*Applied Economic Perspectives and Policy* (2010) volume 32, number 4, pp. 515–548.
doi:10.1093/aepp/ppq023

Featured Article

# The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections

Christopher B. Barrett* and Michael R. Carter

Christopher B. Barrett, Stephen B. and Janice G. Ashley Professor in the Charles H. Dyson School of Applied Economics and Management, Cornell University. Michael R. Carter, Professor in the Department of Agricultural and Resource Economics, University of California-Davis.
* Correspondence to be sent to: cbb2@cornell.edu.

*Submitted January 2009; accepted October 2010.*

**Abstract** *Impact evaluation based on randomized controlled trials (RCTs) offers a powerful tool that has fundamentally reshaped development economics by offering novel solutions to long-standing problems of weak causal identification. Nonetheless, RCTs suffer important and underappreciated pitfalls, some of which are intrinsic to the method when applied to economic problems, others that are the result of methodological boosterism. Among the pitfalls are ethical dilemmas, uncontrollable treatments that result in a 'faux exogeneity,' distortion of the research agenda, and a tendency to estimate interventions' abstract efficacy rather than their effectiveness in practice. We illustrate these points through the literature on smallholder capital access and productivity growth. Ultimately, we argue for a methodological pluralism that recognizes all identification strategies' limitations.*

**Key words:** Ethics, evaluation, identification, randomized controlled trials.

**JEL Codes:** A10, B00, O10, O30.

## Introduction

The challenge of doing development economics, and doing it well,[1] motivates the constant search for novel methodological responses. In searching to establish what policy or project interventions truly work, as well as why, and under what circumstances – and to credibly prioritize those interventions that really do work – the big challenge that

[1]*The study of development concerns human beings as agents whose choices, constrained and conditioned by the external environment, result in behaviors that matter not just to their own well-being, but also, due to externalities and general equilibrium effects, to the aggregate experience of their communities. Economists study human choice subject to scarcity in order to better understand these behaviors and the resulting outcomes. That understanding is in turn expected to have reliable and practical implications about policies and programs that induce changes in behavior by expanding opportunities and relaxing constraints, leading to improved economic welfare.*

development economists face is identifying interventions' causal impacts and key behavioral parameters and structural factors that condition or explain those impacts. One such response, the use of randomized controlled trials (RCTs) to study the impact of specific programming interventions, has over the last decade become an important, if not dominant, methodology in development microeconomics. In contrast to other methods designed to solve similar problems of inference (e.g., panel data methods or Heckman (1992) estimators), RCTs have engendered considerable controversy. Our goal in this article is to pick through this controversy and re-center the debate on how development economics can continue to advance and contribute to the understanding of real-world problems whose importance transcends that of methodological affinity.

It is essential to recall that RCTs arose in response to widespread problems of weak identification in empirical economics, not just in development but in other applied fields as well, including labor, health and public economics. As doubts grew about the effectiveness of controlling for confounding variables in regressions based on observational data (Angrist 1990; Angrist, Imbens and Ruben 1996; Card 1990; Krueger 1999; Lalonde 1986; Leamer 1983; Imbens and Wooldridge 2009), concern also grew that economists could not credibly identify the causal impacts of interventions. This concern fed rising interest in and the use of (natural or designed) experimental approaches to find an alternative, and perhaps superior, method to rigorously estimate program impacts. The rise of RCTs and their associated methods has indisputably raised the bar in applied economic analysis, and appropriately enhanced the scrutiny of identification strategies for inferring causal effects. The most sophisticated proponents never trumpeted these methods as a panacea, merely as progress. Indeed, Leamer (1983) famously cautioned that "[o]ne should not jump to the conclusion that there is necessarily a substantive difference between drawing inferences from experimental as opposed to nonexperimental data," (p.31).

But sometime during the ensuing quarter century or so, the humble search for improved methods of generating believable answers to pressing policy questions gave way to what many now perceive as methodological triumphalism. Much of the controversy surrounding RCTs seems to be an artifact of its most fervent advocates proclaiming the RCT as the "gold standard" marking the apparent end of methodological history, and of other researchers' uncritical acceptance of these exaggerated claims.[2] The more ardent "randomistas'" epistemological claims have provoked sometimes acerbic observations that this line of research is of baser metal than gold (e.g., see Basu 2005; Deaton 2010; Leamer 2010; Ravallion 2009). Our goal here is neither to rehash others' arguments concerning the validity of RCT impact estimates, nor to intensify an already-overheated debate. Instead, starting from the perspective that exogenous variation in policy or control variables is unarguably statistically useful (Imbens 2010), we highlight some less discussed, but no less important, limitations of RCTs as actually applied to economic development problems.

We will argue that some of the RCTs' limitations are intrinsic to the methodology when used in real-world field applications – as opposed to

[2]*While there are certainly devotees of the Heckman (1992) estimators, an identifiable group of 'Heckmanistas', nor of 'Panelistas' did not emerge, as those who use these methods do not pretend that they decisively resolve identification problems the way so many RCT enthusiasts do.*

their use under idealized conditions – while other limitations are extrinsic and can be easily shed once RCTs become better integrated into development economics as *a* way of knowing, not as *the* way of knowing. After reviewing these limitations in the context of two specific bodies of literature, we conclude with thoughts on a balanced, problem-centric approach to development economics. Among other points, we advocate that greater attention be given to the ethics of employing RCTs, the greater utilization of behavioral economic experiments to help resolve some of the same fundamental identification problems that motivate reliance on RCTs, and the more thoughtful use of nonexperimental findings to structure the behavioral models that should underpin good RCT design.

To help frame the discussion that follows, consider the following stylized simultaneous equation model that empirical development economists typically confront:

$$b = g(y, p, s, \pi, \varepsilon_b) \tag{1}$$

$$y = f(b, p, s, \pi, \varepsilon_y), \tag{2}$$

where $b$ is a behavior (e.g., input use or food purchase) and $y$ is a development outcome (e.g., living standards or nutritional status). In addition to depending on each other, behaviors and outcomes may also depend directly on a set of policy variables, $p$, that are amenable to intervention. Behaviors and outcomes may also depend on a set of observable structural determinants, $s$, (e.g., individual, household, community, economy-wide characteristics), as well as on a set of typically unobserved preferences and characteristics, $\pi$, (e.g., honesty, ambition, time and risk preferences), and of course on classical measurement, sampling and specification errors, $\varepsilon_b$ and $\varepsilon_y$.

While the fact that the above two equations are simultaneous and of unknown functional form clearly poses a challenge for statistically identifying the impact of structural and policy factors, more germane to our discussion is that in observational data, policy variables are unlikely to be orthogonal to typically-unobserved preferences and characteristics. For example, credit or new technologies are more likely to be adopted by those with lesser risk aversion or stronger entrepreneurial aptitude. The severity of this unobserved heterogeneity problem will, among other things, depend on whether characteristics that drive adoption are left in the error term, or whether, as we discuss later, behavioral field experiments can be used to measure those characteristics and move them out of the error term and into the observable vector, $s$.

