Oxford Review of Economic Policy, Volume 30, Number 4, 2014, pp. 753–763

Cautionary notes on the use of field experiments to address policy issues

Glenn W. Harrison*

Abstract Field experiments are popular again in policy circles. There are various types of field experiments, with complementary strengths and weaknesses for different policy questions. There is also a lot of needless hype about what field experiments can do. More balance in our use of field experiments is called for.

Keywords: field experiments, randomized evaluations, welfare

JEL classification: C93, C81, D6, I31

I. Introduction

Field experiments have become popular tools, again, in the evaluation of policies in rich and poor countries. There are many good, scientific reasons for this popularity, as argued in Harrison and List (2004), but there is also a marketing gloss that comes along with most applications. This is common with intellectual fads, of course.[1] The challenge is to keep a balanced eye on what field experiments can do that improves on, or complements, other methods, and to call the nonsense for what it is.

The expression 'field experiments' has come to mean two very different things in the economics literature. For one group it just means any experiment conducted in the field that uses randomization; for others it means any experiment that is conducted in the field or that uses field referents, whether or not randomization has been used. This

*Department of Risk Management and Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, e-mail: gharrison@gsu.edu

Harrison is also affiliated with the School of Economics, University of Cape Town, and IZA—Institute for the Study of Labor. Valuable comments were received from Robert Hahn, a referee, and the editors of this issue.

[1] And with all fads, many of the ideas have been around a long time. The father of modern field experiments, Peter Bohm, spent a lot of time examining policy questions, reviewed in Bohm (2003). And 'social experiments' are just another flavour of field experiment, but with a long policy history: see Ferber and Hirsch (1978, 1981) for *surveys* of research, most employing randomization. Hausman and Wise (1985) and Manski and Garfinkel (1992) contain deep discussions of methodological issues with policy studies using field experiments, anticipating many of the statistical issues raised again in recent debates. And in psychology, for example, Deci (1971, p.111) stressed many of the same issues of control in field experiments compared to lab experiments.

doi:10.1093/oxrep/gru037

semantic distinction does have some bite in terms of how people design experiments and what they expect to get out of them, so it is not 'just a semantic distinction'. I will argue for the complementarity of these two types of field experiments, in the interests of being diplomatic here.[2] If one adopts the latter definition from the start, one does not have to worry about complementing a randomized evaluation with structural insights from another type of field experiment. You simply design the field experiment to answer the policy question at issue, and it may or may not include a randomization component.

The problem with most field experiments is that they tend to avoid wanting to make any structural claims about *why* things work. The slogan tells it all when someone says that they only care about *what* works. There is, to be sure, a cursory hand-wave at the theoretical literature on possible behavioural factors at work, but when it comes to what the evidence shows, and is intended to show, we get a net effect inside a theoretical black box. What is needed, in addition, are experiments to provide some structural insight into the processes at work. Which of the Big Three behavioural moving parts—in my view, risk attitudes, subjective risk perception, and time preferences—might account for observed behaviour? Only if we obtain some estimates of these structural parameters will we have any hope of describing why something is working or not, and then going further and undertaking a welfare evaluation.

There are several reasons that most field experiments fall short. The first is that they limit themselves to evaluations of *observables*: this price change in delivering that product leads to what revealed change in demand? The second is that they limit themselves to *average* effects: what is the average change in demand? The third is that their focus has been *partial* equilibrium: what is the effect of this price change for the commodity I am able to randomize (or study the randomization of)?

