The challenge of representative design in psychology and economics

Robin M. Hogarth*

Universitat Pompeu Fabra, Barcelona

robin.hogarth@upf.edu

May 2004

* Robin M. Hogarth is ICREA Research Professor at Universitat Pompeu Fabra, Barcelona, Spain. This research was financed partially by a grant from the Spanish Ministerio de Ciencia y Tecnología.

Abstract

The demands of *representative design*, as formulated by Egon Brunswik (1956), set a high methodological standard. *Both* experimental participants and the situations with which they are faced should be representative of the populations to which researchers claim to generalize results. Failure to observe the latter has led to notable experimental failures in psychology from which economics could learn. It also raises questions about the meaning of testing economic theories in "abstract" environments. Logically, abstract tests can only be generalized to "abstract realities" and these may or may not have anything to do with the "empirical realities" experienced by economic actors.

Keywords: experiments, representative design, sampling

JEL classification: B41

It is hard to deny the value of experiments and, yet, not all experiments produce results that people trust. What, it can be asked, characterizes experiments that generate "valid" knowledge from those that don't? The goal of this paper is illuminate this question.

The answer can be sketched as follows. First, the goal of experimentation is to test theory by which is meant statements of conditional expectation, e.g., if condition *x* holds, effect *y* is more likely to occur. (Of course, *x* and *y* are typically more complicated than this simple notation suggests.)

Second, in conducting experiments, investigators are, in effect, building models of situations, models that Chapanis (1961) appropriately termed "replica" models. That is,

> The essential thing about replica models is that they look like the thing being modeled in some respect….A globe is a replica model of the earth because, in some respects, it looks like the earth. (Chapanis, 1961, pp. 115-116).

Third, the belief people place in the results of experiments depends on how well the replica model captures the intended "reality" on relevant dimensions. In other words, a globe is a good model of the earth if the relevant issue only concerns its being a sphere but not if the investigator is interested in, say, the effects of different climates by regions. In brief, I shall argue that the concept of *representative design* (Brunswik, 1956) can be used to assess whether experiments capture the relevant features of intended realities. By respecting its demands, researchers can greatly increase the value of experiments.

This paper is organized as follows. I first comment on what I mean by "theory." Second, I elaborate on the concept of representative design and point out that, to generalize results, not only should samples of participants be representative, but also samples of tasks or situations. I also emphasize the importance of contextual

variables that are typically ignored in economics. Finally, I conclude that although representative design demands much of experimentalists, it provides a clear rationale for generalizing results.

Some comments on "theory"

The word "theory" can be thought of as describing a belief or conditional expectation (see above) that allows one to take actions in the world or make statements about the "way things are." Indeed, if we did not have many "theories," we would not be able to function. In this sense, we are all great "theorists" and, although general, this definition helps think about the nature of theories in science and how to test them.

First, most theories of behavior (the subject of much of economics, sociology, and psychology), deal with the kinds of actions people take in specific environments. Thus these theories are conditional statements of (a) what kinds of people take (b) what kinds of actions in (c) what kinds of environments.

Second, following Popper (1959), I do not believe that we can prove theories to be correct. We can only disprove theories. However, this is not a major difficulty. There are many theories that are wrong but nonetheless useful. For example, most of our actions are consistent with the so-called "flat earth" theory. This is typically "good enough." But, critically, we also understand when it does not apply (e.g., in dealing with intercontinental travel).

Third, if one follows Popper's logic, we can never attain "correct" knowledge about anything. However, this is no reason for pessimism. For one thing, theories don't have to be 100% "correct" (as noted above). Second, the history of science is replete with examples of beliefs (i.e., theories) that have been disproved subsequently.

Future generations will undoubtedly ridicule many of our current theories – although hopefully we will be considered more enlightened than our predecessors.

Inherent in these arguments is that good theories make accurate predictions. However, in science theories need to do more than just predict. They should be parsimonious, elegant, and lead to surprising (or "interesting") implications. At the same time, there is a need for consistency between theories (i.e., beliefs). Although no theory is "correct," experiments can help choose between theories. In particular, we need to understand *when* theories do and do not make accurate predictions.