While panel data or structural modeling of the adoption/selection process can potentially be used to control for this identification problem, the validity of these methods depends on specific, fairly strong assumptions. RCTs attempt to circumvent these limitations by randomizing the policy instrument, $p$, across the population in an effort to guarantee the statistical independence of $p$ from $\pi$ and $s$. The advantage of such randomization is obvious and well-described by Banerjee and Duflo (2008), as well as Duflo, Glennerster, and Kremer (2008).

The limitations of RCTs are perhaps less obvious. After briefly reviewing the well-known potential power of RCTs in the next section, we go on to outline key intrinsic pitfalls in the use of RCTs in development

economics, including ethical constraints and the common (but generally unrecognized) problem we call 'faux exogeneity', which emerges when study subjects are active economic agents as opposed to passive biophysical subjects who receive blinded medical treatments. In the subsequent section, we examine extrinsic pitfalls to RCTs created when excessive attachment to their putative gold standard quality distorts the research agenda and disconnects analysis from the richer body of evidence and thinking. We then briefly describe the potential power – and parallel pitfalls – of behavioral field experiments as an alternative means to aid reliable statistical identification. The penultimate two sections then illustrate the power and the pitfalls of RCTs in two key bodies of literature, that on access to capital, and that on productivity growth in smallholder agriculture. We conclude the paper with thoughts on a methodological way forward.

## The Power and Promise of Randomized Controlled Trials

Social experiments based on RCTs have become a workhorse of contemporary development microeconomics. As in the labor economics literature, which has followed a parallel course, the laudable objective of RCTs is to resolve the serious econometric problems associated with program placement and selection effects – the uptake of a new intervention is non-random – as well as the endogeneity of key $p$ or $s$ variables.

RCTs are a specific case of longstanding instrumental variables (IV) methods; the randomized variable is simply a strictly exogenous instrument. These methods are widely and appropriately deemed necessary to address the difficulty of identifying the causal effects of programs whose participants differ systematically from non-participants. The absence of an observable counterfactual – what would have happened to the same person, in the same place and time, with and without the program – compels the researcher to make comparisons over time (before-and-after estimation) or with a different subpopulation of nonparticipants. In either case, there are almost surely non-random differences that account for part of the observed differences in the dependent variable, thereby contaminating the resulting estimate of the program's "effect".

The solution adopted by many development economists today is to pilot new programs as RCTs. Statistically, randomized inclusion in the treatment group becomes an instrumental variable that is by construction meant to fulfill the standard conditions for a valid instrument, especially statistical independence from expected program benefits, and more generally from the error structure.[3] RCTs have been deployed en masse by a range of scholars (including ourselves), especially those associated with MIT's Poverty Action Lab, its sister organizations at other universities, and various donor agencies to answer a range of specific questions relating to education, health, finance, agriculture and other areas of interest to many development economists.

A central tenet of the RCT movement is that economic theory is excessively limiting, and that we must be open to surprises. Through clever experimental designs and this openness to letting the data lead the

[3]*All of the familiar critiques of IV estimation apply to experimental methods, including finite sample bias, weak correlation with the endogenous behavior or condition of intrinsic interest, etc.*

researcher, rather than vice versa, leaders of the RCT movement in development economics, such as Abhijit Banerjee, Esther Duflo, Dean Karlan, Michael Kremer and Ted Miguel, have opened up important new areas of inquiry in development economics and helped build bridges to the behavioral economics literature, which has similarly championed healthy skepticism about many longstanding assumptions of neoclassical economic theory. These advances have helped reinvigorate development economics, for which the community owes the randomistas heartfelt, collective gratitude.

## The Intrinsic Pitfalls of Randomized Controlled Trials in Development Economics

Given their clear association with the subdiscipline's resurgence, RCTs have become enormously popular in development economics. As is commonly true of wildly successful new innovations, however, one also recognizes signs of herd behavior, of uncritical uptake, naïve disregard of the product's flaws, and inappropriate use by its adherents. The new product also commonly elicits harsh criticism from users of the prior generation of methods.

RCTs in development economics are the methodological progeny of biomedical trials, based on elementary principles of experimental design. In this section, we consider problems that occur when we attempt to apply these methods to economic problems in which the system under study is a (general equilibrium) behavioral system populated by *agents* who consciously choose their responses, not a biological or physical system in which (typically unconscious) *subjects* such as blood chemistry, cancer cells or a virus respond endogenously following the laws of nature.

### *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. 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.

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 get published by otherwise 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 – let alone the subjects themselves – by putting unsafe drivers on the road illegally. This is 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 similar, but perhaps less egregious example, comes from the study of the so-called Rockefeller Effect (Gugerty and Kremer 2008). 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; and, (3) that compensation be paid to cover any damages incurred. 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.

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.

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, i.e., 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.[4]

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 appropriately 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.

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

*Faux Exogeneity and Other Internal Validity Problems*

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.[5] 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 capacity to comprehend and to act vary in subtle but substantive ways. Unobservable perceptions of a new product, contract, institutional arrangement, 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.

The unobserved heterogeneity problem that one seeks to remedy through randomization can thereby 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.

[5]*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.*

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, Boucher, and Carter (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.

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.[6] 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

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

population (i.e., that RCT estimates have internal validity). Just as the original gold standard depended on a range of strong assumptions – which ultimately proved untenable, leading to the collapse of the gold standard – 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 (2010), 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 biomedical literature and ecological literature 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.""

*The LATE May Miss the Point*

Even if one overlooks the preceding ethical and internal validity problems – or somehow believes that an RCT reduces the magnitude of these problems relative to other feasible methods – RCTs still ultimately only identify the (local) average treatment effect (Deaton 2010). But much of what is interesting in development economics transcends the unconditional mean. Policymaker interest revolves instead around other properties of the distribution of effects, especially the conditional effects (e.g., what is the effect on women or on children or on the poorest quantiles?), as well as the proportion of positive and negative effects and the characteristics of those likely to fall in each of those groups. In our experience, these latter effects have a far greater influence on the ultimate political economy of the scale-up of seemingly successful development interventions than do estimated mean treatment effects.

RCT studies focus on generating consistent and unbiased estimates of treatment effects. 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 little 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.

As we illustrate in subsequent sections on capital access and smallholder productivity, much of the point in development economics research is to uncover the essential heterogeneity of response and the underlying structure that accounts for such heterogeneity, not just the mean marginal response. But RCTs rarely uncover underlying structural features of the mechanisms of greatest interest to private and public sector decision-makers. Typically, they only reveal the LATE; in a more refined form, and with adequate sample size, they may allow for the estimation of a LATE within quantiles or distinct subpopulation strata of interest. If the heterogeneous responses found within a population drive the political economy of policy-making and project scale-up, and if interventions require targeting to employ scarce resources responsibly, it is unclear if the effects identifiable using RCTs shine light in the right place.