II. Looking for keys where the light is better

The problem with just looking at observables is that they tell us nothing about the latent variables that are of interest in welfare evaluation. For that we need to make inferences about consumer surplus, and for that we need to know a lot more about the latent preferences that people bring to their choices, such as risk preferences and time preferences. We also need to know a lot more about the subjective beliefs that people bring to their choices. The reason that there is this dogmatic focus on observables is easy to discern and openly discussed: a desire to avoid having to take a stand on theoretical constructs as maintained assumptions, since maintained assumptions might be wrong. The same methodological precept guides the choice of statistical methods, but that is another story about modelling costs and benefits. One can fill in these blanks in our knowledge about latent preferences and beliefs with theories and guessed-at numbers, or with theories and estimated numbers. But one has to use theory to make conceptually coherent

[2] I speak more directly on this issue in Harrison (2014, section 1). It is also proper to stress that randomized evaluations are excellent methods for doing what they set out to do in the narrow sense explained below. I just want more interesting policy questions answered, and fully expect that the excellent answers to the less interesting questions will be needed as part of that broader policy objective.

statements about preferences and beliefs, and then undertake welfare evaluations. That is the rub: an agnosticism towards theory.

Advocates of field experiments often portray the trade-off here in overly dramatic fashion. Either one uses the methods that avoid these theoretical constructs, or one dives head first into the shoals of full structural modelling of behaviour. This is a false dichotomy, raised as a cheap rhetorical device to still debate over the role of theory, and Heckman (2010) takes aim squarely at it. The latter structural modelling is very hard to do well, and quite easy to do poorly. The former 'reduced form' modelling is fine as far as it goes, but just does not go very far. The missing middle ground becomes apparent when empirical puzzles emerge, leading to casual theorizing and even more casual behaviourism, sadly illustrated in Banerjee and Duflo (2011) and Karlan and Appel (2011).

Moreover, the empirical methods of randomized evaluations are not as agnostic as some would like to claim. Rosenzweig and Wolpin (2000) provide an excellent, patient methodological discussion of the behavioural, market, and technological assumptions needed when one relies on naturally occurring randomization. Leamer (2010) similarly offers a patient methodological review of the statistical issues that lie beneath the inferences about all randomized evaluations, and a reminder that the statistical issues connect to broader behavioural and economic assumptions (e.g. homogeneity). In some cases, the latest generation of randomized evaluations is responding to these concerns. As one might expect, they are responding with more data, and even bigger grants, but at least they are responding: for instance, there are now several evaluations of the 'scaling problem' with field evaluations at a local level.[3]

There are now many examples of structural modelling of latent constructs using data from field experiments and natural experiments. Recent examples include Andersen *et al*. (2008, 2014*a*, *b*), Farber (2008), Crawford and Meng (2011), and DellaVigna *et al.* (2012). Comparable recent examples of structural modelling in the laboratory across a wide range of topics include Gill and Prowse (2012), Andersen *et al.* (2014*c*) and Dixit *et al.* (2014).

III. Gaussianity

Why the fascination with the average? On a good statistical day, it is one measure of central tendency, that is true. But there are many reasons why we are directly interested in knowing the full distribution, not just one *in*sufficient statistic.

First, we might simply care about winners and losers. Assume that there is a modest change in the average, in some direction that the researcher deems a welfare improvement. What if this comes about with large gross changes at the individual level: lots of people do wonderfully from the intervention, and lots of people do terribly from the intervention? One does not have to swerve too far from the strict Utilitarian social welfare concept to doubt if this is, indeed, a social welfare improvement.

[3] The scaling problem arises because of the assumption that the potential outcomes to an individual unit of observation should not be affected by the potential changes in the treatment exposure of other individuals. To economists, this is just a partial equilibrium assumption. See Morgan and Winship (2007, pp. 37ff) for a textbook introduction, and Garfinkel *et al.* (1992) for a history of the concept in economics and statistics.

Trade-offs between efficiency and equity aside, we might also be interested in identifying winners and losers in order to design a better intervention, in the spirit of the compensation criteria of welfare economics. Or to design a more robust intervention that could survive rent-seeking attacks from losers, and hence be more politically sustainable.

Second, we might care about the distribution when evaluating the 'policy lottery' that any intervention affords a decision-maker (Harrison, 2011*a*). One reason is to extend consumer sovereignty in welfare evaluation to consider the risk attitudes of those affected, if there is some statistical risk that any given individual is a winner or a loser. Another reason is to reflect uncertainty aversion or ambiguity aversion, arising from imprecision in estimated effects. Either of those two, which are often confused terminologically, require that one undertake welfare evaluation over the *distribution* of impacts, whatever specific modelling church one attends. These specific modelling altars vary in how they weight the distribution, but they all agree that the essence is to take it into account in some way: that is, in fact, what differentiates uncertainty and ambiguity aversion from 'familiar' risk aversion.

