Chapter 7: Retreat From Radical Skepticism: Rebalancing Theory, Observational Data and Randomization in Development Economics

Christopher B. Barrett and Michael R. Carter

<ls>

Introduction

Not so long ago, the typical empirical development economics seminar reached a moment of specific or structural skepticism in which attendees theorized about specific omitted or ignored structural factors that might plausibly bias reported estimates of the coefficients of interest. The speaker responded with detailed comments on the data generation process, perhaps convincing attendees that their specific concerns were ill-founded. Attendees departed, mentally re-centering reported confidence intervals up or down depending on the nature of the plausible bias, expanding the girth of those same intervals, and doing a Bayesian merger of the adjusted results into their understanding of the pre-existing literature. When the seminar went poorly, the skewered speaker would fall fitfully asleep that night, wishing surveys had measured confounding variables, perhaps later dreaming of random variation in key policy variables or in the instruments that predict their uptake or placement.

While few would dispute the usefulness of random variation in key variables, the meteoric rise over the past decade or so of randomization methods in development economics has replaced structural skepticism with a generalized skepticism, or what Stokes (this volume) calls radical skepticism. Following Stokes, radical skepticism imagines a world in which the number of potentially confounding and omitted factors is unmanageably large or even unknowable. In this world, the burden of proof falls uniquely on the researcher to establish the orthogonality of key variables to the unknown universe of potentially confounding factors. Skepticism becomes generalized, and the only defensible identification strategy is random allocation of the key independent variable across the population of interest. Seminar attendees can intone stock phrases

about identification, rather than theoretically grounded concerns specific to the problem at hand. Discussion of causal pathways becomes secondary to concerns about the indisputably exogenous assignment of households (or firms, or areas) to treatment and control groups. The unbiasedness of point estimates, rather than their mean squared error, becomes the central statistical preoccupation, and Bayesian learning from imperfect evidence seems to be replaced by a binary division of evidence into that which meets a purported statistical gold standard, and that which does not. Skewered seminar speakers still do not sleep well at night.

While this shift from structural to radical skepticism may make for less interesting seminars, it bespeaks a more fundamental problem. As Stokes points out, if radical skepticism is correct, it not only undermines research based on observational data, it also poses an intractable problem of essential heterogeneity to randomized controlled trials (RCTs). If we cannot count, model and potentially measure factors that might be spuriously or otherwise correlated with key variables in observational data, then we can similarly never know if a universe of unknowable factors mediates the effects of even randomly distributed treatments. Well-identified local average treatment effects become data-weighted averages of multiple response regimes or unknowable dimensionality. Generalization to other populations, where the relative preponderance of the regimes may be different, becomes indefensible. Radical skepticism thereby destroys in equal measure the internal validity of observational studies and the external validity of RCTs. One cannot invoke one without unleashing the other.

Ironically, by generalizing the researcher's natural, healthy skepticism about new findings and divorcing it from specific, theoretically grounded concerns, radical skepticism drives a wedge between the use of RCTs in development economics and its use in the agricultural and bio-medical disciplines from whence RCTs originate and draw much of their inspiration. Bhargava (2008) and Stokes (this volume) give examples from coronary medicine in which theoretically grounded causal

models identify specific confounding factors (coronary artery dimension and c-reactive protein levels) that mediate the treatment impacts of cholesterol-lowering drugs. If what economists call the "policy relevant treatment effect" is to be inferred, measurement of these confounders is of equal importance irrespective of whether one is examining observational or RCT data.

We development economists might benefit from greater fidelity to the biological scientists' approach, to integrate in a more balanced fashion the experimental testing of hypotheses about effects associated with specific causal pathways with the theory and insights from observational data that yield those hypotheses and that enable us to generalize from necessarily small samples to the broader development experience. This requires a retreat from radical to structural skepticism and rebalancing the roles of theory, observational data and randomization in development economics research. Such a move will not only make for more meaningful RCTs, it will reopen the door to measuring heretofore unobservables and (skeptically) learning from observational data. The latter is especially important because the understandable excitement about the statistical power of RCTs can blind us to its pitfalls in economic applications. Because of these limitations, we continue to need analysis based on observational data.

In the remainder of this paper, we will first look at four problems that limit what we can (or should) learn about economics from RCTs:

1. Ethical constraints;

2. Faux Exogeneity of uncontrolled treatments;

3. Measuring efficacy rather than effectiveness; and

4. Scale and temporal limits to randomization.

We will then close the paper by returning to the issue of rebalancing theory, measurement and randomization in order to obtain internally and externally valid understandings of the development economics process.

