Experimental Methodology on the Move

A review of *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter. Oxford and New York: Oxford University Press, 2015. Pp. xii + 477. ISBN 978-0-19-532832-5 (hardback).

The volume under review is the second of a series of 'Handbooks of Economic Methodologies' published by Oxford University Press. Like its predecessor (Caplin and Schotter, eds. 2008), it is based on papers and talks presented at the Center for Experimental Social Science of New York University, and its contents reflect in part the research interests of the host institution.

It is worth saying at the outset that these are no 'handouts' in the conventional sense of the term. Instead of providing a series of wide-ranging overviews of recent research, the two volumes contain focused discussions of specific issues that are currently hot in the discipline, with papers debating the pros and cons of controversial proposals, often in an antagonistic way. Although a non-expert should be able to reconstruct the background and make up his or her own mind about these issues, the authors (intentionally) do not make the effort to present them in a fair and objective manner. Instead, they forcefully defend partisan views and attempt to refute alternative positions in a dialectical exchange with their colleagues.

This makes a lively reading, and gives an interesting snapshot of the way in which methodological practices are changing in this area of economics. After a lengthy introduction by the Nobel Laureate Alvin Roth, the handbook is organized in three large sections, devoted to 'Economic Theory and Experimental Economics', 'Psychology and Economics', and 'The Laboratory and the Field'. The chapters (twenty overall) vary in length and depth, with some playing the role of provocative 'target papers' and other, shorter commentaries providing alternative views on the same topics. The contributors are a good mix of first, second, and, sometimes, third-generation experimental economists, including influential scholars such as David Levine, Martin Dufwenberg, Uri Gneezy, Ido Erev, John Kagel, Gary Charness, John List, Colin Camerer, and Glenn Harrison, among others.

After Roth's partly historical and partly forward-looking introduction, the first substantial section (Part II of the volume) features a couple of chapters that outline two popular views about the relationship between theory and experiments in economics. In chapter 2 David Levine and Jie Zhang emphasize the predictive success of economic theory – highlighting the concepts of Nash

Equilibrium and Quantal Response Equilibrium – against the often repeated claim that rational choice theory has been thoroughly refuted by the experimental evidence. In particular they argue that the rational choice framework has got the resources to account for and to predict 'anomalous' behaviour by adding new parameters in its equations, as for example in so-called social preference models.

In chapter 3 Andy Schotter develops a thesis that he had previously outlined in a paper which is one of my all-time favourites in the methodology of experimental economics (Schotter 2006). According to Schotter, the theory of rational choice is 'Strong and Wrong', that is, an admittedly false but useful benchmark to be attacked by experiments. The key idea is that it is difficult to design good experiments unless one has a model that makes precise predictions about the way people should behave in specific circumstances. The theory of rational choice offers this service, makes it possible to observe anomalous behaviour, and consequently to ask specific questions about the factors that make people deviate from the benchmark prediction. Although I find the original paper sharper and more effective than this lengthy chapter, Schotter in my view does a better job at capturing the dynamics of research than Levine and Zhang's static view of theoretical hypotheses (where do they come from? What was the role of experiments in developing the theory?).

The other chapters in this section are of interest mainly to specialists. Andrew Caplin and Mark Dean illustrate a method to elicit 'process data', that is, evidence about the way in which individuals reach their final decisions, by searching the space of options. Muriel Niederle discusses the relationship between experimental design, testing, and the estimation of theoretical parameters. Finally, three short commentaries by Leeat Yariv, Martin Dufwenberg and Uri Gneezy and Pedro Rey-Biel conclude this section.

Ido Erev and Ben Greiner's chapter opens the section entitled 'Psychology and Economics: A Comparison of Methods' (Part III). Their chapter is a good counterpart to Schotter's 'Strong and Wrong' thesis: Erev and Greiner argue that both rational choice theory and the behavioural models built to explain anomalous behaviour suffer from limited applicability. The problem is that such models work mainly in those situations in which the decision-makers know the relevant parameters (such as the probabilities and payoff distributions). But usually people do not have all the relevant information when they are making a decision, and as a consequence their choices tend to go in directions that are orthogonal to those that one would expect from classic experimental results. The general message of Erev and Greiner's chapter is that it is dangerous to

build behavioural models on specific experimental counterexamples. Behavioural economists should invest more energy in making their theories quantitative, identifying parameters that are robust to small changes in contextual information. From an institutional perspective, Erev and Greiner also have interesting ideas about how to make this type of research more attractive to theorists and experimenters.

