Journal of African Economies, Vol. 20, number 4, pp. 626–652
doi:10.1093/jae/ejr030

Randomisation and Its Discontents

Glenn W. Harrison[a,b,*]

[a]Department of Risk Management and Insurance, Georgia State University, Atlanta, GA, USA
[b]Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, Atlanta, GA, USA

* Corresponding author: Glenn W. Harrison. E-mail: gharrison@gsu.edu

Abstract

Randomised control trials have become popular tools in development economics. The key idea is to exploit deliberate or naturally occurring randomisation of treatments in order to make causal inferences about 'what works' to promote some development objective. The expression 'what works' is crucial: the emphasis is on evidence-based conclusions that will have immediate policy use. No room for good intentions, wishful thinking, ideological biases, Washington Consensus, cost–benefit calculations or even parametric stochastic assumptions. A valuable byproduct has been the identification of questions that other methods might answer, or that subsequent randomised evaluations might address. An unattractive byproduct has been the dumbing down of econometric practice, the omission of any cost–benefit analytics and an arrogance towards other methodologies. Fortunately, the latter are gratuitous, and the former point towards important complementarities in methods to help address knotty, substantive issues in development economics.

JEL classification: C93, D03, O12, B41

The Randomised Control Trial (RCT) has become a popular tool in development economics. The key idea is to exploit deliberate or naturally occurring randomisation of treatments in order to make causal inferences about 'what works' to promote some development objective. The expression 'what works' will be dissected and defanged, but the emphasis is on evidence-based

conclusions that will have immediate policy use. No room for good intentions, wishful thinking, ideological biases, Washington Consensus, cost–benefit calculations or even parametric stochastic assumptions. This is *Dragnet* stuff: just the facts.[1] A valuable byproduct has been the identification of questions that other methods might answer, or that subsequent RCTs might address. An unattractive byproduct has been the dumbing down of econometric practice, the omission of any cost–benefit analytics, and an arrogance towards other methodologies. Fortunately, the latter are gratuitous, and the former point towards important complementarities in methods to help address knotty, substantive issues in development economics.

Section 1 reviews the range of experimental procedures available beyond the RCT, the history of randomisation and the RCT, some of the statistical baggage that comes with doing an RCT, but that has no essential role, and the self-appointed status of the RCT as the King of the Evidential Hill. Section 2 untangles the concept of 'what works,' and why randomised evaluations often do *not* work in interesting ways for development policy. Section 3 reviews two book-length 'infommercials' for RCTs in the context of development economics, deliberately taking the cold-hearted *Dragnet* approach and taking their words literally.

Section 4 concludes with what will strike some as a radical proposition: that claims about human behaviour will only be worth believing when they have been replicated in the lab. The reason is simple, and indeed not controversial in other experimental sciences. The lab provides the controls necessary to avoid endless epistemological finger-wagging dismissals of alternative explanations. As one of the strongest advocates of the methodological role of field experiments in economics, in Harrison and List (2004), this might seem to be a contrarian statement (even for me). But it just reminds us of the complementarity of methods in science for different inferential objectives, a point forgotten in much of the current fads over certain experimental methods.

1. Experiments and randomisation

Experiments can help inform the evaluation of policies with uncertain impacts in two ways. The first is by providing some guidance as to latent

[1] The reference is to the oft-repeated command of the terse Sergeant Joe Friday from the TV programme, as he took notes from witnesses to a crime: 'Just the facts, ma'am.' In fact, these lines were never used in the TV programme, despite becoming widely associated with it. The actual lines were better: 'All we want are the facts, ma'am,' and 'All we know are the facts, ma'am.'

structural parameters needed to complete the welfare evaluation. The second is by bypassing the need for all of this structure, in an agnostic manner, and 'letting the data speak for itself' with minimal theoretical assumptions and a reliance on randomisation. We come back to the complementarity of these two uses of experiments later.

1.1 Types of experiments

It is worth identifying the various types of experiments in wide use, since some think of the word 'experiment' as only referring to something that involves randomisation. Harrison and List (2004) propose a taxonomy to help structure thinking about the many ways in which experiments differ. At one end of the spectrum are *thought experiments*, which can be viewed as the same as any other experiment but without the benefit of execution (Sorenson, 1992). Then there are conventional *laboratory experiments*, typically conducted with a convenience sample of college students and using abstract referents.[2] Then there are three types of field experiments. *Artefactual field experiments* are much like lab experiments, but conducted with subjects that are more representative of a field environment. *Framed field experiments* extend the design to include some field referent, in terms of the commodity, task, or context. *Natural field experiments* occur without the subject knowing that they have been in an experiment. Then we have *social experiments*, where a government agency deliberately sets out to randomise some treatment. Finally, there are *natural experiments*, where some randomisation occurs without it being planned as such: serendipity observed.

These categories were never intended to be hard and fast, and indeed the reason for suggesting them was to make clear the criteria defining a flexible continuum of experiments, rather than propose a single bright line to define what one means by a field experiment. One can easily imagine intermediate categories. One important example is the notion of a *virtual experiments* due to Fiore *et al.* (2009), with the potential of generating both the internal validity of lab experiments and the external validity of field experiments.

Randomisation can be used in every one of these types, and is more a method of conducting experiments rather than a defining characteristic of any one type of experiment in the field. For example, if one presents subjects with a series of constructed lottery choices, in order to measure their risk preferences or their subjective beliefs, there is no randomisation

[2] A referent is an object or idea to which a word or phrase refers.

involved. It seems incoherent, and inconsistent with usage in the field of experimental economics, to *not* view that as an experiment.

1.2 Randomised evaluations in development economics

Randomised evaluations in development economics involve the deliberate use of a randomising device to assign subjects to treatment, or the exploitation of naturally occurring randomising devices. Good reviews of the methodology are contained in Duflo and Kremer (2005), Duflo (2006), Duflo *et al.* (2007) and Banerjee and Duflo (2009). Complementary econometric strategies are described in Angrist and Pischke (2009).

One of the claimed advantages of randomisation is that the evaluation of policies can be 'hands off,' in the sense that there is less need for maintained structural assumptions from economic theory or econometrics. In many respects this is true, and randomisation does indeed deliver, on a good, asymptotic randomising day, orthogonal instruments to measure the effect of treatment. This has been well known for a long time in statistics, and of course in the economics experiments conducted in laboratories for decades. But it is apparent that the case for randomisation has been dramatically oversold: even if the original statements of the case have the right nuances, the second generation of practitioners seem to gloss those. Words such as 'evidence based' or 'assumption free' are just marketing slogans, and should be discarded as such. Excellent critiques by Rosenzweig and Wolpin (2000), Deaton (2010), Heckman (2010), Keane (2010a, b), Leamer (2010), and spirited defences by Imbens (2010), cover most of the ground in terms of the statistical issues.