Third, and assuming away the statistical identifiability of whether any given individual is a winner or a loser, and even assuming away equity concerns, we might care deeply about the distributional shape of things to come from an intervention if it makes more people vulnerable to certain thresholds. In poor countries, the most important threshold is the absolute poverty level, defined here as that level of resources below which the individual unit experiences some asymmetric physiological effects. For now, equate resources with income. Imagine an intervention that keeps the fraction below that poverty line the same, bunches a lot of people ε above the poverty line when they were well above it prior to the intervention, and somehow allows the 'rich and famous' to enjoy gains such that the average income of the population increases. Surely the tsunami of vulnerable individuals hovering ε above the poverty line, compared to the baseline, should matter for our welfare evaluation? Again, to see the fundamental point, rule out equity effects, and rule out risk aversion (or even uncertainty aversion) with respect to the estimated impact of *this* intervention in *this* domain. What if there are 'background risks' that might nudge this tsunami of vulnerable individuals below the poverty line, even after we have ascertained the impact of the foreground intervention under study? It is a commonplace in developing countries, magnificently documented by the *Portfolios of the Poor* of Collins *et al.* (2009), that the poor face myriad risks at any given time, and value flexibility in risk management options.[4] This lesson obviously applies as well to the poor in rich countries.

There is no fundamental reason that field experiments, and randomized evaluations, cannot be tasked to look at distributional issues.

IV. Worms, teachers, fertilizer, and savings

One can certainly be interested in worms and whatever they do, absentee teachers and whatever they do not do, the optimal use of fertilizer, wherever it comes from,

[4] An important technical point must be made here: these issues arise naturally and conventionally if one assumes non-additive utility defined over multivariate risk. However, it is an unfortunate commonplace to assume additive utility in many applications, such that the risk over *final* wealth positions is all that matters.

savings rates, and so on. But these are not substitutes for the rigorous measures of welfare from a policy, given by the equivalent variation in income. We need these measures of welfare for the application of cost–benefit analysis familiar to older generations: comparing a *menu* of disparate policies potentially spanning all of these interventions (Harrison, 2011*a*, 2014). How do I decide if it is better to reduce worms, increase teacher presence, use fertilizer better, or increase savings rates, if I do not know the welfare impact of these policies in a way that allows comparability? Of course, the best intervention might be 'costless' to implement, but that is rare.

A related concern is the sample selection effect that comes from only doing field experiments on things that one is allowed to randomize, or that serendipity randomizes for us. What if we care about an intervention in some area that does not permit randomization, such as tariff policy? How do we then trade off interventions in the areas we can study, with those in the areas that we cannot study (or cannot study for the foreseeable decision-making future)?

It is often difficult to design a careful randomized evaluation quickly, not because of any flaws in the method, but because of the logistical constraints of coordinating multiple sites and obtaining necessary approvals. Worrall (2007, pp. 455–9) presents a detailed case study of a surgical procedure which was identified as being 'clearly beneficial' on the basis of observational studies, but where it took years to undertake the requisite randomized evaluation needed for the procedure to become widely recommended and used. Lives were lost because of the stubborn insistence on randomized evaluation evidence before the procedure could be widely adopted. Of course, counter-examples exist, but the costs and benefits of having complementary evaluation methodologies are often lost in the push to advocate one over the other.