After a commentary by Keith Murnigham, two psychologists – Tom Tyler and David Amodio – draw an interesting comparison between psychology and behavioural economics. This chapter is particularly welcome because most cross-disciplinary comparisons so far have focused on the contrast between psychology and economics, assuming simply that the approach of behavioural economists is more or less analogous to the approach of the psychologists who study human decision making. According to Tyler and Amodio, on the contrary, there is a significant difference: while most psychologists are interested in the *direct measurement* of mental states, behavioural economists try to *infer* mental states from behaviour. Both psychologists and behavioural economists however agree that mental states are the main causes of behaviour, and that a proper understanding of human choice depends on our understanding of the underlying psychological mechanisms. Although Tyler and Amodio believe that economists have a lot to learn from psychologists, they are rather sceptical about the prospects of neuroscience. They argue that neuroeconomics is a way to by-pass the study of psychological states – a form of 'mindless' economics, paradoxically. The chapter ends with a plea for triangulation and the use of different techniques, complemented by Gary Charness' appeal (in a separate commentary) to disciplinary specialization accompanied by intensive cross-disciplinary communication.

The third section ('The Laboratory and the Field') is the most innovative of the volume, in my view. The main topic is the relationship between field and laboratory data, but more generally what is at stake is the generalizability (or external validity) of experimental results outside the narrow domain in which they have been collected. When I started working on the methodology of experimental economics, almost two decades ago, the entire literature on this issue comprised no more than half a dozen papers. The situation has changed quite dramatically since then, and external validity has now become a central topic in experimental economics.

The central chapter in this section is Steven Levitt and John List's paper 'What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World'. This chapter plays a similar role to Gul and Pesendorfer's 'The Case for Mindless Economics' in the first handbook. The main difference is that Gul and Pesendorfer's paper was previously unpublished, while Levitt and

List's paper has already appeared in the *Journal of Economic Perspectives* in 2007. (Strangely, although the paper is clearly indicated as a reprint in the title, the original source is not formally acknowledged.) Levitt and List's article has been a hit, and has become widely cited. The point of this reprint is to offer an arena for open discussion that has been missing so far.

For those who are unfamiliar with Levitt and List's article (LL from now on), here is a short summary: LL propose a 'theory' to explain the behaviour observed in classic laboratory experiments such as the Ultimatum Game, the Dictator's Game, the Trust Game, and the Public Goods Game. The 'theory' as a matter of fact is little more than a list of the factors that may influence people's behaviour in these strategic situations, and that should be included in their utility functions. LL review evidence indicating that monetary and non-monetary factors – including negative externalities, social norms, and experimental scrutiny – explain variations that cannot be accounted for by classic theories of social preferences. Last, they argue that these factors make it difficult to extrapolate from laboratory experiments to other situations of interest, and suggest that field experiments offer a better ground for generalizability.

Although the LL article includes many valuable insights, its controversial status derives in part from an ambiguity: on the one hand, it tries to make a methodological point about the external validity of lab experiments; on the other, it puts forward a critique of a particular strand of theoretical research ('social preference' models) that was literally booming when the article was originally published. This ambiguity has made it an elusive target for criticism. Some economists agree with the general methodological point but disagree with its theoretical implications. Others share LL's scepticism concerning social preference models but are keen to defend the generalizability of lab experiments. In general, these two issues have not been disentangled properly.

By reading the other chapters of this section however one can get a clear idea of what is at stake in this debate. The second central contribution is Colin Camerer's chapter, 'The Promise and Success of Lab-Field Generalizability in Experimental Economics'. As in the first handbook, where he replied to Gul and Pesendorfer, Camerer takes the role of the guardian of behavioural economics against LL's doubts. And again, as in the first handbook, Camerer's chapter is a model of how to engage in a methodological debate, making use of all the theoretical and empirical resources that are currently available.