A core pitfall is that experiments typically treat human beings as subjects, not as agents. When measurable outcomes are the core variables of interest, as is typically true in evaluation research, the behavioral mechanisms that yield these outcomes in the non-experimental economy are almost inevitably subordinated in research design. This problem is compounded when the phenomena of interest – such as market equilibria, outcomes that fundamentally depend on collective action, etc. – arise from complex multi-agent interactions not readily reproducible in experiments (as we discuss below with regard to capital access). Furthermore, it is by no means clear that purging agents' endogenous behavioral response is always desirable given that the core question of 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.

Indeed, the endogenous processes that guide resource allocation by human agents, whether by policy-makers or individuals, can ultimately undermine the quest to eliminate endogeneity through research design. 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. Overcorrection for endogeneity may, ironically, render findings consistently irrelevant to the real-world questions concerning the intervention under study.

### *External Validity of Randomized Controlled Trials*

Probably the most widespread critique of experimental evidence revolves around the external validity of results (see, for example, Acemoglu 2009; Deaton 2010; Ravallion 2009; Rodrik 2009). In brief, the problem is that unobservable and observable features inevitably vary at the community level and cannot be controlled for in experimental design due to context matters. 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 affect the measurable impacts of the experiment.

Furthermore, given the essential unobserved heterogeneity within a sample (Heckman, Urzua, and Vytlacil 2006), field experiments generate only point estimates that are effectively an unknown data-weighted average 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.

Out-of-sample predictive and prescriptive analysis requires understanding mechanisms, which in turn requires a (falsifiable) model of behavior and resulting welfare outcomes. We want not just to evaluate the impact of distinct actions but, even more, to know why there is impact, how it arises, and whether it is likely replicable or scalable.

As an example, one of the authors sat on a review panel that considered a proposal to implement an RCT of a novel cash transfer program in a region of the developing world. An economic anthropologist on the panel familiar with the region confidently predicted that the RCT estimate of the reduced form of equation (2) above would find large and positive

treatment effects. However, the anthropologist went on to note that the finding would be entirely an artifact of inter-tribal politics in the region and would tell us absolutely nothing about the way the program's mix of incentives and payments would work elsewhere. Thus, while any analysis is a prisoner of its study area, RCT studies armed with strong instruments appear empowered to overlook the deep exploration of structure and behavior that more conventional approaches require.

## Extrinsic Pitfalls of Randomized Controlled Trials

Like all empirical methods applied to complex social phenomena, randomized controlled trials are subject to intrinsic limitations, suggesting a Bayesian approach to learning in which individual results are skeptically blended[7] with the prior literature. While this observation is almost trite, enthusiasm for the power of RCTs has tended in practice to delegitimize other ways of knowing, providing license to ignore existing literatures, or to avoid critical problems not easily amendable to study with RCTs. While this practice is in no way intrinsic to RCTs, its impact on development economics has become striking, as RCTs have become an almost hegemonic professional discourse. In taking note of these extrinsic problems of RCTs, we hope to help re-center the discipline and merge – not purge – RCTs into the broader development economics enterprise.

### *Distortion of the Research Agenda*

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 2010; 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 interventions is necessarily and appropriately subject to significant political economy considerations (Ravallion 2009). As one moves from smaller, partial equilibrium questions (e.g., "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 general equilibrium and political economy questions, RCTs necessarily become less (or not at all) 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, rather than the importance of the question being asked, seems to drive the research agenda. 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

[7] *By "skeptical blending", we mean that reported standard errors should more realistically be treated as a lower bound on the unknown true imprecision of the parameter estimates.*

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 so 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.[8] 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.[9]

### *Forgetting the Opportunity Cost of Interventions*

Incomplete and problematic policy recommendations can also emerge when RCT enthusiasm delegitimizes other ways of knowing. RCTs typically aim to establish the treatment effect of an intervention against a counterfactual of no intervention. Given the complexity of multi-factorial randomized block design of high order dimensionality, comparisons among multiple candidate interventions – so that the research can establish the opportunity cost of pursuing one intervention, not just the intervention's gross impact – remain very limited in practice due to feasibility constraints. Most sciences with a well-established experimental tradition therefore proceed humbly. Researchers rely on a rich array of theory and observational evidence to generate experimental designs whose results are

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

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

then added to the theory and observational evidence so as to gradually generate a set of prescriptions.

In economics we teach undergraduates the fundamental importance of opportunity costs in explaining and informing choices. In allocating scarce development programming dollars, the question is not whether a particular intervention has an impact relative to a counterfactual of doing nothing, but whether it is more cost-effective than other interventions. While this observation again borders on the trite, the delegitimation of non-experimental evidence severely tempts analysts to make immodest policy and project recommendations from one-off RCT results that are not held to a stern evaluation of the true opportunity cost of an intervention.

Influential work on deworming exemplifies this problem. The heavily-cited field experiment by Miguel and Kremer (2004) has been marketed into a global solution at www.dewormtheworld.org. Although deworming is clearly a desirable and effective intervention relative to doing nothing, how many public health professionals seriously consider deworming the best use of very scarce health care dollars in developing countries? In an informal poll of twenty public health professionals with extensive experience working with children in developing countries, not a single one deemed deworming one of the top three public child health concerns worth investing in, and only two placed it in the top five. Rather, the health care professionals emphasized the importance of early childhood and prenatal nutrition, immunization against infectious diseases, and breastfeeding.[10,11]

As with the discussion in the prior subsection, this problem is not intrinsic to the RCT method per se, and its solution is also to be found in merging RCT results into the broader stream of the existing literature and methods. This observation applies with particular force when we as economists begin to experiment with interventions outside of our area(s) of expertise, where deep bodies of literature can inform us about both mechanisms and impacts, as discussed in the capital access and productivity sections that follow. While the RCT literature will and should continue to grow in the future, the policy at present, when decisions have to made, demands more careful consideration of the opportunity costs posed by alternative interventions.

## Behavioral Field Experiments and Other Approaches to Non-confoundedness

By randomizing the allocation of a development program (across people and/or space), RCTs are designed to eliminate any systematic or spurious relationship between program treatment and other factors that may affect the outcome variable of interest, thereby assuring that program impacts

[10]*The Copenhagen Consensus 2008 evaluation of highest impact interventions corroborates this casual assessment. "Deworming and other nutrition programs at school", (a far broader set of interventions than simply deworming) was ranked as the fifth most-desirable health or nutrition intervention (see http://www.copenhagenconsensus.com/Default.aspx?ID=1318).*

[11]*This problem replicates a pervasive problem in biomedical research, as FDA approval of a new medicine generally requires a company to demonstrate a drug's efficacy relative to a placebo, but not necessarily relative to other existing or potential treatments for the same ailment. We thank Jordan Matsudaira for pointing this out.*

are not confounded with the impacts of these other factors. As discussed above, the actual practice of RCTs typically fails to achieve this idealized state, but nevertheless surely does typically reduce selection and other endogeneity problems. Of course, this same goal can also be achieved by measuring potentially confounding factors and including them directly in the statistical analysis.