One side-effect of the popularity of RCT is the increasing use of Ordinary Least Squares estimators when dependant variables are binary, count or otherwise truncated in some manner.[3] One is tempted to call this the *OLS Gone Wild* reality show, akin to the *Girls Gone Wild* reality TV show, but it is much more sober and demeaning stuff. I have long given up asking researchers in seminars why they do not just report the marginal effects for the right econometric specification. Instead I ask if we should just sack those faculty in the room who seem to waste our time teaching things like logit, count models or hurdle models. I have

[3] Increasing, but not exclusive. For example, Bertrand *et al.* (2010) report marginal effects from a probit model of binary choice. On the other hand, they also report OLS regressions for non-negative, continuous dependant variables such as loan amount. A hurdle model incorporating the effects of both would be the correct specification to capture the net effect of their RCT.

also volunteered that if they ever receive a referee report telling them to estimate and report the right econometric model that they can freely assume I wrote it.

1.3 The origin of randomised control trials

Where did the notion of an RCT start? Fisher (1926) is widely acknowledged as the 'father' of randomisation, and indeed he did the most to systematically develop the methods. But the concept of an RCT actually originated in one of the classic debates of psychometrics: a critique by Peirce and Jastrow (1885) of the famous experiments of Fechner on subjective perceptions of differences in sensation.[4] Fechner had used his own observations of sensations to test his own theories about minimally perceptible differences, much like Fisher's famous tea-drinking lady used cups of tea that she had prepared herself to form her opinions about the effect of having milk included before or after the tea. The risk, of course, is that the experimenter might, deliberately or not, behave in a way that supports his pet theory. Randomisation cures that problem of incentive compatibility in data generation.

Salsburg (2001) contains lively discussions of the famous tea-drinking anecdote, and the tensions between surrounding personalities. Hacking (1988) contains a discussion of the exotic contexts, such as the debunking of telepaths and other psychics, that led to the rise of randomisation as a popular scientific method. Kadane and Seidenfeld (1990, p. 335) note that this concern arises solely because of concerns that the subject is responding strategically to the treatment: 'when there can be no game between researcher and subject, e.g., when the responses are involuntary (because the subjects are plots of land with no interest in the outcomes), this problem doesn't even arise.'

This is also the perspective that Savage (1962, pp. 34, 87, 89) took on the role of randomisation, that it helped two parties communicate. Stone (1969) elaborated on this argument, imagining an inferential dialogue between two Bayesians. The listener wants to use his own priors and the speaker's data to draw his own inferences, and he argues that random sampling is the unique way to do that. Kadane and Seidenfeld (1990, p. 341) disagree, noting that the choice of sample size, for example, presumes something about the priors and loss functions of the speaker.

[4] Regression discontinuity designs originated in psychology as well: see Thistlethwaite and Campbell (1960). Lee and Lemieux (2010) review their many applications in economics.

Indeed, this is related to the vexing issue of heterogeneity of response, much discussed, and often glossed, in the modern RCT literature in development economics. On the other hand, Kadane and Seidenfeld (1990, p. 343) note that as long as the assumptions underlying the choice of sample size are made explicit, in the form of some power analysis, randomisation still allows the listener to draw inferences using his own priors and loss functions. They also stress that it is not the only such method; nor are power analyses popular among RCT applications in development economics.

1.4 The gold standard claim

We often hear that an RCT is the gold standard in medicine, and that this should be what we unwashed social scientists should aspire to. Such claims get repeated without comment, but, to quote a popular political refrain in the USA, advocates of RCTs are entitled to their own opinions but not their own facts.

Two careful studies showed that the alleged differences between an RCT and an observational study were not in fact present. Benson and Hartz (2000, p. 1878) '...found little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or qualitatively different from those obtained in randomized, controlled trials.' Similarly, Concato *et al.* (2000, p. 1887) conclude that the '...results of well-designed observational studies (with either a cohort or a case–control design) do not systematically overestimate the magnitude of the effects of treatment as compared with those in randomized, controlled trials on the same topic.' This does not say one should not use an RCT, just that it should be used when cost-effective compared with other methods, which are often cheaper and quicker to implement.[5]

Timing is an issue that deserves more discussion. It is often difficult to design a careful RCT 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. 455–459) 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

[5] Prior to the popularity of RCTs, in many areas of empirical economics the typical discussion centered on the ability of weak instruments to be able to infer causality: see Rosenzweig and Wolpin (2000), Angrist and Krueger (2001), Stock *et al.* (2002) and Murray (2006). This discussion is avoided by using an RCT, although issues remain about the interpretation of causality, buried in the 'intent to treat' Sicilian defence.

took years to undertake the requisite RCT needed for the procedure to become widely recommended and used. Lives were lost because of the stubborn insistence on RCT evidence before the procedure could be widely adopted. Of course, counter-examples probably exist, but the costs and benefits of having complementary evaluation methodologies are often lost in the push to advocate one over the other, as illustrated by the contrast between Gosset and Fisher (Ziliak, 2008).

2. Showing 'what works' with randomised evaluations?

Turning to the recent wave of applications of randomisation in economics, several concerns have been raised. Experiments are conducted to make inferences, and different types of inferences can call for different types of experiments. To take three types of inference of concern here, one might be interested in evaluating the welfare effects of a treatment for a cost–benefit analysis, one might be interested in understanding behaviour in order to design normative policies, or one might be interested in estimating the (average) effects of a policy (on observables). The last of these is not usually the most important of the three.

2.1 Evaluating welfare effects

One can certainly be interested in worms and whatever they do, absentee teachers and whatever they do not do, the optimal use of fertiliser, wherever it comes from, 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. How do I decide if it is better to reduce worms, increase teacher presence, use fertiliser better or increase savings rates, if I do not know the welfare impact of these policies? Of course, they might be 'costless' to implement, but that is rare.

Related to this concern, there is an important debate over the effects of charging for access to interventions. Kremer and Holla (2009) review the evidence from many RCTs in health and education that suggest that individuals and households do not seem willing to pay for interventions that generate what *seem to be* significant benefits to them at what *seem to be* significant costs. There appears to be a 'jump discontinuity' in willingness to pay that is disconcerting. At first, and second, blush this seems to be a clear revealed preference argument that the welfare benefits of the intervention

are not what the researcher assumes them to be.[6] And it leaves analysts scrambling for behavioural explanations without any empirical basis. After hand-waving about *a priori* plausible behavioural explanations, Weil (2009, p. 121ff) has nothing better to conclude[7] from these RCT studies than that 'the lesson here is that economists have to think more about what households know and what households think.' Is that really the best we can do?