Camerer begins by drawing a distinction between a 'policy view' and a 'scientific view' about external validity. His main point is that from a scientific perspective the problem does not exist, because every theoretical claim has a universal scope of application. The demonstration that a

theory fails in the laboratory is ipso facto a proof that the theory is unsatisfactory, *from a scientific point of view*. This line of argument has a prominent pedigree in experimental economics, dating back to the seminal methodological papers written by Vernon Smith and especially Charles Plott in the 1970s and 1980s. Given that Camerer has been Plott's colleague at Caltech for many years, it is not surprising to find them rehearsed here.

Overall I do not find these arguments convincing, as I have explained elsewhere, mostly because I do not think that the scientific view and the policy view can be sharply separated (Guala 2005, chapter 7; see also Bardsley et al 2009). The choice of which theory is worth testing in the laboratory is strongly influenced by our expectations about which theory is likely to be applicable in a particular context, and the demonstration that a theory works (or not) in the lab is only a preliminary step in a longer research path that must end at some point with a concrete application.

I do, however, agree with Camerer's next argument, namely that the problem of generalizability holds in the same way when a result derived from *field* data is applied to another (similar but not identical) field setting. In the end, Camerer argues, generalizability is an empirical issue that arises whatever source of data you are using. Coherently with this empiricist approach, he surveys more than twenty cases in which the results of laboratory experiments have been compared to similar field settings, and finds that only in two of them there seem to be problems of external validity: contrary to LL's suggestion, "there is actually little evidence that sociality experiments are misleading compared to field settings *when matched with similar features*" (p. 259).

I take this to mean that laboratory experiments are generalizable *locally*, that is, to situations that are similar to those that have been instantiated in the laboratory (see Guala 2005; it is reassuring to see the same cases being cited by Camerer, for example Kagel and Levin's results on the winner's curse, and Plott's 'testbed' experiments on combinatorial auctions).

Although Camerer's systematic analysis leaves the reader with the impression that not much can be saved in LL's arguments, to say that lab results extend to 'similar' conditions is not entirely satisfactory. How similar is 'similar enough'? Real external validity inferences, of course, rarely involve identical circumstances. In principle we would like to have a set of inferential methods to systematically address the issue of applicability also when field conditions are quite different from those found in the lab. One possibility, of course, is that no such methods exist. But then, if this is the case, we should honestly acknowledge that behavioural economists routinely make claims of applicability that go way beyond what can be justified using Camerer's 'localist' approach (for a specific, highly controversial example, see Baumard 2010, Guala 2012).

Here LL's ambiguity becomes relevant again. While LL's claims about the generalizability of experiments have failed to make a breakthrough, their attack against the pretensions of social preference theorists have been met positively by many experimenters. The sensible view, of course is that social preferences are a modelling technique, and their worth is quite independent of the experimental results upon which they are based. LL may be right about social preference models, but their argument is muddled by their decision to lead their campaign simultaneously at the methodological and theoretical level.

To add more tension, LL's critique has been read by many as a not-so-disguised attempt to promote field experimentation, a methodology that has been used intensively by John List over the last decade. LL are careful in their paper to state that field and lab experiments are complementary, and that both have a role to play in a successful research programme. Such a position is shared by virtually all handbook contributors, and yet the very fact that it has to be repeated over and over suggests that LL's contribution has been generally interpreted as an attempt to redirect research from the laboratory to the field. Glenn Harrison, co-author of a seminal methodological paper on field experiments (Harrison and List 2004) says explicitly that the complementarity thesis was lost in subsequent articles (authored by List alone) that stressed too much the inferiority of lab experiments.

Harrison's chapter (co-authored with Morten Lau and Elisabet Rustrom) provides an example of complementary lab-field research, taken from the authors' collaboration with the Danish government on the estimation of citizens' risk attitudes. Although it is fairly technical, this chapter contains useful material for those who are interested in mapping the path that leads from theoretical models to application, through several stages of estimation in the laboratory and in the field. The same applies to John Kagel's chapter, partly devoted to winner's curse experiments and partly focused on gift-exchange experiments, where the similarity between lab and field is much more tenuous. Kagel insists in particular on the observation that human learning is context-specific, a fact that should make us wary of generalizing to dissimilar contexts.

Guillaume Fréchette surveys experiments that compare the behaviour of students with the behaviour of professional subjects, finding that in most cases the difference is insignificant and in those where it is significant it differs in different ways. This chapter is an excellent example of methodology supported by empirical data, and shows that the experimental literature has grown so large in economics that most methodological questions can now be answered in the lab itself. (A good reminder for philosophers.)