Fixed effects estimators are one example of such an approach. Because fixed effect estimators do not directly measure potentially confounding factors, their validity depends on the assumption that the product of the unobserved confounding factor and the coefficient(s) that determine its impact on the outcome variables is time invariant. This assumption is quite reasonable for some types of interventions (e.g., where genetic inheritance may matter to health outcomes), but less for others (e.g., where risk perceptions almost surely evolve over time). Moreover, fixed effects estimators offer an extreme example of how assuring non-confoundedness by measuring factors that may be non-orthogonal to the treatment or intervention variable is likely to be statistically expensive in the sense of requiring larger sample sizes to achieve the same precision. That being said, panel data methods, and the assumptions on which they depend, continue to merit consideration for questions that cannot easily or ethically be studied with RCTs.

Behavioral field experiments offer another approach to resolving the challenge of unconfoundedness by directly measuring individual characteristics that might be correlated with non-randomly allocated programs and interventions. Laboratory experiments of the sort that historically predominate in behavioral economics are rare in development economics where the behavioral experiments tradition overwhelmingly involves field experiments, which have enjoyed a surge of popularity in economics more broadly over the past decade or so (Harrison and List 2004; List, Sadoff, and Wagner 2010). Many of these field experiments focus on measuring preferences and basic behavioral parameters; rarely are they designed as an identification strategy to answer basic development questions.

Perhaps the most famous behavioral field experiments in development economics are the early work by Binswanger (1980; 1981), in which he used lottery games to elicit risk aversion measures in an effort to explain small farms' choice of technology. While Binswanger's work was focused on the impact of risk aversion per se, it can be retrospectively read as an effort to more reliably identify the impact of other factors (structurally or spuriously correlated with risk aversion) that influence technology choice. This is one of the under-utilized powers of behavioral experiments; by controlling explicitly for credibly measured parameters that are unobservable in conventional survey data, they can effectively eliminate much of the confounding that RCTs aim to eliminate while simultaneously shedding useful light on interesting behavioral patterns. Behavioral experiments are not immune, of course, from some of the intrinsic problems associated with experimental methods, as discussed above, but they are often viable where RCTs are not and, moreover, have not fallen prey to the extrinsic problems of RCTs.

Behavioral experiments are especially important if we acknowledge the importance of social context, identity, norms, etc., (Barrett 2005; Cardenas 2009) in shaping a number of behaviors (e.g., asset accumulation) and conditioning program impacts. As one example of the sorts of added insights

one can gain from incorporating experimental methods, Carter and Castillo (2005) use a number of familiar behavioral experiments to gauge norms of altruism and trust within rural communities and use these to condition the estimated effects of endogenous social interactions on households' recovery from the devastation caused by Hurricane Mitch. The experimental data reveal tremendous intra-community heterogeneity based on norms typically unobservable in conventional survey data. This heterogeneity of response is prospectively very important to policy-makers, community leaders or humanitarian organizations trying to facilitate disaster response and recovery.

Risk management is a longstanding topic of deep interest to development economists. Making progress in the description of actual risk management behavior by residents of poor communities, much less on predictive or prescriptive analysis to help guide policy-makers, almost surely depends on improving and extending the profession's use of experiments intended to elicit key behavioral parameters that are surely heterogeneous in population, correlated with observables that are amenable to intervention, and otherwise unobtainable using traditional data collection methods.

Experiments can be used effectively to replicate alternative conditions in realistic ways so as to answer fundamental policy questions. For example, growing enthusiasm for using insecticide treated bednets (ITBs) as an inexpensive means of preventing malaria among the poor in rural Africa has sparked considerable debate about the effectiveness of free ITB distribution programs. Proponents deem it essential to provide ITBs to as many of the poor as possible, and as quickly as possible, while opponents worry that those unwilling to buy an ITB will resell freely given nets, thereby undermining both the intent of the giveaway program and the commercial distribution system for ITBs. Hoffmann, Barrett, and Just (2009) use an experimental auction mechanism to test the hypothesis that poor households will keep and use free ITBs; their experimental results offer strong evidence that the liquidity, income and endowment effects of ITB giveaways virtually eliminate the hypothesized resale behavior, providing rigorous evidence to inform a key policy decision not readily addressed using observational data alone.

Behavioral experiments are of course subject to their own assumptions and limitations, just like RCTs. Our argument here is simply that, like panel data methods, they offer a potentially valuable alternative approach to solving confoundedness problems that bedevil the use of observational data to answer impact evaluation and other development economics questions. Fixed effects estimators and the direct inclusion of behavioral parameters derived from field experiments offer another path to the same end sought through RCTs, and can be used as either complements to RCTs or as substitutes where randomization is infeasible, unethical or impractical.

## Access to Capital and Randomized Controlled Trials: Power and Pitfalls

Building on the prior discussion, this section and the next examine the role that RCTs have played – and might play – in the advancement of

two longstanding areas of research in development economics: smallholder capital access, and productivity. We structure the discussion around key questions in these research areas with the hope that the resulting discussion will help point the way towards a more fruitful integration of experiments into development economics research in the years ahead.

Access to capital, especially by small-scale farmers and other low wealth households, has long been a major intellectual preoccupation of development economics and an area of intense policy intervention. Early thinking and matching interventions were rooted in the perspective that monopolistic lenders exploited and limited the economic advance of households. However, the eventual collapse of the statist credit policies that accompanied this perspective ushered in *laissez-faire* perspectives and policies. These in turn were displaced by theory and practice more attentive to the economics of imperfect information and the possibility that collateral-constrained (low wealth) households might be subjected to non-price rationing and credit market exclusion, even in perfectly competitive markets. The continuing microfinance (MFI) boom – and its search for collateral substitutes and credit allocation processes immune from adverse selection and moral hazard – is a natural outgrowth of this imperfect information perspective.

Against the backdrop of this brief intellectual history, the following four questions retain their salience:

(1) Does the financial market work such that we find households in the non-price-rationed regimes, or do markets work in an efficient, price-rationed manner for all?
(2) If non-price rationing exists, is it systematically biased against any particular set of households (e.g., low wealth households) such that the operation of the competitive economy tends to reinforce initial levels of poverty and inequality?
(3) How costly is non-price-rationing and how much would household input use and income increase if liquidity constraints could be relaxed and non-price-rationing eliminated?
(4) Are there contractual or institutional innovations that can change the rules of access to capital, lessen non-price-rationing, and decrease its cost?