<ls>

<txa> Ethical Constraints

Experimental research involving any sort of researcher-managed intervention requires safeguards to protect the rights and welfare of humans participating as subjects in the study. Four broad classes of ethical dilemmas nonetheless routinely arise in experiments conducted by development economists. While these dilemmas share commonalities with the dilemmas that confront, say, medical trials, many are specific to social and economic interventions. These dilemmas receive distressingly little attention in graduate training and in the literature, and are of concern not only because they violate ethical principles sacrosanct to all serious research, but also because they commonly lead subjects, implementers or both to actively circumvent the research design, thereby undercutting the statistical raison d'être of the initial randomization. While uncomfortable to discuss, raising these issues now seems preferable to waiting for a major disaster to occur that would undermine the research agenda.

The first and most obvious class of ethical dilemma revolves around the unintended but predictable adverse consequences of some experimental designs. The "do no harm" principle is perhaps the most fundamental ethical obligation of all researchers. Most universities and serious research organizations have institutional review boards established to guard against precisely such contingencies. Nonetheless, many highly questionable designs make it through such reviews and the results have been published by leading economists in reputable journals. As but one prominent example involving widely respected scholars, Bertrand et al. (2007) randomized incentives for subjects in India who did not yet possess a driver's license, so as to induce them to bribe officials in

order to receive a license without having successfully completed the required training and an obligatory driver safety examination. The very predictable consequence of such an experiment is that it imperils innocent non-subjects – not to mention the subjects themselves – by putting unsafe drivers on the road illegally. This is an irresponsible research design, yet the study was published in one of the profession's most prestigious journals. Such research plainly signals insufficient attention paid to fundamental ethical constraints on field experimentation within economics.

A second example comes from the study of what Gugerty and Kremer (2008) call the Rockefeller Effect. Taking its cue from John D. Rockefeller, who refused to give money to Alcoholics Anonymous on the grounds that the money would undercut the organization's effectiveness, the Gugerty and Kremer (2008) article explicitly sets out to determine whether grants of money to women's organizations in Kenya distorts them and leads to the exclusion of poorer women and their loss of benefits. Donor groups were providing grants to women's organizations on the presumption that they were doing good. Proving otherwise, and that the Rockefeller Effect is real, could of course be argued to bring real social benefit. However, the ethical complexities of undertaking research designed to potentially harm poor women are breathtaking. Standard human subjects rules require: (1) that any predictable harm be decisively outweighed by social gains; (2) that subjects be fully informed of the risks; (3) that compensation be paid to cover any damages incurred; and, (4) that the study be curtailed if and when it becomes clear that it is causing harm. It remains unclear whether these rules were met in the Gugerty and Kremer (2008) study, which is somewhat chilling given that the study indeed confirms that poor women were harmed by the injection of cash into randomly selected women's groups.

A third example illustrates how attempts to experimentally vary the economic fundamentals of individual behavior can introduce risks of harm to human subjects that do not exist in

observational studies. Jensen (2010) gave students at randomly selected schools information on the estimated returns to schooling that exceeded the students' prior, pre-intervention beliefs. Through this experimental intervention, Jensen hoped to show that perceptions artificially depress educational attainment, a hypothesis that would be challenging to test using observed variation in students' prior beliefs as these beliefs would surely be endogenous to a range of child and parental attributes. To be sure, this is a research and policy question of first-order importance so there is merit to trying to obviate the endogeneity of prior perceived returns to education. The problem is that researcher-provided information on the returns to education might itself be wrong and mislead the subject into, say, overinvestment in schooling.

In his paper, Jensen includes a healthy discussion of this ethical issue (see especially footnote 23), and his protocol included a warning that the estimated returns to education provided by the research team may be an inaccurate predictors of individual outcomes. Somewhat ironically, Jensen relies on observational data to estimate expected returns to education, despite the fact that a standard skeptical perspective might see those estimates (drawn from those who endogenously chose to complete higher levels are schooling) as upwardly biased for those who, without the informational intervention, would have chosen not to complete secondary schooling. Moreover, all students in the Jensen study were given the same unconditional estimate of returns to schooling, rather than conditional estimates adapted to the student's circumstance (e.g., a rural resident with poor local schools). As Jensen observes, provision of this sort of systematic misinformation would not be harmful if we knew that the subjects, pre-intervention, were underinvesting in education.

While that condition may be true, it is also the fundamental issue being studied. Absent that condition, we find ourselves looking at an experimental information treatment that predictably misinforms at least some parents and children with the specific intent of inducing behavioral change.

For some students, the information provided was likely much closer to the (unknown) truth than were the subjects' prior beliefs, generating gains for the subject. But for some others, it misled them into overinvestment in education at nontrivial cost relative to the resource allocation they would have chosen under their uninformed prior. And the researcher cannot even identify the subjects harmed by the experiment, so it is infeasible even to make amends through compensation.

A fourth and final example further illustrates the ethical dilemmas that are specific to efforts to answer core economic questions that are intertwined with purposeful human behavior. In an effort to determine the impact of capital access on the capital constrained, Karlan and Zinman (2009a) worked with a South African paycheck lender to 'de-ration' a randomly selected subset of loan applicants whose credit scores deemed them credit-unworthy. De-rationed individuals included those marginally below the credit score threshold, as well as some individuals well below that threshold. The research design aimed to compare post-loan outcomes of the de-rationed with the non-derationed potential borrowers.