Representativeness and representative design

When people talk about the "representativeness" of experiments they typically refer to what psychologists call "external validity," that is, can the effects observed be generalized outside the experimental setting? Using Chapanis's (1961) replica model analogy, this can be framed by asking how well the model mimics reality on characteristics relevant to the theory. Is the behavior a "good" sample or representation of the intended reality on the relevant dimensions?

Brunswik's (1956) concept of *representative design* provides an intellectually satisfying way of framing this question. An experiment, it is argued, can be viewed as a sample. In this case, sampling theory determines whether results can be generalized appropriately to a particular population. To do this, however, investigators need to specify relevant characteristics of both samples and populations. In addition, sampling should take place on two dimensions. One involves the participants; the other concerns the situations or tasks with which the participants are confronted. Valid inferences can only be achieved by sampling in a representative manner on *both* dimensions.

Describing experiments by whether representative sampling has taken place on both dimensions (yes or no) leads to a 2 x 2 classification as illustrated in Figure 1.

Insert Figure 1 about here

Cell 1 is the ideal situation in which both participants and situations are representative. In cell 2, we have representative participants but not in a representative situation. In cell 3, the situation is representative, but the participants are not. Finally, in cell 4 neither participants nor the situation are representative. Clearly, it would be wonderful if all experiments were in cell 1. However, they are not. The issues therefore center on how important this is in particular cases and what can be done about it.

Sampling participants. One of my colleagues, a professor of statistics, enjoys pointing out that the participants in experiments on our campus are not "representative." Typically aged 18 through 22, and majoring in economics or business, approximately half are female and gained access to our public university by achieving good grades in examinations. In many respects, they are probably "representative" of subject pools for much of experimental economics. The relevant question is whether they can be considered "representative" for purposes of testing economic theories.

If by "representative" my colleague means "representative of the population at large," he is correct. This goal can only be achieved by sampling participants from the general population. However, few – if any – experimentalists do this. Does this matter?

The answer, I believe, depends on the extent to which characteristics of actors are relevant to the theories being tested in the experiment. In many psychological

experiments, for example, basic cognitive processes such as limits of attention and memory span are unlikely to be affected by differential sampling of participants (unless, for example, one was specifically studying skills of particular groups such as the very young or elderly). Thus, results from the "typical" subject pool are unlikely to be distorted. [1]

Economic theory is assumed to be universal and does not specify, for example, demographic characteristics of its actors. What it does do is to specify that actors are informed in the sense that they have some experience of the tasks with which they are confronted. What being informed really means is not clear. But experimentalists can and do typically ensure that participants understand instructions and/or take part in learning trials.

There are, of course, some economics experiments where roles are important and, for which, it is hard to train student participants adequately. As an example, consider situations where students are required to simulate managers and deal with issues such as hiring and firing that may be difficult to conceptualize unless one has had similar experiences. In these situations, it is not clear how to verify that participants are representative.

Parenthetically, I have often heard economists criticize experiments precisely because they are conducted with students. (Results are said not to be "representative.") However, if we take the fictional, theoretical characters that populate models of, say, rational expectations seriously, then nobody can be said to be representative in terms of assumed information processing capacity. In this case, if

[1] On the other hand, some recent evidence indicates that basic cognitive processes such as perceptions of "figure" and "ground" might be much more subject to cultural differences – and thus participants sampled – than previously thought (Nisbett, 2003).

representative experimental subjects are needed to test the theory, the theory is not testable.

Sampling situations. Whereas, in principle, investigators can overcome non-representative sampling of participants by recruiting people with appropriate profiles, the representative sampling of tasks typically provides the greater challenge. I provide two examples. Both involve attempts to measure human cognitive abilities: one to remember; the other to judge distances and the sizes of objects.

The first set of studies attempted to study and measure the limits of human memory. The basic paradigm (Ebbinghaus, 1885) involved having people memorize – in rote fashion – series of so-called nonsense syllables presented in random orders (e.g., DAX, ZUC, etc.). These attempts were successful in that they showed that memory is limited in measurable ways (i.e., the numbers of nonsense syllables that people can repeat correctly under specific conditions). However, memorizing nonsense syllables does not characterize how people encounter and remember information in the natural environment.