The empirical literature that tries to answer these questions faces severe challenges that are the result of the prospect of non-price rationing in loan markets. In the first instance, the fact that there exists double selection in credit markets – borrowers have to want to borrow and lenders have to be willing to lend – heightens concerns over separating the impact of capital access from the impact of the characteristics of those doubly-selected to receive credit.[12] More deeply, econometric work in this area faces a fundamental identification problem common to any market (potentially) in disequilibrium: Observed transactions are the minimum of supply and demand, and without further knowledge or assumptions, it is unclear whether any data point tells us something about the supply curve or the demand curve.

[12]*As Adams (1988) observed long ago, it is difficult to know if the observed higher performance of those receiving loans is due to the impact of liquidity or to pre-existing differences that would have led borrowers to produce more than non-borrowers even in the absence of credit.*

Being able to distinguish credit-rationed households, for whom demand, $D$, exceeds supply, $S$, is also important because theory indicates that basic behavioral relationships are different between the two regimes, implying a switching regime that might have the following form:

$$b_i = \begin{cases} \alpha^c p_i + \beta^c x_i + [e_i + \varepsilon_i] if\ D_i > S_i \\ \alpha^u r_i + B^u z_i + [e_i + \varepsilon_i] if\ D_i \leq S \end{cases}, \quad (3)$$

where the superscript $c$ refers to the constrained (credit-rationed) regime, while the superscript $u$ denotes the unconstrained.[13] As pointed out by Feder et al. (1985), behavior (e.g., input use) or outcomes (e.g., farm income) will depend on prices in the constrained regime ($r$, the interest rate), whereas it will depend on quantities in the constrained regime.

The empirical literature has tried to address both of these identification problems. In analysis that retrospectively seems naïve, one of the authors has written papers that explicitly assumed that all households had positive demand and were in the credit-constrained regime (Carter 1989; Sial and Carter 1996). The effort of those and similar works was to control econometrically for both observed and unobserved differences between borrowers and non-borrowers, and thereby to reliably identify the pattern of access to capital and the impact of changing it.

Leaving aside the adequacy of efforts to control for latent characteristics, this literature was appropriately criticized for failing to correctly sort households into the correct behavioral regimes. Beginning with Feder et al. (1985) and extending through such papers as: Kochar 1998; Bell, Srinivasan, and Udry 1997; Carter and Olinto 2003; and Boucher, Guirkinger, and Trivelli 2009, a variety of econometric and survey techniques have been employed to resolve the two identification problems in order to make reliable inferences on the key questions that make access to capital a vital question. As with all empirical work, the specific findings of this literature can be disputed. However, the aggregate weight of the evidence would indeed seem to indicate that capital access is highly problematic in many regions of the world, especially for low wealth households.

In this context, it is useful to ask what has been and what might be contributed to our understanding of capital access by behavioral field experimental methods and randomized controlled trials. In a paper titled "Giving Credit where Credit is Due," Banerjee and Duflo (2010) argue that RCTs have proven their worth through contributions to the literature on access to capital. After reviewing the contributions of the RCT literature to date, this section will close with reflections on how behavioral field experiments and RCTs might be further utilized to help understand important problems of capital access.

### *Faux Exogeneity and the Extent of Non-price Rationing*

As has been well explored in the literature on credit markets and asymmetric information, non-price-rationing can occur when interest rate increases that might otherwise equilibrate the market induce adverse changes in the borrower population that make expected lender profits go down, not up, with the interest rate increase. While the empirical literature

[13] *Recent theoretical work on 'risk rationing' (Boucher, Carter, and Guirkinger 2008) suggests that the regime structure is even more complex than that illustrated here.*

based on observational data briefly discussed above has attempted to estimate the severity of non-price rationing and its costs, an alternative approach to detecting at least the existence of the forces necessary for non-price rationing is to randomly vary the interest rate and determine whether default increases with higher rates, as would be expected from an asymmetric information perspective. In principle, this approach should shed some light on access to capital question 1 above.

In an ambitious study, Karlan and Zinman (2009b) worked with a South African paycheck lender to randomize interest rates and then observe the resulting default rates. While their actual experimental protocol was somewhat complex – involving solicitations sent to the lender's existing client base – their primary finding is that the interest rate perturbations themselves had no impact on default.

As observed by Banerjee and Duflo (2010), it is a bit hard to know what to make of these results. Had Karlan and Zinman (2009b) found evidence that default increased with price, a necessary condition for non-price-rationing would have been discovered, but it would have said little about the overall severity or incidence of non-price rationing. Observational studies of existing credit markets show that non-price rationing largely operates through pre-emptive self-rationing. That is, individuals who recognize that they will almost surely be denied credit do not bother to go through a costly application process. Gauging the size of this group of individuals, who of course do not show up on lenders' client lists, would require a different approach.

At a somewhat deeper level, it is a bit hard to know how to interpret random interest rate variation. Building off the same South African experiment, Karlan and Zinman (2008) used their random interest rate variation to estimate the price elasticity of credit demand. Interestingly, they found a kink in the demand at the existing market interest rate. Increases above that level reduced demand, but reductions did not 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. 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. This uncontrolled treatment can result in what we called 'faux randomization', and raises a serious issue of interpretation. Unlike medical experiments, human agency and understanding can confound the use of RCT to study economic problems.

### *Randomized Liquidity Injections*

While RCT methods still have a ways to go before they can contribute much to our understanding of non-price-rationing and access to capital (questions 1 and 2 above), an RCT approach that randomly allocated increments of liquidity would seem useful as a way to explore question 3, the cost of liquidity constraints, and the returns associated with relaxing them. In an ambitious project, de Mel, Woodruff, and McKenzie (2008) do

exactly that, randomly providing liquidity gifts to Sri Lankan entrepreneurs. Employing the standard reduced form statistical methods found in the RCT literature to examine the impacts on firms' capital stocks and profits, these authors obtained mixed results: they found significant effects on capital stock, but marginally or insignificant effects of firm profitability (depending on the exact treatment).

In the context of the broader literature, these mixed results are not really surprising, as the de Mel, Woodruff and McKenzie (2008) experiment, despite its randomized variation in liquidity, essentially replicates the methods of the naïve, 1980s econometric literature and assumes that all households are in the credit-constrained regime, expression (3a) above. Put differently, their method ignores the essential heterogeneity implied by alternative excess demand regimes. More pointedly, while their results do tell us what to expect on average from a random distribution of liquidity, they do not identify any policy-relevant parameters, as they say nothing about what the impact would be of an expansion of credit markets in which selectivity based on demand and supply matters. Instead, they give us an unknown, data-weighted average of the impacts of liquidity on those with and without excess demand for capital. By failing to take into account the economic structure of the problem, their results are not only an unreliable predictor of what the relaxation of credit constraints might bring in Sri Lanka, they also are of dubious external validity, as even a random allocation in another environment might yield radically different results if the mix of constrained and unconstrained entrepreneurs were different.