While promising as an approach, the Karlan-Zinman (2009a) study illustrates several intrinsic ethical difficulties of implementing RCTs with real economic institutions. Unlike studies that randomly gift people with liquidity increments,[1] the Karlan-Zinman (2009a) study created real debt for the randomly de-rationed, exposing them not only to the benefits of liquidity, but also to the penalties of prospective default. Given that the lender's scoring model predicted repayment difficulties for the de-rationed, ethical concerns appear important here. From a human subjects' protection perspective, implementing such experiments would thus require full disclosure to the de-rationed and an ability to compensate them for any harm caused for the sake of experimental learning. However, fulfilling these standard human subjects requirements (e.g., by telling a de-rationed study participant that a lender's credit scoring model predicts they will fail, but that the

study will restore their reputation and collateral should they default) would obviously change behavioral incentives and destroy the internal validity of the experiment. This underscores how researchers' ethical obligations often confound the purity of experimental research design.

The second class of ethical problem emergent in many development experiments revolves around the suspension of the fundamental principle of informed consent. This raises the subtle but important distinction between treating human beings as willful agents who have a right to participate or not as they so choose, versus treating them as subjects to be manipulated for research purposes. To avoid the various endogenous behavioral responses that call into question even the internal validity of experimental results (due to Hawthorne effects and the like), many prominent studies randomize treatments in group cluster designs such that individuals are unaware that they are (or are not) part of an experiment. The randomized roll-out of Progresa in Mexico is a well-known example for development economists (Schultz 1994). Even when the randomization is public and transparent, cluster randomization maintains the exogeneity of the intervention, but at the ethically-questionable cost of sacrificing the well-accepted right of each individual participant to informed consent, as well as the corresponding obligation of the researcher to secure such consent. Biomedical researchers have given this issue much thought (e.g., Hutton 2001), but we have yet to see any serious discussion of this issue among development economists.

Informed consent becomes a more serious concern as development agencies push for ever-greater "ownership" of development policies and projects by target populations and their political leaders. In order for communities to "own" the results of a study, they must understand and trust the study design and its implementers. But blinding research subjects to the details of their treatment condition, which is commonly necessary in RCTs – such as Karlan and Zinman (2009a) or Jensen (2010) – requires expressly withholding information from them. That is hardly a strategy for

building trust between researchers and the subject communities that we then hope will internalize and act upon the research findings. Thus, in addition to the ethical concerns of strategically under-informed consent, practical concerns arise that experimental treatment may subtly undermine subject communities' willingness to adopt results as their own and to subsequently act as those more-statistically-pure estimates suggest.

A third class of problem revolves around the role of blindedness in experiments. In most natural sciences, the entire response of physical material can be attributed to the treatment to which it has been subjected. But when humans are the subjects, response is a complex product of both the treatment itself and the perceived difference in treatment between oneself and other subjects. Hence the importance of blinding subjects – and, in best practice, "double blinding" researchers as well – regarding their placement within a study's control or treatment group. But whereas biomedical researchers can commonly develop and distribute to a control group a placebo identical to the experimental treatment medicine, few RCTs make any effort to blind subjects. Indeed, in many cases it would be infeasible to do so, as the economic treatment of interest involves obviously differential exposure to a new product, institution, technology or resource.

This matters for both ethical and statistical reasons. The well-known placebo effect associated with treatment has an important corollary, in other words, that those who know themselves to be in a control group may suffer emotional distress when subjected to discernibly different treatment and that such distress can have adverse biophysical consequences that exaggerate the differences between control and treatment groups. Clinical researchers are deeply divided on the ethics of unblinded research.[2]

Moreover, the emotional suffering inflicted by unblinded treatments often induces active efforts to undo the randomized assignment. Subjects have been known to enroll themselves in

multiple trials until they get a lucky assignment draw as part of the treatment cohort, and implementers discreetly violate the assignment rules as a merciful response to randomly assigned emotional and physical suffering. In this way ethical dilemmas quickly turn into statistical problems as well; the clean identification of randomization gets compromised by human agency to overcome the perceived inequity of differential treatment.

The fourth class of ethical dilemma arises from abrogating the targeting principle upon which most development interventions are founded. Given the scarce resources and fiduciary obligations of donors, governments and charitable organizations entrusted with resources provided (voluntarily or involuntarily) by others, there is a strong case to be made for exploiting local information to improve the targeting of interventions to reach intended beneficiaries (Alderman 2002; Conning and Kevane 2002). The growing popularity of community funds and community-based targeting involves exploiting precisely the asymmetric information that randomization seeks to overcome.