The issue of the amount of time needed for field experiments in general is a deep one in terms of the trade-offs implied for policy-makers. The risks of doing 'quick and dirty' policy evaluations can be severe, both in terms of the economic costs of getting it wrong as well as the political costs of being the one to get it wrong. But there are, arguably, sensible ways of doing 'quick and clean' policy evaluations, or at least quicker and cleaner evaluations. The US Federal government, for instance, has been actively discussing ways to conduct 'small clinical trials' for years (e.g. Evans and Ildstad, 2001; and Campbell, 2005), and recently issued draft 'guidance for industry' on the matter (Food and Drug Administration, 2010*a*, *b*). Three statistical ideas arise naturally and could be more broadly evaluated for economic policy: the use of sequential testing procedures (e.g. Whitehead, 1997); the use of adaptive designs for treatments and sample sizes (e.g. Jennison and Turnbull, 2011); and the use of Bayesian methods to pool data from prior studies (e.g. Goodman, 2005; Louis, 2005; and Berry, 2005).[5]

[5] The obvious experimental design in economics is to take some treatment that has been demonstrated using traditional methods and see if one can use these alternative methods to come to the same conclusion, with the same statistical reliability but more quickly and typically with smaller samples. These methodological tests could be most easily undertaken in the lab, then applied in the field. Some recent work in lab settings does explore adaptive estimation of economic preference parameters: see Wang *et al.* (2010) and Toubia *et al.* (2013).

V. Causality

The concept of causality is central to the attraction of randomized evaluation methods, as indeed it is to all social science. There are two fundamental methodological concerns with the manner in which causality is treated (Harrison, 2013).

The first concern is that heterogeneity of subject response demands some attention to causal mechanisms when that heterogeneity might *interact* with the treatment. The need for some assumptions or structural attention to these mechanisms is at the heart of some of the seemingly arcane statistical issues surrounding randomized evaluation: see Heckman (1992), Deaton (2010), Keane (2010*a*,*b*), and Leamer (2010) for careful statements of the issues. Many students forget that 'sample selection' is fundamentally a problem of accounting for the effects of heterogeneity, and field experimenters are very casual about the use of corrections for sample selection.[6]

The second concern is that causal statements are almost always limited to *directly* observable outcomes in applied work, both in terms of treatments and in terms of outcomes.[7] Unfortunately, many of the things that interest economists involve things that are observable only as latent constructs. Answers to the following causal questions are of direct policy interest, and involve latent constructs.

1. What is the effect of a new microinsurance contract on individual or social welfare?
2. What contributory role do risk attitudes, time preferences, or subjective beliefs play in the level of take-up of a microinsurance policy?
3. Is an increase in the take-up of a microinsurance policy an indicator of improved individual or social welfare?

All of these involve some latent variable, and hence require some theory about what that variable is and how we measure it. For instance, what do we mean by welfare and how do we measure it? Instead, we find answers to simpler causal questions such 'what is the effect of a subsidy on take-up of a microinsurance policy?' This is an important, interesting, intermediate question, but simply does not characterize the full range of causal questions of interest to economists (Ross, 2014).

One implication of restricting attention to directly observable variables is that one often has to assume that a change in some directly observable outcome is a good thing in terms of welfare. Is increased take-up of a microinsurance product a good thing for the insured and society? This is actually a difficult and important question to answer, involving subtle questions about the measurement of preferences and constraints, as well as normative judgements. Collins *et al.* (2009) showed that the 'price of money' in poor countries is not always what it appears to be if one looks at the readily observable interest rate on a loan. The same is true of insurance products in general: one needs to study the detailed product carefully before one can evaluate its value to a client or

[6] These problems become particularly severe when the selection is on the risk attitudes of treated and untreated subjects, so-called 'randomization bias' (Harrison *et al.*, 2009; Harrison and Lau, 2014).

[7] This is quite apart from the problem that many of the characteristics of behaviour that make people interestingly heterogeneous are latent. This type of heterogeneity turns the first concern into a 'perfect storm', since one typically needs theory, auxiliary experiments, *and* some form of structural econometric modelling as 'can openers' to get at those latent traits.

society. All evaluations of microinsurance products with field experiments, however, assume that increased take-up is a good thing for all.