The issue is subtle, however, as Kremer and Holla (2009) stress.[8] Payment can change the nature of the intervention in qualitative ways, even for tiny amounts of money. An old example, from the father of field experiments, Peter Bohm, illustrates this well.[9] In 1980 he undertook a field experiment for a local government in Stockholm that was considering expanding a bus route to a major hospital and a factory. The experiment was to elicit valuations from people who were naturally affected by this route, and to test whether their aggregate contributions would make it worthwhile to provide the service. A key feature of the experiment was that the subjects would have to be willing to pay for the public good if it was to be provided for a trial period of 6 months. Everyone who was likely to contribute was given information on the experiment, but when it came time for the experiment virtually nobody turned up! The reason was that the local trade unions had decided to boycott the experiment, since it represented a threat to the current way in which such services were provided. The union leaders expressed their concerns, summarised by Bohm (1984, p. 136) as follows:

> They reported that they had held meetings of their own and had decided (1) that they did not accept the local government's decision not to provide them with regular bus service on regular terms; (2) that they did not accept the idea of having to pay in a way that differs from the way that "everybody else" pays (bus service is subsidized in the area) – the implication being that they would rather go without this bus service, even if their members felt it would be worth the costs; (3) that they

[6] Some advocates of RCTs openly refer to this as the 'moronic revealed preference argument' in conversation. It is true that there is some subtlety to the issue, as we discuss, but the revealed preference argument is only 'moronic' in the sense that it can be understood by anyone.

[7] To be fair, he did not conduct the study in question, and was just trying to make sense of it.

[8] A balanced statement of pricing implications beyond the immediate RCT is provided by Ashraf *et al.* (2010, section VI).

[9] Dufwenberg and Harrison (2008) provide an appreciation of the methodological significance of Bohm's pioneering work.

> would not like to help in realizing an arrangement that might reduce the level of public services provided free or at low costs. It was argued that such an arrangement, if accepted here, could spread to other parts of the public sector; and (4) on these grounds, they advised their union members to abstain from participating in the project.

This fascinating outcome is actually more relevant for experimental economics in general than it might seem. When certain institutions are imposed on subjects, and certain outcomes tabulated, it does not follow that the outcomes of interest for the experimenter are the ones that are of interest to the subject. And, most critically, running field experiments forces one to be aware of the manner in which subjects select themselves into tasks and institutions based on their beliefs about the outcomes.

This process might be a direct social choice over institutions or rules, it might be Tiebout-like migration, it might be a literal or behavioural rejection of the task, it might be literal or behavioural attrition once the task is understood, it might be the evolution of social norms to resolve implicit coordination problems or it might be some combination of these. This is an active and exciting area of research in laboratory experiments now, and one that draws on insights from field experiments such as those conducted by Bohm (1984).

The point is that we design better lab experiments when we worry about what one just cannot ignore in the field experiment, and those lab experiments in turn inform our inferences about the field experiment. Until someone can do better than saying that 'economists have to think more about what households know and what households think' when there is some important behavioural puzzle in the field, I will give up the external validity of the field instantly and head back to the lab for a spell.

2.2 Designing normative policies

If we are to design normative policies, and understand the opportunity cost of doing so, we need to understand *why* we see certain behaviour. The apparent jump discontinuity in willingness to pay discussed above should send chills through those casually sliding from alleged 'cost effectiveness' to a recommendation that scarce resources be allocated to any project. Weil (2009) illustrates what happens when we have no complementary information on preferences or beliefs to guide our thinking. For example, consider an RCT for bed nets to prevent malaria that showed take-up rates of 40%, 'even when the subsidized price is sixty

cents for a bed net that lasts five years and prevents a certain number of episodes of illness or possibly death of a child' (Weil, 2009, p. 121). Is 40% low? Who knows? Kremer and Holla (2009) and Weil (2009) think so. But here is the extent of the understanding of the issue:

> Of course, any behavior can be rationalized by some combination of discount rates, value placed on child health, and so on. But it is extremely hard to do so in this case. (Weil, 2009, p. 121).

How do we know it is hard to do so? Did someone ask the respondents what their time preferences were, what their subjective beliefs were, what their conditional willingness to pay for an avoided illness or even death was?

> When Kremer and Holla try to think of behavioral models with some kind of procrastination going on, I become less sympathetic to their argument, partly because of the very unusual things going on here. Would the typical persons in the subject population exhibit a lot of procrastination in other aspects of life? (. . .) Or is this procrastination manifested only in the types of situations explored in these studies? (Weil, 2009, p. 122).

Can't we fill these massive rhetorical holes with data?

> If it is the latter, that points to some other sources of the behavior, a prime candidate being some sort of information problem. That is, when I do the calculation, it is clear to me that the typical subjects in a trial should be buying this bed net for sixty cents. But maybe I have a different information structure than these persons do. Maybe they do not believe the net lasts five years, or that it works at all, or that mosquitoes cause malaria, or something like that. (. . .) Somehow these informational problems are getting tied up with the behavioral response. So I am not ready to look at the full panoply of behavioral models to rationalize this behavior. (Weil, 2009, p. 122).

Huh? Subjective beliefs are not behavioral any more? And we have to wonder rhetorically about these key ingredients into the individual valuation of the mosquito-net-purchase lottery?

The frustration with this open-ended thinking comes from the knowledge that we have had the tools for a long time to answer these questions, in some measure. This type of *ex post* 'analysis' is like a doing brain surgery with a divining rod. Or, to quote Smith (1982, p. 929), 'Over twenty-five years ago, Guy Orcutt characterized the econometrician as being in the same predicament as that of an electrical engineer who has been charged with the task of deducing the laws of electricity by listening to the radio play.'

Again, the antidote for this empty rhetoric about what might account for the low willingness to pay is to complement the next RCT with some

field experiments to identify the 'moving behavioral parts' in these decisions, or to just head back to the lab and sort it out there.

The manner in which this 'sell them cheap or give them away' debate has been carefully framed by proponents of RCTs tells us a lot about the dangers of thinking about the welfare effects in a partial manner. Here is how Karlan and Appel (2011, p. 246) put it:

> Since prices didn't change the kinds of people who got nets or the way they were used, the difference between cost-sharing and free distribution could be easily summarized: a lot fewer people ended up protected, and the providers of the nets saved some money. Unfortunately, they weren't saving much. Each net cost about six dollars to produce, so when Population Services International sold nets to Kenyans for seventy-fixe cents, according to their prevailing policy, they were already bearing the vast majority of the cost. Covering the last seventy-five cents would have increased their cost per net by about 13 percent, but then that could have served four times the people!

The problem with this glib punchline is perhaps obvious, but the obvious apparently needs saying. What about the opportunity cost of the $6 shelled out every time one of these nets is given away? The relevant opportunity cost is not $0.75, unless we are in an absurd second best world where the donor only cares about how many nets it distributes.[10] Principles of economics, please.

When we do get a 'cost effectiveness analysis' of this issue, from Cohen and Dupas (2010, section V), it only examines the effectiveness of this programme when subsidies range from 90 to 100% of the cost, and there are some different assumptions about the mortality effects of the nets.