By explicitly refusing to exploit private information held by study participants, randomized interventions routinely treat individuals known not to require the intervention instead of those known to be in need, thereby predictably wasting scarce resources. Indeed, in our experience the unfairness and wastefulness implied by strict randomization in social experiments often sows the seeds of some implementers' breach of research design. Field partners less concerned with statistical purity than with practical development impacts commonly deem it unethical to deny a "control group" the benefits of an intervention strongly believed to have salutary effects, or to knowingly "treat" one household instead of another when the latter is strongly believed likely to gain and the former not. Well-meaning field implementers thus quietly contravene the experimental design,

compromising the internal validity of the research and reintroducing precisely the unobserved heterogeneity that randomization was meant to overcome.

<ls>

<txa> The Faux Exogeneity of Uncontrolled Treatments

The core purpose of RCTs is to use random assignment in order to ensure that the unconfoundedness assumption essential to identifying an average treatment effect holds (Imbens 2010). In the abstract, this is a strong argument for the method. Problems arise, however, when pristine asymptotic properties confront the muddy realities of field applications, and strict control over fully exogenous assignment almost inevitably breaks down, for any of a variety of reasons discussed below or in the preceding section on ethical dilemmas. The end result is that the attractive asymptotic properties of RCTs often disappear in practice, much like the asymptotic properties of other IV estimators. We term this the "faux exogeneity" problem.

In retrospect, the seminal deworming study carried out by Miguel and Kremer (2004) may have misdirected subsequent researchers in that it was based on a medical treatment in which it was possible to know exactly what had been given, *and received*, by the treated subject.[3] However, when randomization is used for larger, economics-oriented topics (e.g., changing agents' expectations by offering them new contract terms or technologies), the true treatment received by subjects becomes harder to discern. Subjectively perceived treatments are likely non-randomly distributed among experimental subjects whose capacities to comprehend and to act vary in subtle but substantive ways. Unobservable perceptions of a new product, contract, institutional arrangement, information, technology or other intervention vary among participants and in ways that are almost surely correlated with other relevant attributes and expected returns from the treatment.

An obvious example is the previously mentioned study of perceived returns to education (Jensen 2010). The experimenter knows the information provided to treated subjects but cannot possibly know what change, if any, this information induced in their prior beliefs. Another example of this problem can be found in Karlan and Zinman's (2008) effort to determine price elasticity of credit demand using randomly distributed price variation in an established paycheck lending market. In this study, researchers surprised randomly selected loan applicants by offering them loans at rates other than the usual market rate. In analyzing the resulting data on loan uptake, Karlan and Zinman find a kink in the demand at the existing market interest rate. Increases above that level reduced demand, but reductions do not symmetrically increase it. A possible interpretation of this odd finding is that potential borrowers did not find the announcement of a price below the usual market price to be credible ("there must be something in the fine print"). While a medical experiment can largely control the treatment (e.g., so many milligrams of a drug injected into the blood stream), manipulating prices and other phenomena is more complex. Although the price announcement was randomized, we really do not know what the treated subjects effectively perceived. Some may have reacted suspiciously to a seemingly good deal (after all, there is supposed to be no free lunch), and others (perhaps those be able to read loan contract language) may have received the intended treatment, acting as if the market price really had decreased. This uncontrolled treatment (some treatment group subjects received the intended treatment, others did not, and which was which was likely correlated with subjects' education and sophistication) can result in what we called 'faux randomization', and raises a serious issue of interpretation. Unlike medical trails, human agency and understanding can confound the use of RCT to study economic problems.

As this example illustrates, the correlation between treatment (lower interest rates) and confounding factors (borrower sophistication) that one seeks to remedy through randomization can creep back in (Heckman, Urzua and Vytlacil 2006; Heckman 2010.). In our view, it is far better to be

aware of and explicit about likely bias due to unobserved heterogeneity than to hide it under the emperor's clothes of an RCT that does not truly randomize the treatment to which agents respond; this is crucially distinct from the treatment the experimenter wishes to apply.

Note that this unobservably heterogeneous treatment problem differs from the well-recognized compliance problem, which induces the important distinction between the average treatment effect (ATE) in the population of interest and the local average treatment effect (LATE) that is identified only for the subpopulation which complies with the treatment (Angrist, Imbens and Rubin 1996). Proponents of LATE estimates – which are not specific to RCTs but are more general to all IV estimators – routinely argue that LATE is the policy-relevant parameter because monitoring compliance is difficult and ineffective. This is true. But in the presence of unobservable heterogeneous treatments within the compliant subpopulation, even the LATE estimate becomes uninformative. Using the "intent to treat" approach to return to the ATE estimate likewise fails to overcome the problem. This point obviously applies generally, not solely to RCTs. In our experience, however, random assignment too often fosters overconfidence such that claims of clean identification blind the researcher to this problem.

A somewhat similar problem can result from the use of side payments designed to bolster the voluntary uptake of a new program within a treatment group, the so-called "encouragement design". While such payments may be absolutely essential if an RCT is to achieve any measure of statistical power, in the presence of essential heterogeneity (i.e., some agents will benefit more than others from an intervention) encouragement designs can result in a different population, with different expected benefits, than the population that would eventually take up the intervention absent of the subsidy built in to the experiment to encourage uptake of the treatment condition.