Just to be complete with our understanding of what words mean and do not mean, the word 'cause' is defined by the *Oxford English Dictionary* (2nd edn) as 'that which produces an effect; that which gives rise to any action, phenomenon, or condition'. So there is nothing here that limits the concept of cause and effect to things that are directly observable. This position should not be controversial, since it is 'programme effects' that field experimenters are interested in (e.g. Banerjee and Duflo (2009, p. 152)), and some of those effects might be directly observable and some might be observable only as latent constructs.

The next generation of field experiments will illustrate the value of combining tasks that allow one to estimate latent structural parameters with interventions that allow the sharp contrast between control and treatment. The next generation of econometric analysts will use the insights from these structural models to inform their understanding of the distributional impacts of interventions, rather than just the average impact. They will also use these structural parameters to gauge the sample selection issues that plague randomized interventions of sentient objects rather than agricultural seeds. And both groups of researchers will find themselves heading back to the lab to validate their behavioural theories, experimental designs, and econometric methods applied to field data. There they will find time to talk to theorists again, who have produced some beautiful structures needed to help understand subjective risk and uncertainty.

VI. Nudges, small and big

An important feature of some of the field applications of behaviour revealed in experiments is the use of 'nudges' to avoid or mitigate biases (Thaler and Sunstein, 2008; Sunstein, 2013). Although much of the core behavioural evidence for bias comes from lab settings, the policy application is in the field, and there are now many field experiments examining these as a policy approach. White (2014, p. 22) raises this issue in the broader debate in an important way:

> Random assignment is a method. It has no inherent ideology. One could as easily randomly assign incentive structures under a central planning regime as in a market economy. But in practice, proponents of randomization are part of the current atheoretical approach to economics dominant in much of the United States in which the behavioural assumptions required for modelling are abandoned in favour of empiricism (Harrison, 2014). This point of view necessarily supports development through nudges rather than big pushes, as the latter requires more behavioural assumptions, assumptions that are embodied in theory. This debate is reflected clearly in Easterly's (2007) critique of planners—the big push of Sachs' Millennium Villages being one of his main targets—compared to seekers who favour small scale innovation and experimentation.
> I think the best we can say is that the jury is still out on this one. I am sympathetic to the idea that structural transformation requires deeper-seated changes. However, challenged that the Indian public health system is broken so what is

> needed is systemic reform rather than giving away plates as an incentive to parents to bring their children to be immunized, Esther Duflo replied that it would take years to achieve such reform, so what is the harm in giving out some plates to get children immunized now.
> I do see her point, but in the end, what policy makers and programme makers need to know is if particular programmes works.

This is all very well put. But we need to be careful not to presume that the opposite is true: that 'nudges' can be supported by atheoretical insights. White (2014) does not say this, but it might easily be inferred from what he writes. Again, I believe the core problem is the use of the slogan 'what works', as suggested at the end of the above extract. That one can, indeed, just focus on 'what works' is far from evident, although many people take it for granted. Take the question of take-up of insurance again: it is far from obvious, without knowing why people take up a product, that this is a good or bad thing for them.

This is one reason I take sharp, personal aim at many specific claims of this 'I only care about what works' ilk in Harrison (2011*b*). If one is passionate about policy goals, even goals on a small scale, we cannot be casual and agnostic about the proper use of theory and econometrics because we are looking at local nudges. This simply does not follow from the excellent point that White (2014) makes, that global, non-nudges do require some theism, and, indeed, huge swaths of it.

Another neglected dimension to the policy evaluation of field experiments and nudges is longitudinal. Do we see effects persist after the academic paper touting them has been published, and has garnered numerous citations? The sorry fate of Sachs' Millennium Villages field experiment should be a clarion call to policy-makers to follow the old motto *festina lente* (hasten slowly): see Munk (2013). Once we evaluate field experiments longitudinally, we have to take sample attrition seriously as an informative outcome, and that is rarely done (Harrison and Lau, 2014).

VII. Conclusions

Specialization is only valuable when it leads to gains from actual trade. Economists have become a very specialized lot, even if we just compartmentalize into theory, experiments, and econometrics for present purposes. And each of these compartments can be comfortable cells to work within, brim with elegance and beauty within their own four walls. But the risk of intellectual Balkanization is evident, most clearly in the manner in which advocates of field experiments eschew theory or structural econometrics.[8] The pity is that one can simply do much better economics, and answer the main questions about policy and causality, by embracing both.