One can imagine, however, a more structural approach to the de Mel, Woodruff and McKenzie (2008) experimental data in which appropriate information and/or econometric methods were used to distinguish households based on their actual constraint regime. Had de Mel, Woodruff and McKenzie (2008) been positioned to answer the first two capital access questions highlighted above (degree and incidence of liquidity constraints), they might have been able to undertake this approach. Such a mixed approach would require a retreat from purely RCT methods (especially as constraint regimes are likely correlated with difficult-to-observe individual attributes), but it would arguably be more informative. We see perhaps here the costs that occur when an RCT approach ignores the broader literature and its lessons already learned.

Karlan and Zinman (2009a) also try to create exogenous variation in liquidity increments and in principal come closer to identifying policy-relevant parameters. For their study, the authors worked with the South African paycheck lender used in their price variations studies 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. By focusing only on households that reveal credit demand by applying for loans, Karlan and Zinman (2009a) avoid some of the naiveté of de Mel, Woodruff and McKenzie (2008) and the pre-1990s econometric literature.

While perhaps promising as an approach, difficulties with the Karlan-Zinman (2009a) study illustrate several intrinsic difficulties of implementing RCTs with real economic institutions. First, unlike the de Mel, Woodruff and McKenzie (2008) cash gifts, 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 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.

A second problem revealed by the Karlan and Zinman (2009a) experiment is substantial non-compliance by the lender with the randomization scheme. In 47% of the cases in the experiment, loan officers refused to de-ration applicants in accordance with the protocol. De-rationing was thus anything but random. While one could imagine an incentive scheme that would have induced loan officers to follow the de-rationing protocol, the problems confronted by Karlan and Zinman (2009a) illustrate the problems of working with real economic institutions and actors who exercise agencies, especially when trials are not double-blinded.

In summary, by jettisoning the considerable knowledge acquired over decades of theoretical and empirical research based on observational data, these RCTs fall well short of answering the fundamental questions about capital access for the poor. Indeed, they have a long ways to go to catch up with what we already knew from the pre-existing literature. Doing better will likely require mixed methods that both take full advantage of the power of experiments and fully confront their pitfalls.

### *Purging the Error Term with Field Experiments*

While RCT methods attempt to deal with statistically problematic correlation between the error term and key variables like liquidity by randomizing the latter, an alternative approach is to try to purge the error term of the latent components that create this confounding correlation. While panel data methods can control for potentially time-invariant characteristics like intrinsic entrepreneurial ability, they may prove unsatisfactory if ability evolves through learning-by-doing processes, or through training, or if the returns to entrepreneurial ability increase over time through the introduction of new technologies or markets. In this context, direct measurement of traditionally latent characteristics would seem most useful.

The rapidly growing body of behavioral economics literature has developed games designed to reveal everything from risk aversion, to rates of time preference, to trust, to expectations, to entrepreneurial ability. There has been a recent proliferation of field experiments that seek to use experimental measures to statistically explain real world behavior. Karlan (2005) played trust games with microfinance borrowers in Peru and found that experimentally-measured trustworthiness predicts loan repayment rather well. While these methods have not yet been combined with observational

data and used to explore credit rationing and the impacts of changing it, they do suggest some paths forward.

*Changing the Structure and Rules of Access to Capital*

While many RCTs operate under tightly controlled, sometimes artificial conditions, a number of newly and recently launched projects work directly with financial firms to study the capital access impacts of innovative instruments designed for eventual full scale rollout. Relying on spatially randomized rollout strategies or encouragement designs, these studies are of the market rather than a synthetic construct that works in opposition to market rules and expectations. In addition to sidestepping some of the problems described above, by working with real market participants, these studies are informative about effectiveness, as opposed to abstract efficacy (see the discussion above).

While promising, these studies face a number of challenges, not the least of which is the pure complexity of working with real market participants. One example of this kind of work is microfinance credit reporting bureaus that link microfinance institutions (MFIs) and the conventional banking sector, potentially creating a ladder from high cost MFI credit to lower cost banking credit (de Janvry, Sadoulet, and McIntosh 2010). Utilizing a randomized rollout in the implementation and announcement of an MFI credit bureau in Guatemala, de Janvry, Sadoulet and McIntosh (2010) show that the creation and announcement of credit reporting has incentive effects on MFI borrower behavior. Determining whether the bureau ultimately improves the access to and cost of credit will require more work, likely including reliance on non-RCT methods to establish credit rationing status before and after intervention.

Giné, Goldberg and Yang (2010) is a second example of a study that examines the impact of information innovations on behavior in real credit markets. Working with a lender in rural Malawi, Giné, Goldberg and Yang (2010) explore the impact of fingerprinting technologies on loan repayment in an environment in which the absence of a strong national identity system otherwise makes it hard for lenders to employ dynamic incentives to assure loan repayment. The authors find that fingerprinting indeed boosts loan repayment rates, but only for those borrowers who would ex ante be predicted to have lower loan repayment rates. Similar to the de Janvry, Sadoulet and McIntosh (2010) study, determining whether or not this innovation really alters credit access and rationing will require additional work.

Finally, there are recent and newly initiated research projects that employ randomization methods to determine whether index insurance contracts – which attempt to remove covariant risk from the system – reduce both supply-side quantity and demand-side risk rationing in agricultural credit markets.[14] While the theoretical case for interlinking credit and insurance is strong (Carter, Cheng, and Sarris 2010), evidence is needed to determine whether index-based insurance can crowd out risk rationing and crowd in new sources of agricultural credit supply to smallholders, especially because these kinds of contracts are novel and

[14]*The BASIS Collaborative Research Support Program has launched an Index Insurance Innovation Initiative to support such projects. See http://i4.ucdavis.edu for details.*

intrinsically complex. Similar to RCTs that examine the impacts of information innovations, these studies will also require observational data to determine the ex ante constraint status of study participants.

*Advancing the Access to Capital Agenda*

In considering the power and pitfalls of randomization methods, access to capital is an especially interesting research area, both because of its intrinsic importance and because theory itself identifies essential response heterogeneity that simply cannot be ignored if generalizable, policy-relevant conclusions are to be drawn from the research. The implied complexity of projects that take ideas from theory to real market participants is daunting and time-consuming. Successful research in this area will almost surely require well-informed approaches that draw on the lessons of theory, observational data-based econometrics, and RCTs in order to be successful.

## Productivity Growth: Markets and Technologies

Significant, broad-based and sustained improvements in living conditions ultimately result from productivity improvements, whether these arise through technological or institutional change, gains from market-based exchange, or some combination thereof. The study of productivity growth, which includes questions such as, what ignites innovation, who adopts new innovations or enters new markets, when and why, and what is the distribution of gains from advances in productivity, has therefore always been central to development economics. But microeconomic processes of productivity growth are intrinsically subject to both placement effects – a technology, market or institution is appropriate to and available in a non-random subset of possible sites – and selection effects that complicate inference with respect to what factors or farmer characteristics induce increased uptake, thus leading to faster productivity growth. Experimental designs would seem to lend themselves well to obviating these problems. So how much are we now learning about productivity growth through RCTs and how much might we learn through more innovative efforts to integrate RCTs and behavioral field experiments into development economists' research designs?