Note that this is a fundamentally different problem than medical researchers confront when employing payments to encourage participation. Participants in medical studies presumably have no idea whether their particular biological system will respond more or less favorably to a treatment than the system of the average person. We would not therefore expect that higher payments would bring in people who know that they will benefit less from the treatment. In contrast, many economic interventions (e.g., access to a new financial contract or technology) depend precisely on participants understanding and evaluating the returns to the new treatment. Mullally et al. (2010) illustrate this problem and the bias it imparts to estimated average treatment effects, using an encouragement design employed to evaluate an agricultural insurance program in Peru. The fact that bias may creep in need not obviate the study. Indeed, by specifically modeling the bias, Mullally et al. show that the biased estimator is more informative in the mean square error sense than is the unbiased estimator obtainable without encouragement at reasonable sample sizes. They make, in other words, the very old fashioned point that biased estimators can be more informative and useful than noisy, unbiased estimators, a point to which we return in the conclusion to this paper.

Another source of faux exogeneity arises due to the challenges of implementing RCT designs in the field. Intended random assignments are commonly compromised by field teams implementing a research design, especially when government or NGO partners have non-research objectives for the intervention that must be reconciled with researchers' aims to cleanly identify causal effects. The ethical concerns raised in the preceding section are but one common source of conflicting aims. Corruption, incomplete comprehension of research methods, logistical complications, etc., also lead to imperfect implementer compliance with the intended research design, and thus to sampling bias.

Note that this compliance problem differs from the problem of non-compliant subjects that partly motivates IV estimation of LATE. This problem creeps in earlier in the research, routinely emerging when implementers select survey respondents for observational studies, and thus compromising the claimed integrity of the data collection. In the pre-RCT research environment, this was (at least) equally commonplace but less fatal of a flaw than when true randomization is itself the source of identification. These crucial details of how design deviates from implementation are almost never reported in papers that employ experimental methods, unlike in the natural sciences, where the exact details of experiments are systematically recorded and shared with reviewers and made publicly available to readers for the purpose of exact replication.[4] Indeed, when research is subcontracted to implementing partners, study authors commonly do not even know if such sampling bias exists in the data.

Unobservably heterogeneous treatments, encouragement bias and sampling bias in economic studies undercut the 'gold standard' claim that RCTs reliably identify the (local) average treatment effect for the target population (i.e., that RCT estimates have internal validity). Just as the original (monetary) gold standard depended on a range of strong assumptions, so does the claim of internal validity depend on multiple, strong, often-contestable assumptions. As with studies based on conventional, observational data, the development economics community needs to interrogate the underlying identifying assumptions before accepting RCT results as internally valid.

The preceding general point is not novel. Heckman (1992; 2010), Deaton (this volume) and Leamer (1983; 2010) discuss a variety of statistical limitations to the internal validity of RCT estimates that merit brief mention. One that we especially highlight, because we find it such a commonplace problem, is that randomization bias is a real issue in the typically small samples involved in RCTs. The identical equivalence of control and treatment subpopulations is an

asymptotic property only. The power calculations now routine in designing experimental studies necessarily tolerate errors in inference just as non-experimental studies do. And, unlike many quasi-experimental studies such as those that rely on propensity score matching, RCT studies frequently fail to confirm that control and treatment groups exhibit identical distributions of observable variables. This problem is easily fixed and the best RCT studies carefully check for balance. But the frequency with which this is ignored in RCT-based studies today betrays a dangerous overconfidence that pervades much of the RCT practitioner community today.

Given the likelihood of randomization bias in small samples, experimental approaches must take special care to balance control and treatment groups based on observables. But there is no standard practice on how to best do this and not all methods of randomization perform equally well in small samples. Bruhn and McKenzie (2009) find that pairwise matching and stratification outperform the most common methods used in RCTs in smaller samples. As a result, standard errors reported in RCT studies that do not control for the used randomization method are commonly incorrect, leading researchers to incorrect inferences regarding treatment effects.

In summary, RCTs are invaluable tools for biophysical scientists, where the mechanisms involved are more mechanical than is the case in behavioral and social sciences, and where virtually all conditions can be controlled in the research design. Human agency complicates matters enormously, as is well known in the agricultural and medical literatures on experiments. It is often unclear what varies beyond the variable the researcher is intentionally randomizing. Hawthorne Effects are but one well-known example. As a result, impacts and behaviors elicited experimentally are commonly endogenous to environmental and structural conditions that vary in unknown ways within a necessarily highly-stylized experimental design. This faux exogeneity undermines the claims of clean identification due to randomization. In our experience, this is the rule in RCTs, rather than

the exception. As Leamer (2010, p.33) vividly writes: "[y]ou and I know that truly consistent estimators are imagined, not real. . . . [But some] may think that it is enough to wave a clove of garlic and chant "randomization" to solve all our problems just as an earlier cohort of econometricians have acted as if it were enough to chant "instrumental variable.""