2.3 Evaluating intra-distributional effects

Figure 1 illustrates why we should not be lashing our inferential might to the mast of 'the average effect.' Each panel shows the distributional impact, compared with baseline, of a policy intervention in terms of some normalised income measure. The top panel shows an average effect which is larger than the bottom panel, and would be the preferred 'evidence based' policy if one were to focus solely on average effects. But it has a larger standard deviation, so there are plausible levels of risk aversion that would suggest

[10] In fact, it does have a much broader mission. The web site www.psi.org states it: 'To measurably improve the health of the poor and vulnerable people in the developing world, principally through the social marketing of health products, services and communications.' I am guessing there is more to this mission than bed nets *per se*.

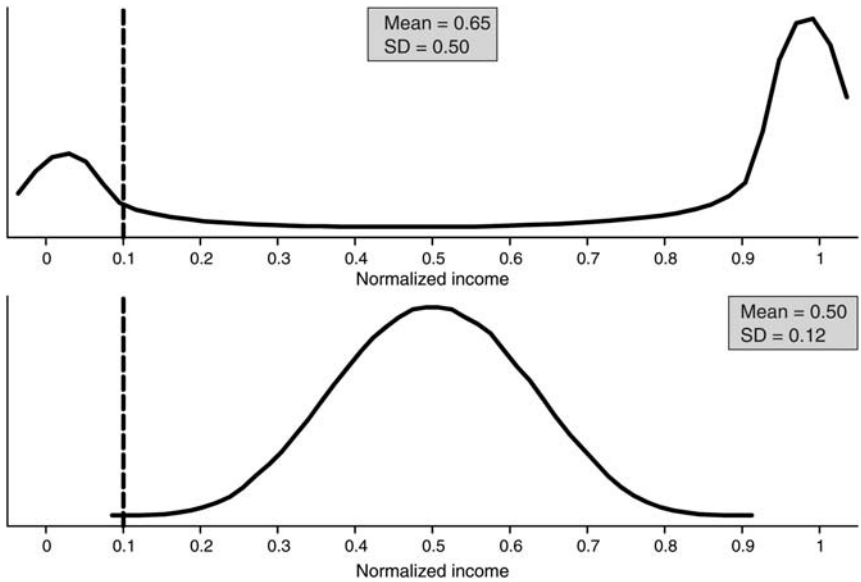

Figure 1: Why Average Effects Are Not Everything.

that the policy lottery with the highest average return is *not* the best one in certainty-equivalent welfare terms. Moreover, what if the welfare impact of income levels was not uniform, such that any income level below the value of 0.1 entailed relatively high costs? Let this be an absolute poverty line, below which there is some asymmetric physiological deterioration. Then any policy that increased the chance of this outcome, even with the promise of a better income on average and even if the affected agents were risk neutral, might be a disaster. A pity, but we cannot avoid worrying about the whole distribution if we are to do a proper welfare analysis.

Of course, once one raises issues about intra-distributional effects, we can hear the Randomistas cursing those pesky unobservables, since they generate all manner of problems. Actually, they are probably just cursing Heckman (2010), or even just cursing heterogeneity itself! Anyone that does not appreciate the significance of the concern with heterogeneity should work through the arithmetic of the 'Vietnam Draft example' in Keane (2010a, p. 5), and see how unreliable Wald estimators can quickly become.[11]

[11] The Wald (1940) estimator was developed for the purpose of estimating an unconditional, bivariate regression when both the regressor and the regressand are subject to

The problem of randomisation bias, and the way in which it allows unobservables to affect inference, is well known. For example, when experimenters recruit subjects they offer them, in effect, a lottery of earnings, offset by a fixed show-up fee. By varying the show-up fee between subjects, and measuring the risk attitudes of these that show up, one can directly demonstrate the effect of randomisation bias from this recruitment procedure (e.g. Harrison *et al.*, 2009). Turning to the RCT setting, it is well known in the field of clinical drug trials that persuading patients to participate in randomised studies is much harder than persuading them to participate in non-randomised studies (e.g. Kramer and Shapiro, 1984). The same problem applies to social experiments, as evidenced by the difficulties that can be encountered when recruiting decentralised bureaucracies to administer the random treatment (e.g. Hotz, 1992). Heckman and Robb (1985) note that the refusal rate in one randomised job-training programme was over 90%, with many of the refusals citing ethical concerns with administering a random treatment.

But apart from the statistical issues, which are bad enough, there is an important reason for wanting to keep track of the intra-distributional effects: we care a lot about 'winners' and 'losers' from policy. No policy maker can afford to ignore these equity effects, and if it is at all possible to come up with policy alternatives that mitigate losses, that is usually extremely attractive. At the very least one would like to be able to identify those individuals, and then one would like to be able to simulate policies that can mitigate losses. The simulation technology for this

measurement error. It essentially splits the sample using some statistic that does not depend on the standard deviation of the errors, and generates point estimates of the intercept and single covariate coefficient. The measurement errors are not assumed to be Gaussian. But they are assumed to be serially uncorrelated, not correlated with each other, homoskedastic and to have finite variance. Sample selection processes could, of course, wreak havoc with the first two assumptions. Wald (1940, p. 298) explained clearly that one should not use his method for prediction, and that one should use OLS for that. In other words, and to use his notation, if X and Y are the observed data, and one has used his estimator to infer that $Y = a + bX$, it does not follow that $y = a + bx$ is an unbiased estimator of y when a new, out-of-sample value of x is observed or evaluated, as in a counter-factual policy exercise. In general, the consistency of the Wald estimator is extremely fragile. Randomistas who insist on using it should review Neyman and Scott (1951) and Pakes (1982), which are completely absent from textbooks on the Wald estimator and its use in RCTs, such as Angrist and Pischke (2009).

'policy reform without tears' exercise is well known in trade policy evaluations, but of course requires some sort of structural insight into behaviour (e.g. Harrison *et al.*, 2002, 2003, 2004).

3. Marketing

It is an unfortunate development in recent decades that several popular areas of economics have been oversold. Behavioural economics is the first, to the point where some now call for a 'behavioural counter-revolution' by experimental economics to reclaim the field (Harrison, 2010). Neuroeconomics is the second, with many outrageous claims by its early proponents, and embarrassing empirical methods (Harrison, 2008a, b). Randomised evaluation in development economics is just the latest.

Two recent books state the claim for randomised evaluation methods in development.[12] Each offers remarkable insight into the manner in which substantive insight can be knowingly distorted by academic entrepreneurs. One needs to be blunt, and quote at length, since so many of the best and brightest 'Padawan Learners' in the profession seem to be blinded by the light. Step *away* from the light...

3.1 Poor Economics is, in the end, just good economics of the poor

Banerjee and Duflo (2011, p. 3) argue that *good* development policies should not wait for *perfect* development policies, and that one should

> start to think of the challenge as a set of concrete problems that, once properly identified and understood, can be solved one at a time. Unfortunately, this is not how the debates on poverty are usually framed. Instead of discussing how best to fight diarrhea or dengue, many of the most vocal experts tend to be fixated on the 'big questions': What is the ultimate cause of poverty? How much faith should we place in free markets? Is democracy good for the poor? Does foreign aid have a role to play? And so on.