When answering those questions, it is essential to bear in mind that technological and institutional innovation arises through a combination of scientific luck and deliberate processes induced by evolving profit incentives or by strategic investments provided by not-for-profit entities such as governments or philanthropists (Hayami and Ruttan 1985; Ruttan 1997). Bench scientists serendipitously discover how to introduce a valuable missing trait into an otherwise-attractive cultivar, firms invest in finding more efficient processes to save on increasingly scarce factors of production, and end users experiment with and adapt available methods, sometimes seemingly just for their own edification.[15] Since relatively little

[15]*For example, the System of Rice Intensification (SRI), a novel rice production method that sharply increases yields without any new purchased inputs (Barrett et al. 2004), was developed by a missionary priest in Madagascar who had trained many years earlier as an agronomist. Fr. Henri de Laulanié ran experiments on his home plots as much to maintain his scientific skills as to uncover the*

innovation is undertaken or even funded by agencies that commission careful impact evaluation studies before the diffusion of a new discovery begins, the opportunities to use RCTs to study productivity growth are necessarily limited. Hence, the vast majority of microeconomic research on the patterns and impacts of productivity growth has necessarily relied on observational data collected after an institutional or technological innovation has begun diffusing or a new market has emerged, although there are valuable opportunities to study pilot efforts experimentally.[16]

In spite of the limited scope available to study the origin and diffusion of innovations experimentally, more could be done using experiments, perhaps especially to inform research prioritization. As Schultz (1964) pointed out long ago, small-scale farmers are typically "poor but efficient". Economists can help policy-makers, business leaders, farming communities and poor households determine how best to stimulate innovation that benefits less well-off households by making new technologies, institutions and markets available to them. We can also help establish which innovations are likely to yield the greatest returns, whether measured in terms of increased economic surplus, poverty reduction, or some other outcome indicator.

One underused method is to ask the intended beneficiaries what change they would most value. Behavioral and environmental economists have long employed a variety of methods for rigorously eliciting the valuation of characteristics not (yet) available in the market. Experimental methods have become especially popular and show real promise for applications to technology and institutional development in developing countries. For example, Lybbert (2005) ran experimental games in south India that established, contrary to prevailing beliefs among crop breeders, that farmers valued higher mean yield growth far more highly than reduced downside yield risk or yield stability, and that this pattern is surprisingly unaffected by household wealth or risk exposure. Given the vast sums spent on agricultural, medical and other forms of research intended to generate innovations to ameliorate problems faced by poor households, there would seem considerable scope for more of this sort of rigorous, experimental elicitation of the research priorities of the poor.

### *Understanding Uptake and Participation*

Once a new technology, market or institution has emerged, understanding its diffusion is central to identifying behavioral responses that lead to productivity gains, as well as the distribution of welfare benefits from the innovation. Econometric problems bedevil much of the literature on technology adoption (Besley and Case 1993). This is partly due to unobserved heterogeneity in individual attributes not readily gathered in conventional surveys (e.g., risk and time preferences, skill) but that can be elicited through behavioral experiments. Unfortunately, there are few such studies to date in the developing world.[17]

*dramatically yield-increasing suite of techniques he ultimately discovered, according to colleagues we spoke with at Tefy Saina, the SRI extension group he founded.*

[16] *Ashraf, Giné, and Karlan (2009) offer an interesting example of an experimental design to uncover the welfare and crop choice effects of a project in Kenya that attempted to stimulate smallholder entry into high-value export crop markets.*

[17] *Engle-Warnick, Escobar, and Laszlo (2007); and Liu (2010) are rare exceptions.*

Selection problems also arise due to heterogeneity in environmental conditions and observable individual characteristics such as wealth, labor or land endowments, educational attainment, location, social networks, etc. An enormous literature has therefore explored these determinants of the heterogeneous and incomplete adoption patterns of new agricultural technologies exhibited by developing country farmers.[18] Inter-household heterogeneity in transactions costs, social connections, wealth and other observable or elicitable characteristics also seem to explain much of the heterogeneity in smallholder market participation patterns (Barrett 2008; Barrett et al. 2010).

Just as with lending contracts, individuals' subjective perceptions of new markets and technologies can vary markedly due to various levels of confidence in, or understanding of the product, inter-household heterogeneity in available alternatives, differences in the social networks and the presence or absence of external change agents, etc. (Luseno et al. 2003; Moser and Barrett 2006; Barrett et al. 2010; Conley and Udry 2010; Maertens 2010). Much has been learned over time about, for example, eliciting subjective distributions (Manski 2004; Delavande, Giné, and McKenzie forthcoming) and identifying social network effects (Bramoullé, Djebbari, and Fortin 2009; Conley and Udry 2010; Santos and Barrett forthcoming). Combining these non-experimental methods with experiments can contribute significantly to advancing rigorous inference about what drives the uptake of productivity-enhancing innovations. As with the use of behavioral field experiments to elicit otherwise unobservable individual characteristics, however, such integration of methods remains strikingly underdeveloped in the literature.

Given the considerable endogeneity and selection problems intrinsic to technology adoption and market participation questions, there is considerable potential for RCTs to help improve causal inference. But, as in the literature on capital access, in order to be useful RCTs must build on and integrate pre-existing knowledge generated non-experimentally, and must accommodate the essential heterogeneity one would naturally expect in response to new innovations.

Consider the case of inorganic fertilizer uptake by small maize farmers in western Kenya. Duflo, Kremer, and Robinson (2008; 2009) ran a series of RCTs over several seasons to explore why small farmers typically do not purchase and apply mineral fertilizers in spite of extension service recommendations to do so and apparent high average marginal returns to fertilizer application. Ultimately, the authors find that the average returns to fertilizer use are indeed quite high and conclude that differences in farmers' rate of time preference explain variation in fertilizer purchase behavior. Other researchers, using non-experimental methods, however, establish convincingly that many maize farmers in this same area face low returns to fertilizer, even if the average returns significantly exceed the cost of purchase (Marenya and Barrett 2009b; Suri forthcoming).

What accounts for the discrepancy? One likely reason is that Duflo, Kremer, and Robinson (2008; 2009) overlook perhaps the most obvious and longstanding explanation the soil science literature offers: that crop yield response to fertilizer – and thus its profitability – depends on ex

[18] *Feder, Just, and Zilberman (1985) offer a thoughtful, albeit now dated, survey of the Green Revolution era evidence.*

ante soil conditions. Marenya and Barrett (2009a, b), studying similar western Kenyan maize farmers to those in the Duflo, Kremer, and Robinson (2008; 2009) experiments, establish that the returns to fertilizer application increase sharply (and nonlinearly) with soil organic matter, that the poorest households are most likely to cultivate the lowest quality soils offering the lowest marginal returns, and that differences among plots and farms in soil organic matter explain significant variation in both the returns to fertilizer, as well as its purchase and application. If poverty affects time preferences, as is widely believed, the Duflo, Kremer, and Robinson (2008; 2009) results could be entirely the result of heterogeneous land quality.