<ls>

<txa> Measuring Efficacy Rather than Effectiveness

The aim of most RCTs is to eliminate the endogeneity that commonly plagues explanatory variables of interest in observational data. But it is by no means clear that purging agents' endogenous behavioral response is always desirable. Often the question of greatest interest is what will happen in response to real people's non-random responses to the introduction of a policy, project or technology. Precise answers to the wrong question are not always helpful. Put differently, when we impose exogenous allocations we do not, in fact, replicate real human behavior. Indeed, we violate the most fundamental proposition of microeconomics: that resource allocation is endogenous.

The crucial distinction between the impact of an exogenously imposed treatment and of a treatment allowing for full endogeneity is reflected in the epidemiology and public health literature as the difference between efficacy – the study of a treatment's capacity to have an effect, as established under fully controlled, ideal conditions – and effectiveness – the study of induced change under real-life conditions, as in clinical practice. Economists who seek to inform agents making real decisions in the real world ultimately need to be able to address questions of effectiveness, not merely efficacy. The latter is a first, not the final, step toward scientific demonstration of interventions that work and are ready for widespread promotion.

The biomedical sciences clearly recognize this, as reflected, for example, in the Society for Prevention Research's formal standards of evidence, meant to "provide guidance for the research

community to generate and test evidence-based prevention programs to improve the public health" (Flay et al. 2005, p.152). These standards explicitly note that "effective programs and policies a subset of efficacious interventions" (Flay et al. 2005, p. 151). Once again, greater fidelity to the literature of the biophysical sciences that have more developed experimental traditions would induce development economists to retreat from radical skepticism and to more creatively balance observational and experimental data, both informed by precise structural theories of change. Overcorrection for endogeneity may, ironically, render findings that are unbiased but irrelevant to the real-world questions concerning the intervention under study.

<ls>

<txa> Scale and temporal limits to randomization

As noted by other commentators, one shortcoming of experimental methods is that only a non-random subset of relevant topics is amenable to investigation via RCTs. For example, macroeconomic and political economy questions that many believe to be of first-order importance in development are clearly not candidates for randomization (Basu 2005; Deaton this volume; Rodrik 2009). Nor are infrastructure issues or any other meso- or macro-scale intervention that cannot be replicated in large numbers. Furthermore, the placement of these issues is necessarily and appropriately subject to significant political economy considerations (Ravallion 2009). As one moves from smaller scale, partial equilibrium questions (for example, "Which type of contract generates a greater response from a microfinance institution's clientele?", or "What is the marginal effect of cash versus food transfers on recipients' nutritional status?") to larger scale, general equilibrium and political economy questions, RCTs necessarily become less useful.

The fact that RCTs are not appropriate for all questions is not a criticism of the methodology per se. However, it becomes a serious problem when RCTs are seen as the way of

knowing, and the applicability of the RCT method seems to drive research agendas rather than the importance of the question being asked. While it would be unfair to single out individual papers, most readers of development economics literature can easily recall papers that, in their zealous quest for exogenous variation, prove points utterly obvious to laypersons. More worrisome is when leading development economists tell policy-makers that the questions they ask which are not amendable to analysis by RCTs are the 'wrong questions,' or that economists know nothing until an RCT has been implemented and has generated a point estimate.

To reiterate, these observations are not an argument against RCTs. Instead, they are an argument for rebalancing the research agenda and recognizing the complementarity of different ways of knowing. Even questions apparently amenable to exclusively RCT analysis may be less amenable than they seem at first glance. An example from the evaluation of Mexico's well-known Progressa program may help clarify this point. Given the amount of time required to accumulate and realize returns to human capital, exclusive reliance on RCTs to evaluate program like Progressa can be problematic. The RCT analysis of the Progressa cash transfer program identified a statistically significant increase of 0.7 years of schooling (Schultz 1994). However, inferring the (long-run) economic significance of this increase was inevitably left to other methodologies that assembled best estimates of the long-term earnings impact of this additional schooling, an exercise necessarily requiring a series of assumptions, including those of a general equilibrium nature (e.g., Behrman, Sengupta, and Todd 2005). The point is that the best research recognizes and exploits the fundamental complementarities among methods.[5] No method has a unique claim to being able to answer most important questions on its own.