Imagine, for example, that you have just heard a good story that you would like to tell others. It is unlikely that you would remember the story verbatim, (i.e., as though memorizing nonsense syllables). Instead, you would explicitly use context and meaning to remember key features of the story (i.e., characters, specific actions, a time line of events, and the conclusion), and when re-telling the story, you would construct your narrative from this context and meaning. In other words, your memory relies heavily on contextual knowledge that enabled you to understand the story in the first place and also allows you to construct your version when you tell it (Bartlett, 1932). The nonsense syllables paradigm does not represent how people learn in their

natural environment. Indeed, this paradigm misled psychological researchers for years.

The second study was conducted by Brunswik (1944). His interest lay in investigating the ability to estimate distances and the sizes of objects and, contrary to the memory researchers, he investigated this in the person's natural environment (the University of California at Berkeley). The methodology consisted in having the person (a student) followed by an associate over a period of four weeks during which she was instructed to behave in her normal fashion. However, at irregular – or random moments – the participant was asked to estimate the sizes of objects that happened to be in her visual field as well as the distances from the objects. The associate then checked the accuracy of the estimates by measuring them. Note that in this study the experimenter did not choose the specific "experimental tasks" *per se* but instead defined a *process* by which these were a random sample of the participant's experience. Thus, although the study involved but a single participant, valid inferences could be made about her ability to judge distances and the sizes of objects.

Although conceptually sound, Brunswik's methodology is demanding and this may, in part, explain why it has not been used more in psychological research. However, recent years have seen growing use of what is now referred to as the *Experience Sampling Method* (ESM) using modern technology (e.g., beepers, mobile telephones, palmtop computers) to prompt participants to respond to questionnaires at random moments (see, e.g., Csikszentmihalyi & Larson, 1987; Hurlburt, 1997). For example, in a recent study I used mobile telephones to sample the decision making behaviors of undergraduate students and business executives (Hogarth, in press). Specifically, I was interested in how much feedback people receive and/or expect to receive on the many decisions they make each day. By sampling individual

participants up to four times a day for up to two weeks, I obtained random samples of the participants' decisions and was able to characterize the feedback they received or expected to receive on their decisions. This, in turn, is information that can be used to develop a deeper understanding of the nature of the decision making environments that people inhabit. Incidentally, although not a major study, the methodology generated some 1,200 data points (i.e., decisions) from 34 participants such that good estimates could be obtained at both individual and group levels.

Attention to how tasks are sampled is also relevant to the considerable controversy in the decision making literature concerning whether violate axioms of rational choice. In much of this research, people's decisions are examined in isolated situations and comparisons are typically made between groups of participants that receive different versions of the same problem (see, e.g., Kahneman & Tversky, 1979). The intriguing – and disturbing – results of this *between subjects* research is that different presentations of the same substantive problems systematically produce different choices. However, the validity of these results have been questioned on the grounds that, outside the laboratory, people typically make series of choices and receive feedback in what could be described more as a *within subjects* design for which the static *between subjects* design is inappropriate (cf., Gigerenzer, 1996; Hogarth, 1981). As in many debates, I believe that neither side is "correct." More importantly, people confront decisions in their lives that can be modeled by both *between-* and *within subjects* research designs. What should be criticized in the original research is the failure to specify the population of conditions outside the laboratory of which the experimental tasks can be considered a representative sample.

Recent work on risk taking by Weber, Shafir, and Blais (2004) provides a nice illustration. These researchers distinguish two ways in which people can acquire

information about risky choices. In one – typically used in experiments – all information is provided in the form of static descriptions. In the other – probably more frequent outside the laboratory – people acquire information experientially across time. This leads to an important difference in how people handle risks that occur infrequently. Faced with static descriptions, people are made aware of small probabilities of incurring risks and may even "over weight" them. On the other hand, because, by definition, outside the laboratory infrequent events occur infrequently, people have little experience with them. Thus, in these circumstances risks associated with small probabilities are not over weighted (Hertwig, Barron, Weber, & Erev, in press).