In fact, the fixation here is on the observation that Jeffrey Sachs and William Easterly disagree with each other! Seriously, that is it. That is the basis for claiming that there is a need for some 'radical rethinking'

[12] In another book, Banerjee (2007) collects contributions from a lively debate on *Making Aid Work*. He argues (p. 7) the case for evidence-based RCTs as the basis for giving aid to poor countries, as an antidote for the 'lazy thinking' found in every alternative method, and the World Bank in particular. The commentators react sharply, and well enough.

of the economics of the poor. The illogic is stunning. Instead of 'sweeping answers' they argue that the randomisation approach they favour 'will not tell you whether aid is good or bad, but it will say whether particular instances of aid *did some good or not*.' (p. 4, emphasis added).

The striking thing about this book is that it tries to be radical and behavioural, but only succeeds in the rhetoric. The substantive analyses are actually very mainstream, and rich for it. It is as if the authors box themselves into an expository corner, where they want to stress how innovative and novel their methods are, but when it comes down to the nuts and bolts of explaining behaviour and drawing conclusions, they turn out to be simply good development economists.[13]

Unfortunately, they are not very good behavioural economists. The literature is read uncritically, and applied when convenient and without qualification. Take discounting behaviour, for example. We are told that

> research in psychology has now been applied to a range of economic phenomena to show that we think about the present very differently from the way that we think about the future (a notion referred to as "time inconsistency"). [...] the human brain processes the present and the future very differently. In essence, we seem to have a vision of how we should act in the future that is often inconsistent with the way we act today and will act in the future. One form that "time inconsistency" takes is to spend now, at the same time as we *plan* to save in the future. In other words, we hope that our "tomorrow's self" will be more patient than "today's self" is prepared to be. [...] A group of economists, psychologists, and neuroscientists worked together to establish that there is in fact a physical basis for such disjunction in decisionmaking. (pp. 64, 194, 195)

Are we asserting here that everyone is always time inconsistent in all settings? Or that the average person is? Or that we can pull this behavioural label out of the hat when it helps to explain behaviour? Is there no debate on these matters? There is, reviewed in Andersen *et al.* (2011a, b) and Harrison (2008a; pp. 316–319).

[13] Hence one would agree completely with Deaton (2010, p. 426) that the 'the analysis of projects needs to be refocused toward the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected work.' The jury is out on his next line: 'The best of the experimental work in development economics already does so because its practitioners are too talented to be bound by their own methodological prescriptions.'

3.2 More than good intentions, and less than good economics

Karlan and Appel (2011) claim that two insights allow one to figure out what works. The first (p. 7) is to get into the field and understand behaviour. In practice, this means casually applying labels derived from artificial lab behaviour to 'hard to explain' behaviour observed in the field. But it is a good motherhood statement to say that one looks at actual behaviour rather than predicted, rational behaviour, even if that means nothing.

The second insight, of course, is randomised evaluation (p. 8), since that 'lets us compare competing solutions [...] and see which one is most effective.' The key word here is 'effective,' and no welfare metric is offered for that: it is simply the most easily observed metric. They then claim that 'Creative and well-designed evaluations can go even further, and help us understand *why* one works better than another.' This is important, if true. In fact, it rarely is.

The motivation in Karlan and Appel (2011) for being driven to do something different is strikingly familiar to Banerjee and Duflo (2011): Sachs and Easterly do not agree. See for yourself:

> Sachs and his supporters regale us with picture-perfect transformational stories. Easterly and the other side counter with an equally steady supply of ghastly the-world-is-corrupt-and-everything-fails anecdotes. The result? Disagreement and uncertainty, which leads to stagnation and inertia – in short, a train wreck. And no way forward. [We] propose that there actually is a way forward.

Real Promised Land stuff. But notice the common, superficial rhetorical device. Two vocal academics disagree with each other on development policy, so we infer that policy makers are incapable of action. It is time to be blunt: this is extraordinarily arrogant and insulting stuff, even by the standards of modern economic proselytising.

We learn from Karlan and Appel (2011, p. 254) that economics has nothing to say about decisions to do with sex. Really! Here is how they put it:

> Sex is a primal activity. It's in our biology. In some sense, we are most definitely Humans – and most definitely not Econs – when we're doing the deed. In that space of urge, impulse, and heavy breathing, a lot fades into the background. [...] The throes of passion simply are not the best setting to run through a cost-benefit analysis

Huh? What kind of drivel is this? The terminology of Econs and Humans is used to differentiate the straw-man of traditional economic reasoning from the Noble Savage of the real world of behavioural economists. Lets get

serious about what these stereotypes consist of. The Econ is defined (p. 6) as someone who, when

> they need to choose between two alternatives, [. . .] weigh all the potential costs and benefits, compute an 'expected value' for each, and choose the one whose expected value is higher. In addition to keeping a cool head, they are very methodical and reliable calculators. Given accurate information about their options, they always choose the alternative most likely to give them their greatest overall satisfaction.

OK, time out. This fiction cannot be perpetuated any longer, or else we lose touch with basic principles of economics.[14] It is expected utility, not expected value. And their probabilities on the risky outcomes they face are generally subjective. Nothing in subjective expected utility says that these subjective probabilities should match those an actuary might use. And, the ultimate klunker, in the final breathe, we *define* their preferences by their choices. So it is tautology to say that the chosen alternative is the one that provides the highest expected subjective expected utility, since we define subjective probabilities and utilities so that this is the case. Is this nonsense actually being taught anywhere?

But it gets worse. Hard to imagine, but listen to this. We are told that 'Behavioral economics expands on narrow definitions of traditional economics in two important ways. The first is simple: Not everything that matters is dollars and cents.' Huh? You mean we need to think about the utility of money instead of money itself? Or that we need to think about the utility of outcomes such as health, sexual performance, family stability and the well-being of our loved ones, as well as money? Gee, traditional economics has always allowed that, when I last looked. Sometimes we do not model it, to be sure, but it is hardly a 'new economics' to argue that

[14] The terms come from the popular *Nudge* book of Thaler and Sunstein (2008, p. 6), who do equally poorly in defining stereotypes. They start with the even sillier term *homo economicus*, which is clear sign in an academic context that someone is engaged in polemics rather than science: 'Whether or not they have ever studied economics, many people seem at least implicitly committed to the idea of *homo economicus*, or economic man – the notion that each of us thinks and chooses unfailingly well, and thus fits within the textbook picture of human beings offered by economists. If you look at economics textbooks, you will learn that homo economicus can think like Albert Einstein, store as much memory as IBM's Big Blue, and exercise the willpower of Mahatma Gandhi. Really.' No, not really. It is sad to see Chicago professors not knowing what is actually being taught in good schools, what is actually in most textbooks, and apparently confusing the inevitably anodyne *Principles* mega-texts with what economists actually do and teach. Fortunately, the substance of their challenging discussion of behavioural institutional design does not rely on this caricature.

we should. This is like rejecting the whole of production theory because data shows that production functions are not Leontief.