RCTs can in principle be exploited over a long-term time horizon, as shown by the recent follow-up to the early 1970s INCAP study of childhood nutritional intervention in Guatemala

(Hoddinott et al. 2008; Maluccio et al. 2009). While the ability to observe adult earnings (and other outcomes) 40 years after the intervention is striking, the unavoidably large attrition rate (and the assumptions therefore required to make inference) placed this study into the Bayesian hopper as one further, important piece of (quasi-experimental) evidence that needed to be mixed with the extensive research on education in developing countries using non-experimental methods in order to generate important, highly policy-relevant findings.[6]

<ls>

<txa> Finding Our Balance in Development Economics

The limits to RCTs in development economics by themselves mandate a return to methodological pluralism if we are continue to answer the important questions. That said, there is a large class of economic development problems whose answers can be ethically and feasibly pursued using randomization methods. RCT studies focus on generating consistent and unbiased estimates of treatment effects of development interventions. In the biophysical sciences from which the RCT tradition arises, this often works because basic physio-chemical laws ensure a certain degree of homogeneity of response to an experiment. But in the behavioral sciences, such as economics, there is even less reason to believe in homogeneity of response to a change in environmental conditions. Furthermore, there is such heterogeneity of microenvironments that one has to be very careful about model mis-specification. These concerns apply to all research but seem especially overlooked in the current RCT fashion.

These observations relate closely to what is perhaps the most widespread critique of experimental evidence, namely that RCT estimates lack external (out of sample) validity. There are two dimensions to this external validity critique. The first dimension is that unobservable and observable features inevitably vary at the community level and cannot be controlled for in experimental design due to context matters (see, for example, Acemoglu 2009; Deaton this volume;

Ravallion 2009; Rodrik 2009). For example, is an agency that is willing to implement an experimental design for a pilot program likely to be representative of other agencies that might implement it elsewhere? Probably not, and in ways that almost surely confound the measurable impacts of the experiment.

The second dimension of the external validity critique postulates the existence of treatment effects that vary systematically with unobserved individual heterogeneity within the sample (Heckman, Urzua, and Vytlacil 2006). As discussed above, in this context, RCTs generate point estimates that are unknown, data-weighted averages across subpopulations of multiple types with perhaps zero population mass on the weighted mean estimate. As is true of any research method that pools data from distinct subpopulations, there is a nontrivial probability that no external population exists to whom the results of the experiment apply on average. Collecting the data experimentally does not solve this problem. If the inferential challenge largely revolves around essential heterogeneity rather than around endogeneity, experiments that address only the latter issue can at best claim to solve a problem of second-order relevance.

Ironically, these critiques of unobserved and unknown confounds are the very same ones used by some RCT advocates to delegitimize learning from observational data. To slightly misquote Shakespeare's Hamlet, radically skeptical RCT engineers have been hoisted with their own petard. While one conclusion might be to retreat entirely from empirical economic research, a more promising path is to retreat from radical skepticism and let theory and careful observation – what Frank (2007) terms "economic naturalism" – guide an understanding of the causal process and name the potential confounds that can cripple inference from both observational data and from RCTs.

As we have argued in a companion paper (Barrett and Carter 2010), advances in behavioral economics that render observable measures of key sources of economically relevant heterogeneity (such as degrees of risk aversion and time preferences) hold out promise for studies using both

experimental and observational data. Of course, in the end, no measures are beyond reproach. We need to treat them all with healthy, structural skepticism and retreat from the radical skepticism that prevails in much of economics today. Rebalancing the mix of theory, observational data and randomization in development economics will not deliver an elusive gold standard. But it just might return us to the humility of Bayesian learning and the relentless pursuit of, not unbiasedness per se, but of confidence intervals that are sufficiently narrow to reliably guide policy advice and the design of development interventions that can improve the human condition without predictably harming those whom we hope will benefit from our research.

<ls>

<txa> Acknowledgements

Portions of this paper are taken with permission from Oxford University Press from our 2010 paper "The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections," *Applied Economic Perspectives and Policy* 32(4):515-548.

<ls>

<txa> References

Acemoglu, Daron. 2009. Theory, General Equilibrium, Political Economy and Empirics in Development Economics. *Journal of Economic Perspectives* 24(3): 17-32.

Alderman, Harold. 2002. Do local officials know something we don't? Decentralization of targeted transfers in Albania. *Journal of Public Economics* 83(3): 375-404.

Angrist, Joshua, Guido Imbens, and Donald Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91(434): 444-472.

Barrett, Christopher B. and Michael R. Carter. 2010. The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections," *Applied Economic Perspectives and Policy* 32(4):515-548.

Basu, Kaushik. 2005. The New Empirical Development Economics: Remarks on Its Philosophical Foundations. *Economic and Political Weekly* XL (40): 4336–4339.

Behrman, Jere R., Piyali Sengupta, and Petra E. Todd. 2005. Progressing through PROGRESA: An Impact Assessment of Mexico's School Subsidy Experiment. *Economic Development and Cultural Change* 54(1): 237-275.

Bertrand, Marianne, Simeon Djankov, Rema Hanna, and Sendhil Mullainathan. 2007. Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption. *Quarterly Journal of Economics* 122(4): 1639-76.