I will only believe claims about field behaviour when I see then in the lab, a seemingly radical position for an advocate of field experiments to take. Of course we are all interested in external validity, even if the concept has some hairy threads to it that we had

[8] This is not just limited to advocates of randomized evaluation. My favourite example, from a book on theory that I admire greatly in many ways, is from Wakker (2010, p. 267): 'Statistics textbooks do indeed recommend using more than eight observations to fit four parameters.' Yup.

better not tug at. But when we see academics at the top of our profession drowning in 6 inches of behavioural water with statements such as 'the lesson here is that economists have to think more about what households know and what households think', then we have to retire to what we know how to do in the lab. Not forever, and with some red-blooded field-based questions to address as best we can when we actually have a chance of making a knowledge claim of methodological substance.

The concept of 'the lab' means several things. To some it just means experiments conducted in the traditional manner in experimental economics, with convenience samples of university students facing instructions that are gutted of field referents so as to pose the abstract choice task as cleanly as possible (Harrison and List, 2004). To others, it means experiments conducted in the field with artefactual choice tasks (Viceisza, 2012). What I have in mind, as a complement to the usual field experiment, is just that we have experiments that control some of the things that the full-blown experiment does not, or cannot, control. This was the logic behind the long list of factors that Harrison and List (2004) used to define the varieties of field experiments: there is no single bright line differentiating the field from the lab. For instance, one can use 'virtual reality' representations of choices to mimic many of the field referents that subjects face, but in the context of a controlled laboratory setting (Fiore *et al.*, 2009; Harrison *et al.*, 2011; Dixit *et al.*, 2014). I really do not care if these are called lab or field experiments as long as we see the importance of varying the degree of control on stimuli to better understand behaviour in terms of coherent theoretical concepts (Harrison, 2005).

These are exciting times in economics, with more attention, more dollars, and some of the brightest minds being devoted to real problems of poverty and public policy than we have seen in a long time. But the gratuitous methodological precepts that seem to come along with the use of field experiments remind one more of the quaint trappings of the Catholic Church, Freemasonry, and Basel III than we should be comfortable with as scientists. The marketing claims of advocates of randomized control are easy to identify, but they come cloaked in seductive phrases such as 'evidence-based economics' or only being interested in 'what works'. How can one stand against such things?[9] Well, rather easily if you see what they entail in terms of avoiding many of the tough problems of economics.

References

Andersen, S., Fountain, J., Harrison, G. W., and Rutström, E. E. (2014*c*), 'Estimating Subjective Probabilities', *Journal of Risk and Uncertainty*, **48**, 207–29.

— Harrison, G. W., Lau, M. I., and Rutström, E. E. (2008), 'Eliciting Risk and Time Preferences', *Econometrica*, **76**(3), 583–619.

— — — — (2014*a*), 'Discounting Behavior: A Reconsideration', *European Economic Review*, **71**, 15–33.

— — — — (2014*b*), 'Dual Criteria Decisions', *Journal of Economic Psychology*, **41**, 101–13.

Banerjee, A. V., and Duflo, E. (2009), 'The Experimental Approach to Development Economics', *Annual Review of Economics*, **1**, 151–78.

[9] As Groucho Marx once said, 'The secret of life is honesty and fair dealing. If you can fake that, you've got it made.' Harrison (2011*b*) critically reviews the marketing claims made in Banerjee and Duflo (2011) and Karlan and Appel (2011).