The correspondence between laboratory and external reality is, of course, an issue of great concern in many economics experiments. Thus, for example, incentives are typically real with participants being rewarded according to the level of their performance. Moreover, in market experiments the rules under which participants operate are usually quite realistic. There is little doubt that many of these replica models are good "representations" of markets that involve multiple actors and the possibility to learn from experience.

But, there are also many phenomena for which experiments are not well-suited for generalizing results. As a case in point, consider judgments of willingness-to-pay and willingness-to-accept on issues as trivial as small gambles to those as consequential as compensation awards in civil trials. Whereas the response mechanisms can be justified by economic theory, experimental participants are not machines that necessarily produce appropriate responses. Indeed, because most people's experience with such mechanisms is limited, their responses are often

sensitive to normatively irrelevant considerations. For good examples, see Sunstein, Hastie, Payne, Schkade and Viscusi (2002).

Economists are adept at handling the requirements of classic, factorial experimental designs. However, a major problem with factorial experiments is that the basic logic (i.e., the orthogonal variation of variables) precludes generalizing results outside the laboratory (Brunswik, 1956). The rationale does seem impeccable. By varying one variable at a time and holding all others constant, one can isolate the effects. However, outside the laboratory all other variables are not constant and variables are not orthogonal. Thus, estimates of the sizes of effects (based on the experiment) are subject to other forces. You can design factorial worlds within an experiment but this may not have much to do with what happens outside it.

Context in experiments. An intriguing aspect of many experiments in economics is that tasks are described in abstract terms. The rationale is that because the theory is stated in abstract terms, it should be tested in an abstract way.

It is important to emphasize that this argument does meet the criterion of the representativeness of tasks. That is, the abstract experimental task (or sample) is representative of the abstract theory (or population). However, in this case, the reasoning is somewhat circular and possibly even self-fulfilling. Surely, in economics the theory is useful to the extent that it predicts behavior in the economy and not just abstract representations?

Thus, "abstract" experiments are interesting but only as a first step. The next, and more important step, is to investigate the theory in situations that are representative of the economy at large. This does not mean, of course, that tests need to be done on all possible objects. One can sample situations relevant to the theory being generalized. For example, if the theory is being used to investigate issues in the

market for pig-iron as opposed to, say, paintings, then why not use stories that involve pig-iron or similar products?

The importance of this point is bolstered by the fact that one of the major lessons of research in psychology over the last 50 years has been the importance of context. People react differently to structurally identical problems that are presented in different ways, for example, abstract vs. concrete, or two different contexts. Consider Wason's (1966) famous card problem.

> Imagine that you have in front of you four cards that show (from left to right) the letter *A*, the letter *B*, the number *4*, and the number *7*. You are informed that each card has a letter on one side and a number on the other. You are also informed of the rule, "If a card has a vowel on one side, it has an even number on the other side." Given the four cards in front of you, *which and only which* cards would you need to turn to verify whether the rule is correct? (Hogarth, 2001, p. 116).

The modal responses to this problem (replicated many times) are the cards showing the letter *A* and the number *4* and the card showing just the letter *A*. In its abstract form, the problem requires testing a logical rule of the format *if p then q*, that is, if vowel then even number. In this case, the correct answer is to test the first and the fourth cards, thereby providing both possible confirming and disconfirming checks. On the other hand, when the structurally identical problem is presented in more familiar contexts, people have little difficulty in answering the question correctly. For example, consider the following variation (Gigerenzer & Hug, 1992).

> Imagine that you work in a bar and have to enforce the rule that, in order to drink alcoholic beverages, patrons must be over twenty-one years old. You observe four "young" people in the bar: the first is drinking beer; the second is drinking Coke; the third is twenty-five years old; and the fourth is sixteen years old. Whom do you check in order to verify that the rule is being enforced? (Hogarth, 2001, p. 116).

People do not see the structure of problems directly. Rather, they infer the structure from context (Einhorn & Hogarth, 1981) and this conditions the meaning they attribute to their perceptions.