But there is even more to the new economics. Instead of using a cost–benefit analysis,

> Sometimes we have different priorities. Other times we are distracted or impulsive. We sometimes slip up on the math. And, more often than we would like to admit, we are shockingly inconsistent. [. . .] Instead of deducing a way to think from a core set of principles, behavioral economics builds up a model of decision-making from observations from people's actions in the real world. As we will see throughout this book, this way of thinking can help us design better policies.

Now we have to start taking words seriously at some point. At the outset, 'different priorities' than what? The preferences that explain observed behaviour, in accord with the principles of revealed preference in traditional economics. Which ones, then? Impulsive? If you mean someone has a high discount rate, compared with some expectation, then that is fine: we know how to evaluate present discounted utility in traditional economics, and we can even do it with big discount rates as well as little discount rates. Maybe by 'impulsive' you mean 'non-exponential' or 'hyperbolicky' in the way we discount the future. Fair enough, but lets then have a serious debate about the evidence for hyperbolicky behaviour, since it is strikingly absent when one does careful experiments in the field for reasonable stakes (Andersen *et al.*, 2011b).

Again, the problem is that for all the hype about behavioural economics, it is just applied uncritically whenever some other explanation does not come quickly to mind. This is just lazy scholarship, common enough in behavioural economics but a serious matter when it is tossed around in development policy. For example, work habits of a driver in a developed country are likened (p. 89) to the apparent inefficiencies exhibited by New York taxi cab drivers analysed by Camerer *et al.*, 1997. Is it too much to ask that the critics of this study, most notably Farber (2005), be even acknowledged?

3.3 Relentless institutional destruction

Banerjee (2007, p. 7ff.) has a particular dislike of what goes on at the World Bank. He writes about

> A sad and wonderful example of how deep this lazy thinking runs is a book the World Bank brought out in 2002 with the express intention, ironically, of rationalizing the business of aid-giving. The book [. . .] was meant to be a catalogue

> of the most effective strategies for poverty reduction. . .[. . .] While many of these are surely good ideas, the authors of the book for not tell us how they know that they work.

Since the World Bank had not used randomised evaluations for any of these projects, or at most one, it is branded as 'resistant to knowledge.' Huh? How does that conclusion follow? Maybe stupid, maybe incompetent, but resistant to knowledge? When someone rejects that characterisation, he is viewed as just wanting to be respected rather than being right (p. 111). This is surely the shallow end of the pool of rhetoric.

Anyone who knows the public choice context of the activities of the World Bank need not be surprised at any of these criticisms. But it has also been a willing funding agency for many randomised evaluations as the skill sets and opportunities become available, as noted by Goldin *et al.* (2007, p. 29ff).

Indeed, the deeper issue is the decline in the World Bank's use of cost–benefit analysis in recent decades, whether or not that analysis is complemented by an RCT. The Independent Evaluation Group (2010) of the World Bank offers a thoughtful analysis of this process, honestly pointing out some of the abuses of cost–benefit analysis. Banerjee (2007, pp. 15, 122) is absolutely right to fume about the fool writing lines for a World Bank publication, referring to the 'success of an initiative' to provide computer kiosks in rural areas of India when it had just noted several lines earlier that very few had in fact been connected. It is easy to say that this is obviously not what cost–benefit analysis is meant to be, some anodyne *ex post* rationalisation of projects initiated for geo-political or long-forgotten reasons. The fact that it has become that in many circles, including regulatory policy in the USA, says nothing about its conceptual importance for welfare evaluation.

3.4 Relentless institutional self-promotion

The books by Banerjee and Duflo (2011) and Karlan and Appel (2011) are thinly disguised vehicles to promote their own organisations, and the importance of their existence. In one book there are constant reminders about J-PAL researchers and, guess what, in the other book there are constant reminders about IPA researchers. We are told by Karlan and Appel (2011, p. 28) that their organisation, Innovations for Policy Action (IPA), is the intellectual saviour for debates over development aid:

> . . .economists Jeffrey Sachs and Bill Easterly have butted heads for years over a very simple but elusive question: Does aid really work? At the root of their differences is a disagreement over what constitutes "evidence," and that's the rub. Until recently, the debate about aid effectiveness has been tied up in complicated econometrics and a mire of controversial country-level data. The cutting-edge research that IPA has done in evaluating the effective the effectiveness of specific development programs is finally giving us a new way to think about this question.

Wow! Where do we send the Nobel?

There is a discontinuity of their history on the origins of these intellectual-life-saving organisations, and they cannot both be right. Banerjee and Duflo (2011, p. 14) note that

> In 2003, we founded the Poverty Action Lab (which later became the Abdul Latif Jameel Poverty Action Lab, or J-PAL) to encourage and support other researchers, governments and nongovernmental organizations to work together on this new way of doing economics, and to help diffuse what they have learned amount policy makers. The response has been overwhelming. By 2010, J-PAL researchers had completed or were engaged in over 240 experiments in forty countries around the world, and very large numbers of organizations, researchers, and policy makers have embraced the idea of randomized trials.

Karlan and Appel (2011, p. 26) tell a somewhat different story. In their version of history, Karlan was the prime mover. They note that he

> . . .saw a void, a need for a new kind of organization with a head for academia, but with its feet squarely in the real world. It would serve as a loudspeaker and an advocate for policy-relevant research, and be full of people ready and eager to help generate research results, and, most important, it would work to scale-up the ideas that are proven to work. [Karlan] pitched the idea to [his] graduate advisers, Abhijit Banerjee, Esther Duflo, and Sedhil Mullainathan. They agreed that such an organization was sorely needed and, even better, they agreed to join the board [. . .]. Development Innovations was born, though its name would soon change. A year later, in 2003, Abhijit, Esther, and Sendhil started MIT's Poverty Action Lab [. . .], a center at MIT and network of like-minded researchers from around the world. J-PAL has an equally strong fervor for finding rigorous solutions to the problems of poverty. [. . .] From the beginning, Abhijit, Esther, Sendhil and [Karlan] knew how closely the two organizations would work together, so [they] changed the name of Development Innovations to Innovations for Poverty Action (IPA), and continue working together to this day. Each year IPA has managed to at least double in size, starting with $150 total revenue in 2002 [. . .], to $18 million in grants and contracts income in 2009. We now have some four hundred employees and projects in thirty-two countries.

Maybe there could have been two *Facebook*s after all?

Fogel (1962) undertook a famous study of what would have happened to American economic growth if the railroads had never been invented, and concluded 'not much'. Tongue in cheek, McAfee (1983) discussed the counterfactual problem for economic historians: what if Fogel had not written his article? Again the conclusion was 'not much'. Maybe in the absence of J-PAL and IPA we might have come across these insights anyway.

4. I will only believe it when I see it in the lab

We now have many rich models of behaviour, potentially allowing structural understanding of decisions in many settings of interest for the design of development policy. But we also realise that there are some basic confounds to reliable inference about behaviour. These are not side technical issues. Risk attitudes can involve more than diminishing marginal utility, and we have no significant problems identifying alternatives paths to risk aversion through probability weighting. Loss aversion is much more fragile, until we can claim to know the appropriate reference points for agents. Time preferences can be characterised, and appear to hold fewer problems than early experimental studies with lab subjects suggest.