Judd Kessler and Lise Vesterlund devote a whole chapter to elaborate a point that is made in passing by many contributors, namely that the issue of external validity often focuses on whether *quantitative* results are generalizable, while there isn't much disagreement on the fact that *qualitative* results (i.e. the sign or direction of effects) tend to hold across different contexts. Kessler and Vesterlund claim that there is no evidence of lab effects being reversed in different settings, and justify the emphasis on qualitative effects with the observation that idealized models containing simplifying assumptions are rarely instantiated exactly in real-world circumstances. Whenever disturbing factors are at work, we can at most hope that the direction of the effect is stable. Finally, Kessler and Vesterlund make a useful distinction between failures of experiments that are due to methodological flaws and those that are due to differences in the environments in which the results are applied: only the first are an indictment of the experimental approach, while the latter only signal an imperfect understanding of the conditions in which the results ought to be applied.

The handbook ends with a chapter by Omar Al-Ubaydli and John List, which is partly a new contribution and partly a reflection on the debate triggered by LL. The chapter starts by saying that lab and field are complementary (again) but then proceeds to build a framework that seems especially designed to make field experiments look better than lab experiments for the purpose of generalizability. The framework is presented formally, although I find most of the notation redundant, given the rather simple claims that the authors want to make. On the positive side, I appreciate the emphasis on causality and causal inference, an aspect that is often lost in a theory-driven discipline like economics.

Nonetheless, some of the claims made in this chapter are contentious. For example: it is not true that additive separability of the causal factors is unnecessary in field experiments, as Al-Ubaydli and List claim (p. 437), because randomization can lead to confounding results when uncontrolled factors interact non-additively with the main treatment variable (see Guala 2005, 132-4). There is no fast and cheap substitute to tight experimental control, and the authors' faith in randomized trials (p. 424) contrasts sharply with the recent wave of concerns about randomization in the bio-medical sciences (see e.g. Cartwright 2011).

The last fifteen pages of this chapter are commentaries and replies to LL's critics. There is a lengthy re-analysis of Camerer's empirical evidence about the generalizability of social preference experiments, a detailed rebuttal of Camerer's interpretation of List's own data from field experiments, and a largely convincing critique of the inflated external validity claims made by some

behavioural economists (especially Ernst Fehr on gift-exchange experiments). Finally, the volume ends with a programmatic statement that I wholeheartedly endorse and that makes one hope that more progress will be made on external validity during the next decade:

> We believe that at this point the field can move beyond strong statements that lab or field results will always replicate. This type of reasoning seems akin to standing on the stern of the Titanic and saying that it will never go down after the bow sinks below the water surface. Rather, it is now time to more fully articulate theories of generalizability and bring forward empirical evidence to test those theories. Building a bridge between the lab and the field is a good place to start. We hope that this volume moves researchers to use AFEs [Artefactual Field Experiments], FFEs [Framed Field Experiments], and NFEs [Natural Field Experiments] to bridge insights gained from the lab with those gained from modelling naturally occurring data. (457)

Nautical analogies apart (experimental economics looks more like the Victory than the Titanic, so far!) it is an appropriate way to end this handbook, and also a good greeting for a volume that every philosopher or social scientist interested in experimental methodology will have to read.

References:

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., & Sugden, R. (2009). *Experimental economics: rethinking the rules*. Princeton: Princeton University Press.

Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind and Society,* 9, 171–92.

Caplin, A. & Schotter, A. (eds. 2008). *The foundations of positive and normative economics*, New York: Oxford University Press.

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *The Lancet,* 377, 1400-1401.

Guala, F. (2005). *The methodology of experimental economics*. New York: Cambridge University Press.

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences,* 35, 1-59.

Gul, F. & Pesendorfer, W. (2008). The case for mindless economics. In A. Caplin and A. Schotter (eds.) *The foundations of positive and normative economics.* New York: Oxford University Press.

Harrison, G. W. & List, J. A. (2004). Field experiments. *Journal of Economic Literature,* 42, 1009-45.

Schotter, A. (2006). Strong and wrong: on the use of rational choice theory in experimental economics. *Journal of Theoretical Politics,* 18, 498-511.