Economic phenomena are not immune to these kinds of effects. For example, much has been made of the gambling metaphor of choice that characterizes behavior in terms of probabilities and utilities. However, there is abundant evidence that choice – as well as the assessment of risk – is highly dependent on context (see, e.g., Loewenstein, Weber, Hsee, & Welch, 2001). Nor can market discipline be counted on to eliminate "irrelevant" effects. For instance, consider a recent investigation of markets where some firms incur costs to ensure that their goods are produced ethically (e.g., no child labor). Participants playing the roles of consumers are sensitive to this contextual factor and pay an "ethical premium" for their purchases. Indeed, this behavior reduces their own earnings from the experimental sessions (Rode, Hogarth, & Le Menestrel, in preparation).

Concluding comments

For experimental studies to be judged as "representative" – and thus to generalize – claims need to be established for this (relative to the theory being tested) on the dimensions of *both* experimental participants and situations. This, in turn, requires specifying what the theory has to say in respect of both people and situations. Moreover, tests should be made of the sensitivity of results to the selection of participants and experimental situations. Often, understanding an effect can only be achieved when it can be turned "on" and "off" in experimental settings. For example, how big does a loss need to be before people are concerned by loss aversion?

Theories that are "abstract" pose their particular problems – does this mean that the findings should hold up under all possible environments or none? If economists wish to apply abstract theories to concrete situations then the latter need to be sampled in the testing process.

Finally, experimental economics provides but one set of tools. If theories are important, they should be tested by different methodologies. Indeed, the principles of representative design can be profitably used in programs of research that deliberately employ both experimental and field (empirical) data in testing and extending theories. These, however, often require collaboration between researchers in different silos of academic disciplines the difficulty of which may be the greatest obstacle in generalizing results.

References

Bartlett, F. C. (1932). *Remembering*. London,UK: Cambridge University Press.

Brunswik, E. (1944). Distal focusing of perception: Size constancy in a representative sample of situations. *Psychological Monographs*, 56 (254), 1-49.

Brunswik, E. (1956). *Perception and the representative design of experiments (2d ed.).* Berkeley, CA: University of California Press.

Chapanis, A. (1961). Men, machines, and models. *American Psychologist*, 16, 113-131.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, 175, 526-536.

Ebbinghaus, H. (1885). *Memory*. (Translated by H. A. Ruger & C. E. Bussenius.) New York, NY: Teachers College, 1913. Paperback ed., New York: Dover, 1964.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, 32, 53-88.

Gigerenzer, G. (1966). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, 103, 592-596.

Gigerenzer, G., & Hug, K. (1992). Domain specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, 43, 127-171.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (in press). Decisions from experience and the effect of rare events. *Psychological Science*.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional consequences of judgmental heuristics. *Psychological Bulletin*, 90, 197-217.

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL: The University of Chicago Press.

Hogarth, R. M. (in press). Is confidence in decisions related to feedback? Evidence – and lack of evidence – from random samples of real-world behavior. In K. Fiedler & P. Juslin (eds.), *In the beginning there is a sample: Information sampling as a key to understand adaptive cognition*. Cambridge, UK: Cambridge University Press.

Hurlburt, R. T. (1997). Randomly sampling thinking in the natural environment. *Journal of Consulting and Clinical Psychology*, 67 (6), 941-949.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.

Loewenstein, G.F., Weber, E. U., Hsee, C. K., Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127(2), 267-286

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently*. New York, NY: The Free Press.

Popper, K. R. (1959). *The logic of scientific discovery*. New York, NY: Basic Books.

Rode, J., Hogarth, R. M., & Le Menestrel, M. (in preparation). *Are consumers prepared to pay for fair trade? An experimental investigation*. Barcelona, Spain: Universitat Pompeu Fabra.

Sunstein, C. R., Hastie, R., Payne, J. W., Schkade, D. A., & Viscusi, W. K. (2002). *Punitive damages: How juries decide*. Chicago, IL: The University of Chicago Press.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology*. Harmondsworth, UK: Penguin.

Weber, E. U., Shafir, S., & Blais, A-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, 111 (2), 430-445.

Figure 1 -- Representativeness of experiments: By participants and situations

Situations: **Participants:**	**Representative**	**Not-representative**
Representative	**1**	**2**
Not-representative	**3**	**4**