4.1 wwSs: what would Savage say?

But the 600 pound gorilla confound is the subjective belief that decision-makers hold in many settings. This is the one that is widely ignored.[15] What Savage (1971, 1972) showed was that, under some (admittedly strong) assumptions, one could infer a subjective probability *and* a utility function from observable behaviour. The subjective probability and the utility function would each be well behaved in the classical senses, but one could not, in general, make claims about the one without knowing or assuming something about the other.[16]

[15] An important exception, from the behaviourist side of the Force, is Köszegi and Rabin (2008). They write (p. 196ff) about the 'impossibility of Skinnerian welfare economics' in the absence of measurement of subjective beliefs.

[16] Machina and Schmeidler (1992) provide an important extension to show that 'probabilistic sophistication,' in the sense of making decisions with underlying probabilities that obey the usual axioms of probability, did not require EUT. In particular,

The suggestion is not that casual assumptions about possible subjective beliefs should be used to rationalise 'rational behaviour' in every setting, but that inferences about cognitive failures, and the need for nudges, hinge on our descriptive knowledge of what explains behaviour. If we rule out some factor, then something else may look odd, just as a balloon bulges on one side if you press it on the other side. To take a simple example, assume that there is a risk premium, but one uses either a model that assumes that 100% of the observed behaviour is due to diminishing marginal utility or a model that assumes that 100% of the observed behaviour is due to probability pessimism. The first model will generate concave utility functions and impose zero probability weighting, and the second model will generate convex probability weighting functions and impose linear utility functions: both will likely explain the risk premium tolerably well. But the two models can have very different implications for the design of development policy that 'works' in any interesting sense.

Of course, in some settings it is simply not possible to 'go back to the well' and elicit information of this kind. However, there is no reason why one cannot use information from one sample, even from a different population if necessary, to condition inferences about another sample, to see the effect.

4.2 The lab and the field, not the lab versus the field

Preferences and beliefs have been elicited reliably in lab settings and in the field, although the myriad of contexts of the field mean that each application is in some important sense unique. The question to be asked is why these methods are not used more frequently in RCT evaluations of policies. This is beginning, but the attempts to elicit preferences and beliefs in existing randomised evaluations have been casual at best. Here we have a hypothetical survey question about risky behaviour, there we have an unmotivated question about beliefs, and rarely do we try to elicit time preferences at all. The potential complementarity between these methods is obvious, and conceded by all, but there seems to be relatively little appetite for careful field experiments to elicit preferences and beliefs. In part this derives from the way in which randomised evaluations have been marketed

---

Rank-Dependent Utility models, with increasing probability weighting functions, would suffice.

and promoted intellectually, as an antidote to the need to make structural economic or econometric assumptions.

This proposal to treat structural estimation in the field as complementary to a randomised treatment is not the same thing as saying that one should build full structural models of the effect of the intervention, although this is not ruled out. Advocates of randomised interventions often pose a false dichotomy between 'all-in theological' modelling via structural assumptions or 'agnostic eyeballing' of the average effects: Heckman (2010), in particular, takes aim squarely at this false tradeoff. The former is very hard to do well, and quite easy to do poorly. The latter is fine as far as it goes, but just does not go very far.

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 randomised 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.

References

Andersen, S., G.W. Harrison, M.I. Lau and E.E. Rutström (2011a) 'Discounting Behavior and the Magnitude Effect', Working Paper 2011-01, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

Andersen, S., G.W. Harrison, M.I. Lau and E.E. Rutström (2011b) 'Discounting Behavior: A Reconsideration', Working Paper 2011-03, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

Angrist, J.D. and A.B. Krueger (2001) 'Instrumental Variables and the Search for Identification: From Supply and Demand to National Experiments', *Journal of Economic Literature*, 15 (4): 69–85.

Angrist, J.D. and J.S. Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Ashraf, N., J. Berry and J. Shapiro (2010) 'Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia', *American Economic Review*, 100: 2383–413.

Banerjee, A.V. (2007) *Making Aid Work*. Cambridge, MA: MIT Press.

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

Banerjee, A.V. and E. Duflo (2011) *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: Public Affairs.

Benson, K. and A.J. Hartz (2000) 'A Comparison of Observational Studies and Randomized, Controlled Trials', *New England Journal of Medicine*, 342 (25): 1878–86.

Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir and J. Zinman (2010) 'What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment', *Quarterly Journal of Economics*, 125 (1): 263–306.

Bohm, P. (1984) 'Are there practicable demand-revealing mechanisms?', in H. Hanusch (ed.), *Public Finance and the Quest for Efficiency*. Detroit: Wayne State University Press.

Camerer, C., L. Babcock, G. Loewenstein and R. Thaler (1997) 'Labor Supply of New York City Cabdrivers: One Day at a Time', *Quarterly Journal of Economics*, 112: 407–41.

Cohen, J. and P. Dupas (2010) 'Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment', *Quarterly Journal of Economics*, 125 (1): 1–45.

Concato, J., N. Shah and R.I. Horwitz (2000) 'Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs', *New England Journal of Medicine*, 342 (25): 1887–92.

Deaton, A. (2010) 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, 48 (2): 424–55.

Duflo, E. (2006) 'Field experiments in development economics', in R. Blundell, W. Newey and T. Persson (eds), *Advances in Economics and Econometrics: Theory and Applications*, vol. 2. New York: Cambridge University Press.

Duflo, E., R. Glennerster and M. Kremer (2007) 'Using randomization in development economics research: a toolkit' in T.P. Schultz and J. Strauss (eds), *Handbook of Development Economics*, vol. 4. New York: North-Holland.

Duflo, E. and M. Kremer (2005) 'Use of randomization in the evaluation of development effectiveness', in G. Pitman, O. Feinstein and G. Ingram (eds), *Evaluating Development Effectiveness*. New Brunswick, NJ: Transaction Publishers.

Dufwenberg, M. and G.W. Harrison (2008) 'Peter Bohm: Father of Field Experiments', *Experimental Economics*, 11 (3): 213–20.

Farber, H.S. (2005) 'Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers', *Journal of Political Economy*, 113 (1): 46–82.

