

Harrison, G. & E.E. Rutstroem (2001), Doing it both ways - experimental practice and heuristic context. Behavioral and Brain Sciences 24(3), 413-4.

Harrison, G.W. & J.A. List (2003). Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winner's Curse. Manuscript.

Harrison, G.W. & J.A. List (2003a). What constitutes a field experiment in Economics? Manuscript.

Henrich, J. (2000). Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining Among the Machiguenga of the Peruvian Amazon. American Economic Review 90(4), 973-79.

Henrich, J. (2000a). Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining Among the Machiguenga of the Peruvian Amazon. Uncut version.

Henrich, J. (2001). Challenges for everyone: Real people, deception, one-shot games, social learning, and computers. Behavioral and Brain Sciences 24(3), 41415

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H Gintis, & R. McElreath (2001), In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. American Economic Review 91(2), 73-78.

Henrich, J., R. Boyd, S, Bowles, C. Camerer, E. Fehr, H, Gintis, & R. McElreath (2002), Cooperation, Reciprocity and Punishment: Experiments from 15 small-scale societies. Book manuscript, chapter 2: Overview and Synthesis.

Hertwig, R. & A. Ortmann (2001), Experimental Practices in Economics: A Methodological Challenge for Psychologists? Behavioral and Brain Sciences, 24(3), 383-403.

Hoffman, E., K. McCabe, & V. Smith (2000), The impact of exchange context on the activation of equity in ultimatum games. Experimental Ecoomics 3(1), 5-9.

Hoffman, E. & M. Spitzer (1985), Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distribution Justice, The Journal of Legal Studies14(2), 259-97.

Koehler, J.J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. Beavioral and Brain Sciences, 19(1), 1-53.

Kreps, D.M. (1990). A course in microeconomic theory. Princeton University Press.

List, J.A. (2001). Do Explicit Warnings Eliminate the Hypthetical Bias in Elicitation Procedures: Evidence from Field Auctions for Sportscards. American Economic Review 91(4), 1498-1507.

List, J.A. & D. Lucking-Reilly (2000). Demand Reduction in Multiunit Auctions: Evidence from a Sportscard Field Experiment. American Economic Review 90(4), 961-72.

Mas-Collell, A., M.D. Whinston, & J.R. Green (1995). Microeconomic Theory. Oxford University Press: New York.

McKelvey, R.D. & T.R. Palfrey (1995). Quantal Response Equilibria for Normal Form Games. Games and Economic Behavior 10(1), 6-38.

McKelvey, R.D. & T.R. Palfrey (1998). Quantal Response Equilibria for Extensive Form Games. Experimental Economics 1(1), 9-41.

Ofir, C. (1988). Pseudodiagnosticity in Judgement Under Uncertainty. Organizational Behavior and Human Decision Processes, 42, 343-363.

Ortmann, A. & G. Gigerenzer (1997). Reasoning in Economics and Psychology: Why Social Context Matters. Journal of Institutional and Theoretical Economics 153(4), 700-10.

Ortmann, A. & R. Hertwig (2000). One-off scenarios as individuating information, repeated-game contexts as base rate information: On the construction and deconstruction of anomalies in economics. Manuscript, presented at the ESA meetings, New York, June 2000.

Ortmann, A. & R. Hertwig (2002). The Costs of Deception: Evidence from Psychology. Experimental Economics, 5(2), 111-131.

Ortmann, A., R. Hertwig, & G. Harrison (in prep), Experimental Methods in Psychology: A Challenge for Economists?

Pfungst, O. (1911). Clever Hans: The Horse of Mr. von Osten. Thoemmes Press: Bristol.

Platt, M.L. (2002). Neural correlates of decisions. Current Opinion in Neurobiology, 12, 1-8.

Rosenthal, R. & K.L. Fode (1963). The effect of experimenter bias on the performance of the albino rat. Behavioral Science, 8, 183-189.

Rosenthal, R. & R.L. Rosnow (1991). Essentials of behavioral research: Methods and data analysis, 2nd edition. McGraw Hill.

Rosenthal, R. & D.B. Rubin (1978). Interpersonal expectancy effects: The first 345 studies. The Behavioral and Brain Sciences, 1, 377-415.

Roth, A.E. et al. (1991). Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. American Economic Review, 81(5), 1068-95.

Roth, A.E. (1995). Bargaining Experiments. In: The Handbook of Experimental Economics. Princeton University Press: Princeton, NJ.

Samuelson, P.A. (1963). Discussion contribution. American Economic Review, 53(2), 231-36.

Samuelson, P. & W. Nordhaus (1985), *Economics* 12th edition. McGraw Hill Company: New York.

Sanfey, A.G. et al. (2003). The Neural Basis of Economic Decision-Making in the Ultimatum Game. Science, 300, 1755-58.

Schall, J.D. (2001). Neural Basis of Deciding, Choosing and Acting. Nature Reviews Neuroscience, 2, 33-42.

Schultz, W. (2000). Multiple reward signals in the brain. Nature Reviews Neuroscience, 1, 199-207.

Schultz, W. & A. Dickinson (2000). Neuronal coding of prediction errors. Annual Reviews Neuroscience, 23, 473-500.

Smith, V.L. (1976). Experimental Economics: Induced Value Theory. American Economic Review 66 (2), 274-9.

Smith, V. (2001). From old issues to new directions in experimental psychology and economics. Behavioral and Brain Sciences 24(3), 428-9.

Smith, V. (2002), Method in Experiment: Rhetoric and Reality. Experimental Economics 5(2), 91 - 132.

Starmer, C. (1999), Experiments in Economics: should we trust the dismal scientists in white coats? Journal of Economic Methodology 6(1), 1-30.

Todd, P.M. & G. Gigerenzer (2000), Precis of Simple Heuristics That Make Us Smart. Behavioral and Brain Sciences 23, 727 - 80.

Wozniak, R.H. (1999), Oskar Pfungst: Clever Hans (The Horse of Mr. Von Osten) (1907; English 1911). Extract from *Classics in Psychology, 1855 - 1914: Historical Essays* at www.thoemmes.com/psych/pfungst.htm