Banerjee, A. V., and Duflo, E. (2011), *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, New York, Public Affairs.
Berry, D. A. (2005), 'Introduction to Bayesian Methods III: Use and Interpretation of Bayesian Tools in Design and Analysis', *Clinical Trials*, **2**(4), 295–300.
Bohm, P. (2003), 'Experimental Evaluations of Policy Instruments', in K. G. Mäler and J. R. Vincent (eds), *Handbook of Environmental Economics*, Vol. **1**, Amsterdam, Elsevier, 438–60.
Campbell, G. (2005), 'The Experience in the FDA's Center for Devices and Radiological Health with Bayesian Strategies', *Clinical Trials*, **2**(4), 359–63.
Collins, D., Morduch, J., Rutherford, S., and Ruthven, O. (2009), *Portfolios of the Poor: How the World's Poor Live on $2 a Day*, Princeton, NJ, Princeton University Press.
Crawford, V. P., and Meng, J. (2011), 'New York City Cab Drivers' Labor Supply Revisited: Reference-dependent Preferences with Rational-expectations Targets for Hours and Income', *American Economic Review*, **101**, 1912–32.
Deaton, A. (2010), 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, **48**(2), 424–55.
Deci, E. L. (1971), 'Effects of Externally Mediated Rewards on Intrinsic Motivation', *Journal of Personality and Social Psychology*, **18**(1), 105–11.
DellaVigna, S., List, J. A., and Malmendier, U. (2012), 'Testing for Altruism and Social Pressure in Charitable Giving', *Quarterly Journal of Economics*, **127**(1), 1–56.
Dixit, V., Harrison, G. W., and Rutström, E. E. (2014), 'Estimating the Subjective Risks of Driving Simulator Accidents', *Accident Analysis and Prevention*, **62**, 63–78.
Easterly, W. (2007), *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*, New York, Oxford University Press.
Evans, C. H., and Ildstad, S. T. (eds) (2001), *Small Clinical Trials: Issues and Challenges*, Washington, DC, National Academy Press.
Farber, H. S. (2008), 'Reference-dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers', *American Economic Review*, **98**(3), 1069–82.
Ferber, R., and Hirsch, W. Z. (1978), 'Social Experimentation and Economic Policy: A Survey', *Journal of Economic Literature*, **16**(4), 1379–414.
— — (1981), *Social Experimentation and Economic Policy*, New York, Cambridge University Press.
Fiore, S. M., Harrison, G, W., Hughes, C. E., and Rutström, E. E. (2009), 'Virtual Experiments and Environmental Policy', *Journal of Environmental Economics and Management*, **57**(1), 65–86.
Food and Drug Administration (2010*a*), *Adaptive Design Clinical Trials for Drugs and Biologics*, Washington, DC, US Department of Health and Human Services, February.
— (2010*b*), *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*, Washington, DC, US Department of Health and Human Services, February.
Garfinkel, I., Manski, C., and Michalopoulos, C. (1992), 'Micro Experiments and Macro Effects', in C. Manski and I. Garfinkel (eds), *Evaluating Welfare and Training Programs*, Cambridge, MA, Harvard University Press.
Gill, D., and Prowse, V. (2012), 'A Structural Analysis of Disappointment Aversion in a Real Effort Competition', *American Economic Review*, **102**(1), 469–503.
Goodman, S. N. (2005), 'Introduction to Bayesian Methods I: Measuring the Strength of Evidence', *Clinical Trials*, **2**(4), 282–90.
Harrison, G. W. (2005), 'Field Experiments and Control', in J. Carpenter, G. W. Harrison, and J. A. List (eds), *Field Experiments in Economics*, Greenwich, CT, JAI Press, Research in Experimental Economics, Vol. 10, 17–50.
— (2011*a*), 'Experimental Methods and the Welfare Evaluation of Policy Lotteries', *European Review of Agricultural Economics*, **38**(3), 335–60.
— (2011*b*), 'Randomisation and Its Discontents', *Journal of African Economies*, **20**(4), 626–52.
— (2013), 'Field Experiments and Methodological Intolerance', *Journal of Economic Methodology*, **20**(2), 103–17.
— (2014), 'Impact Evaluation and Welfare Evaluation', *European Journal of Development Research*, **26**(1), 39–45.