Fiore, S.M., G.W. Harrison, C.E. Hughes and E.E. Rutström (2009) 'Virtual Experiments and Environmental Policy', *Journal of Environmental Economics & Management*, 57 (1): 65–86.
Fisher, R.A. (1926) 'The Arrangement of Field Experiments', *Journal of the Ministry of Agriculture*, 33: 503–13.
Fogel, R. (1962) 'A Quantitative Approach to the Study of Railroads in American Economic Growth', *Journal of Economic History*, 22: 163–97.
Goldin, I., F.H. Rogers and N. Stern (2007) 'Forum', in A.V. Banerjee (ed.), *Making Aid Work*. Cambridge, MA: MIT Press.
Hacking, I. (1988) 'Telepathy: Origins of Randomization in Experimental Design', *Isis*, 79: 427–51.
Harrison, G.W. (2008a) 'Neuroeconomics: A Critical Reconsideration', *Economics & Philosophy*, 24 (3): 303–44.
Harrison, G.W. (2008b) 'Neuroeconomics: Rejoinder', *Economics & Philosophy*, 24 (3): 433–44.
Harrison, G.W. (2010) 'The Behavioral Counter-Revolution', *Journal of Economic Behavior and Organization*, 73: 49–57.
Harrison, G.W., J.L. Jensen, M.I. Morten and T.F. Rutherford (2002) 'Policy reform without tears', in A. Fossati and W. Weigard (eds), *Policy Evaluation With Computable General Equilibrium Models*. New York: Routledge.
Harrison, G.W., M.I. Lau and E.E. Rutström (2009) 'Risk Attitudes, Randomization to Treatment, and Self-Selection Into Experiments', *Journal of Economic Behavior and Organization*, 70 (3): 498–507.
Harrison, G.W. and J.A. List (2004) 'Field Experiments', *Journal of Economic Literature*, 42 (4): 1013–59.
Harrison, G.W., T.F. Rutherford and D.G. Tarr (2003) 'Trade Liberalization, Poverty and Efficient Equity', *Journal of Development Economics*, 71: 97–128.
Harrison, G.W., T.F. Rutherford, D.G. Tarr and A. Gurgel (2004) 'Trade Policy and Poverty Reduction in Brazil', *World Bank Economic Review*, 18 (3): 289–317.
Heckman, J.J. (2010) 'Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy', *Journal of Economic Literature*, 48 (2): 356–98.
Heckman, J.J. and R. Robb (1985) 'Alternative methods for evaluating the impact of interventions', in J. Heckman and B. Singer (eds), *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press.
Hotz, V.J. (1992) 'Designing an evaluation of JTPA', in C. Manski and I. Garfinkel (eds), *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press.
Imbens, G.W. (2010) 'Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)', *Journal of Economic Literature*, 48 (2): 399–423.
Independent Evaluation Group (2010) *Cost-Benefit Analysis in World Bank Projects*. Washington, D.C.: World Bank.

Kadane, J.B. and T. Seidenfeld (1990) 'Randomization in a Bayesian Perspective', *Journal of Statistical Planning and Inference*, 25: 329–45.

Karlan, D. and J. Appel (2011) *More Than Good Intentions: How a New Economics is Helping to Solve Global Poverty*. New York: Dutton.

Keane, M.P. (2010a) 'Structural vs. Atheoretic Approaches to Econometrics', *Journal of Econometrics*, 156: 3–20.

Keane, M.P. (2010b) 'A Structural Perspective on the Experimentalist School', *Journal of Economic Perspectives*, 24 (2): 47–58.

Köszegi, B. and M. Rabin (2008) 'Revealed mistakes and revealed preferences', in A. Caplin and A. Schotter (eds), *Positive and Normative Economics: A Handbook*. New York: Oxford University Press.

Kramer, M. and S. Shapiro (1984) 'Scientific Challenges in the Application of Randomized Trials', *Journal of the American Medical Association*, 252 (19): 2739–45.

Kremer, M. and A. Holla (2009) 'Pricing and access: lessons from randomized evaluations in education and health', in W. Easterly and J. Cohen (eds), *What Works in Development: Thinking Big and Thinking Small*. Washington, D.C.: Brookings Institution Press.

Leamer, E.E. (2010) 'Tantalus on the Road to Asymptopia', *Journal of Economic Perspectives*, 24 (2): 31–46.

Lee, D.S. and T. Lemieux (2010) 'Regression Discontinuity Designs in Economics', *Journal of Economic Literature*, 48 (2): 281–355.

Machina, M.J. and D. Schmeidler (1992) 'A More Robust Definition of Subjective Probability', *Econometrica*, 60 (4): 745–80.

McAfee, R.P. (1983) 'American Economic Growth and the Voyage of Columbus', *American Economic Review*, 73 (4): 735–40.

Murray, M.P. (2006) 'Avoiding Invalid Instruments and Coping with Weak Instruments', *Journal of Economic Perspectives*, 20 (4): 111–32.

Neyman, J. and E.L. Scott (1951) 'On Certain Methods of Estimating the Linear Structural Relation', *Annals of Mathematical Statistics*, 22: 352–61.

Pakes, A. (1982) 'On the Asymptotic Bias of Wald-Type Estimators of a Straight Line when Both Variables are Subject to Error', *International Economic Review*, 23 (2): 491–7.

Peirce, C.S. and J. Jastrow (1885) 'On Small Differences of Sensation', *Memoirs of the National Academy of Sciences for 1884*, 3: 75–83.

Rosenzweig, M.R. and K.I. Wolpin (2000) 'Natural 'Natural Experiments' in Economics', *Journal of Economic Literature*, 38: 827–74.

Salsburg, D. (2001) *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York: Freeman.

Savage, L.J. (1962) 'Subjective probability and statistical practice', in L.J. Savage (ed.), *The Foundations of Statistical Inference*. London: Methuen.

Savage, L.J. (1971) 'Elicitation of Personal Probabilities and Expectations', *Journal of American Statistical Association*, 66: 783–801.

Savage, L.J. (1972) *The Foundations of Statistics*. 2nd edn. New York: Dover Publications.
Smith, V.L. (1982) 'Microeconomic Systems as an Experimental Science', *American Economic Review*, 72 (5): 923–55.
Sorenson, R.A. (1992) *Thought Experiments*. New York: Oxford University Press.
Stock, J.H., J.H. Wright and M. Yogo (2002) 'A Survey of Weak Instruments and Weak Identification in Generalized Method of Movements', *Journal of Business and Economic Statistics*, 20 (4): 518–29.
Stone, M. (1969) 'The Role of Experimental Randomization in Bayesian Statistics: Finite Sampling and Two Bayesians', *Biometrika*, 56 (3): 681–3.
Thaler, R.H. and C.R. Sunstein (2008) *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.
Thistlethwaite, D.L. and D.T. Campbell (1960) 'Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment', *Journal of Educational Psychology*, 51 (6): 309–17.
Wald, A. (1940) 'The Fitting of Straight Lines if Both Variables are Subject to Error', *Annals of Mathematical Statistics*, 11: 284–300.
Weil, D.N. (2009) 'Comment', in W. Easterly and J. Cohen (eds), *What Works in Development: Thinking Big and Thinking Small*. Washington, D.C.: Brookings Institution Press.
Worrall, J. (2007) 'Why There's No Cause to Randomize', *British Journal of the Philosophy of Science*, 58: 451–88.
Ziliak, S.T. (2008) 'Guinnessometrics: The Economic Foundation of "Student's" t', *Journal of Economic Perspectives*, 22 (4): 199–216.