Bhargava, Alok. 2008. Randomized controlled experiments in health and social sciences: Some conceptual issues," *Economics and Human Biology* 6:293-298.
Bruhn, Miriam and David McKenzie. 2009. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics* 1(4): 200-232.
Conning, Jonathan and Michael Kevane. 2002. Community-based targeting mechanisms for social safety nets: A critical review. *World Development* 30(3): 375-94.
Deaton, Angus. 2010. Instruments, Randomization, and Learning About Development. *Journal of Economic Literature* 48(2): 424-455.
de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. Returns to Capital in Microenterprises: Evidence from a Field Experiment. *Quarterly Journal of Economics* 123(4): 1329-1372.
Flay, Brian R., Anthony Biglan, Robert F. Boruch, Felipe González Castro, Denise Gottfredson, Sheppard Kellam, Eve K. Mościcki, Steven Schinke, Jeffrey C. Valentine and Peter Ji 2005. Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination. *Prevention Science* 6(3): 151-175.
Frank, Robert H. 2007. *The Economic Naturalist: In Search of Explanations for Everyday Enigmas*. New York: Basic Books.
Gugerty, Mary Kay and Michael Kremer. 2008. Outside Funding and the Dynamics of Participation in Community Associations. *American Journal of Political Science* 52(3): 585-602.
Harmon, Amy. 2010. New Drugs Stir Debate on Rules of Clinical Trials. *New York Times* September 19, 2010. http://www.nytimes.com/2010/09/19/health/research/19trial.html?_r=1&emc=eta1.
Heckman, James J. 1992. Randomization and Social Policy Evaluation. In Charles F. Manski and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, MA: Harvard University Press.
Heckman, James J. 2010. Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature* 48(2): 356–98.
Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. Understanding Instrumental Variables in Models with Essential Heterogeneity. *Review of Economics and Statistics* 88(3): 389-432.
Hoddinott, John, John A. Maluccio, Jere R. Behrman, Rafael Flores, and Reynaldo Martorell. 2008. Effect of a Nutrition Intervention During Early Childhood on Economic Productivity in Guatemalan Adults. *The Lancet* 371(9610): 411-416.
Hutton, Jane L. 2001. Are Distinctive Ethical Principles Required for Cluster Randomized Controlled Trials? *Statistics in Medicine* 20(3): 473–488**.**
Imbens, Guido W. 2010. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48(2): 399-423.
Jensen, Robert. 2010. The (Perceived) Returns to Education and the Demand for Schooling *Quarterly Journal of Economics* 125(2): 515-548.
Karlan, Dean and Jonathan Zinman. 2008. Credit Elasticities in Less Developed Countries: Implications for Microfinance. *American Economic Review* 98(3): 1040-1068.
Karlan, Dean and Jonathan Zinman. 2009a. Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. *Review of Financial Studies doi: 10.1093/rfs/hhp092*
Leamer, Edward E. 1983. Let's Take the Con Out Of Econometrics. *American Economic Review* 73(1): 31-43.
Leamer, Edward E. 2010. Tantalus on the Road to Asymptotia. *Journal of Economic Perspectives* 24(2):31-46.

Maluccio, John A., John Hoddinott, Jere R. Behrman, Reynaldo Martorell, Agnes R. Quisumbing, and Aryeh D. Stein. 2009. The Impact of Improving Nutrition During Early Childhood on Education Among Guatemalan Adults. *Economic Journal* 119(537): 734-763.

Miguel, Edward and Michael Kremer. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72(1): 159-217.

Mullally, Conner, Stephen R. Boucher, and Michael R. Carter. 2010. Perceptions and Participation: Mistaken Beliefs, Encouragement Designs, and Demand for Index Insurance. Unpublished manuscript, University of California, Davis.

Ravallion, Martin. 2009. Should the Randomistas Rule? *BE Press Economists' Voice.*

Rodrik, Dani. 2009. The New Development Economics: We Shall Experiment, but How Shall We Learn? In *What Works in Development: Thinking Big and Thinking Small.* Jessica Cohen and William Easterly, eds. Washington, D.C.: Brookings Institution Press.

Schultz, T. Paul. 2004. School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics* 74(1): 199-250.

Stokes, Susan C. 2010. A Defense of Observational Research. *This Volume.*

<ls>

<txa>notes

[1] De Mel, Mackenzie and Woodruff (2009) give away random amounts of liquidity, but at the cost of failing to distinguish liquidity-constrained from liquidity-unconstrained households, as Barrett and Carter (2010) discuss.

[2] See Harmon (2010) for an example from a current controversy in oncology research.

[3] Even in the Miguel and Kremer (2004) case, incomplete treatment due to non-random school attendance on days in which treatments were administered leads to bias of unknown sign. The authors report incomplete uptake but never fully explore its implications.

[4] An important exception is the Karlan and Zinman (2009) study discussed below.

[5] One of us learned this lesson the hard way when a Finance Ministry official greeted evidence that cash transfers seem to statistically significantly increase height by two centimeters with the dreaded question, "So what?"

[6] A forty-year wait for an answer is, of course, often not practical. The INCAP studies are rare examples.