Harrison, G. W., and Lau, M. I. (2014), 'Risk Attitudes, Sample Selection and Attrition in a Longitudinal Field Experiment', Working Paper 2014-04, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.
— List, J. A. (2004), 'Field Experiments', *Journal of Economic Literature*, **42**(4), 1013–59.
— Haruvy, E., and Rutström, E. E. (2011), 'Remarks on Virtual World and Virtual Reality Experiments', *Southern Economic Journal*, **78**(1), 87–94.
— Lau, M. I., and Rutström, E. E. (2009), 'Risk Attitudes, Randomization to Treatment, and Self-selection into Experiments', *Journal of Economic Behavior and Organization*, **70**(3), 498–507.
Hausman, J. A., and Wise, D. A. (eds) (1985), *Social Experimentation*, Chicago, IL, University of Chicago Press.
Heckman, J. J. (1992), 'Randomization and Social Program Evaluation', in C. Manski and I. Garfinkel (eds), *Evaluating Welfare and Training Programs*, Cambridge, MA, Harvard University Press.
— (2010), 'Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy', *Journal of Economic Literature*, **48**(2), 356–98.
Jennison, C., and Turnbull, B. W. (2011), *Group Sequential and Adaptive Methods for Clinical Trials*, Boca Raton, FL, Chapman & Hall.
Karlan, D., and Appel, J. (2011), *More Than Good Intentions: How a New Economics is Helping to Solve Global Poverty*, New York, Dutton.
Keane, M. P. (2010*a*), 'Structural vs Atheoretic Approaches to Econometrics', *Journal of Econometrics*, **156**, 3–20.
— (2010*b*), 'A Structural Perspective on the Experimentalist School', *Journal of Economic Perspectives*, **24**(2), 47–58.
Leamer, E. E. (2010), 'Tantalus on the Road to Asymptopia', *Journal of Economic Perspectives*, **24**(2), 31–46.
Louis, T. A. (2005), 'Introduction to Bayesian Methods II: Fundamental Concepts', *Clinical Trials*, **2**(4), 291–4.
Manski, C., and Garfinkel, I. (eds) (1992), *Evaluating Welfare and Training Programs*, Cambridge, MA, Harvard University Press.
Morgan, S. L., and Winship, C. (2007), *Counterfactual and Causal Inference: Methods and Principles for Social Research*, New York, Cambridge University Press.
Munk, N. (2013), *The Idealist: Jeffrey Sachs and the Quest to End Poverty*, New York, Doubleday.
Rosenzweig, M. R., and Wolpin, K. I. (2000), 'Natural "Natural Experiments" in Economics', *Journal of Economic Literature*, **38**, 827–74.
Ross, D. (2014), *Philosophy of Economics*, Basingstoke, Palgrave Macmillan.
Sunstein, C. (2013), *Simpler: The Future of Government*, New York, Simon & Schuster.
Thaler, R., and Sunstein, C. (2008), *Nudge: Improving Decisions About Health, Wealth and Happiness*, New Haven, CT, Yale University Press.
Toubia, O., Johnson, E., Evgeniou, T., and Delquié, P. (2013), 'Dynamic Experiments for Estimating Preferences: An Adaptive Method for Eliciting Time and Risk Parameters', *Management Science*, **59**(3), 613–40.
Viceisza, A. C. G. (2012), *Treating the Field as a Lab: A Basic Guide to Conducting Economics Experiments for Policymaking*, Washington, DC, International Food Policy Research Institute.
Wakker, P. P. (2010), *Prospect Theory for Risk and Ambiguity*, New York, Cambridge University Press.
Wang, S., Filiba, M., and Camerer, C. F. (2010), 'Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters', Working Paper, Division of Humanities and Social Sciences, California Institute of Technology.
White, H. (2014), 'Current Challenges in Impact Evaluation', *European Journal of Development Research*, **26**(1), 18–30.
Whitehead, J. (1997), *The Design and Analysis of Sequential Clinical Trials*, 2nd edn, New York, Wiley.
Worrall, J. (2007), 'Why There's No Cause to Randomize', *British Journal of the Philosophy of Science*, **58**, 451–88.