As this example illustrates, it matters less whether one uses an experimental design or pays attention to the prior non-experimental literature and takes the time to measure variables that cause essential heterogeneity in anticipated responses and impacts. Experiments have powerful potential to correct for selection effects (including in the Marenya and Barrett (2009 a, b) studies), but not if methodological hubris induces researchers to ignore other literatures that may not meet the RCT litmus test and to ignore common problems of essential heterogeneity that can be anticipated based on those literatures.

One promising experimental approach involves encouragement designs (also called "randomized outreach") based, for example, on the randomized distribution of additional information on the innovation or market, or of discounts for the purchase of new products. This provides a credible instrument for identifying an "intent to treat" effect in an environment of incomplete, endogenous adoption or participation ("noncompliance"). Beyond this impact evaluation benefit, however, encouragement designs based on discounts for commercially distributed innovations (e.g., fertilizer, index insurance) also permit the identification of crucial price elasticity parameters that are otherwise difficult, at best, to identify given the limited price variation typically observed in a new product's pilot phase. Since pricing can greatly affect uptake, encouragement designs are an example of an experimental method that can help generate credible identification not only of causal impacts, but also of a key elicitable characteristic (the price elasticity of demand) that matters for prescriptive analysis: how should agro-dealers price fertilizer or underwriters price insurance, and what, if any, subsidy level most effectively stimulates uptake? New experimental designs offer the opportunity to answer these sorts of questions more reliably and quickly than is feasible with standard observational data.

Perhaps the biggest problem plaguing studies of diffusion of innovations and market access revolves around the unavoidable tension between external validity and the ethics of violating the targeting principle. Technologies naturally diffuse first to areas where they offer the highest returns; firms likewise naturally seek out locations that offer the cheapest, most reliable suppliers. Experimental designs that ignore these inherent placement effects and attempt to introduce innovations or new markets in randomly selected areas waste scarce resources; indeed, they violate the targeting principle. But if researchers study innovations where and when they occur – whether experimentally or not – serious external validity concerns arise. For example, Ashraf, Giné, and Karlan's (2009) experimental study of the crop choice and welfare impact of an

intervention to help smallholder farmers access high-value export markets generates convincing estimates of the effect in the specific part of the Kirinyaga District, a high agronomic potential area with good access to the international airport outside Nairobi. However, those estimates do not project to other areas, especially dissimilar ones.

Furthermore, much depends on the real institutions that condition access to information, inputs (e.g., improved seeds or fertilizer), and contracts. Because institutional performance is heterogeneous and the complexity of experimental trials naturally induces researchers to work with relatively effective field partners, this introduces yet another important source of external validity problems in experimental studies of productivity growth. RCT estimates of impacts in small trials run by unusually competent field agencies almost surely overstate the effects one can reasonably anticipate from more complex, spatially-dispersed, large-scale programs that engage an array of partners of varied – and average lower – quality.

Unless they egregiously violate the targeting principle, experimental methods cannot overcome the placement problem intrinsic to technology adoption (Besley and Case 1993), smallholder access to emerging value chains (Barrett et al. 2010), institutional change, or other sources of productivity growth. By contrast, large-scale representative surveys with retrospective reconstruction of histories of, for example, technology adoption or market entry, can address this problem if credible instruments exist to identify the placement effects (Moser and Barrett 2006; Michelson 2010). If one wants to understand the larger-scale welfare impacts of a new innovation, the external validity problem poses a serious challenge.

### *Estimating the Welfare Effects of Productivity Growth*

Most development economists' interest in productivity growth is not intrinsic; rather, it is instrumental. We are interested in productivity growth because of its expected impacts on welfare indicators such as incomes, nutritional status or asset holdings. But the pathway from innovation through diffusion to induced welfare change is a complex one and heavily dependent on general equilibrium effects. As such, the role of RCTs in studying the welfare effects of smallholder productivity interventions is necessarily limited. While behavioral games can quite effectively reproduce strategic interactive behaviors among a small number of agents around a small number of actions, experimental designs cannot yet handle the complex, multi-agent, multi-sectoral interactions required to effectively reproduce realistic general equilibrium effects from exogenous treatments.

To be sure, well-executed RCTs can do a reasonably good job of establishing the short-term, direct gains of, for example, technology adoption to adopting households compared to those not adopting. But RCTs and behavioral experiments are ill-suited to capturing the indirect effects that arise through equilibrium shifts in input (e.g., labor) or output (e.g., food) markets. And history tells us that this is where most of the welfare gains to innovation accrue, as consumer surplus due to lower prices, rather than in producer surplus, and as gains in real employment levels and real wages (Evenson and Gollin 2003; Minten and Barrett 2008). Ironically, the RCT methods so widely championed today for rigorous impact evaluation are intrinsically ill-suited to estimating the magnitude or distribution of

welfare gains resulting from arguably the greatest driver of economic development – technological change.

Despite the long history and central place of productivity growth research in development economics, experimental methods remain underexploited. There are many good reasons for this: natural limitations arising from the often-random nature of discovery and diffusion; considerable essential heterogeneity issues; the intrinsic tradeoff between external validity and violation of the targeting principle; and the centrality of general equilibrium effects to establishing welfare impact estimates. There remains untapped power in RCTs and behavioral field experimental methods for productivity research, perhaps especially in the elicitation of targeted beneficiaries' preferences for innovations that might guide research prioritization, and of key behavioral parameters that condition the uptake of innovations, as well as the use of encouragement designs to help evaluate and inform the commercial distribution of innovations and prospective subsidy policies. But productivity research is an area where methodological heterodoxy should and will remain the norm; there is no solid case for the preferential use of RCTs, as is currently favored by some donors and academics.[19]

## Conclusions

Randomized controlled trials have, appropriately, come to play an important, even essential role in development economics. But we are troubled by what seems an increasingly naïve promotion of RCTs, as if there exists any implementable method that generates "perfect identification". Like methods based on observational data, RCT designs rarely emerge unscathed from their encounter with ethically and institutionally complex field settings populated by willful research implementers and subjects. Furthermore, even well-designed and executed RCTs often cannot generate useful estimates of the parameters of greatest relevance or in a sufficiently general way so as to be useful, especially on their own, absent of explicit interlinkage with more general theorizing and the broader body of (largely non-experimental) empirical results. RCTs offer an important new tool that development economists need to embrace enthusiastically, but with eyes wide open to their limitations.

Development economics requires healthy methodological pluralism that recognizes all identification strategies' intrinsic limitations. The best way to generate useful new insights to help answer the most pressing questions about improving the human condition is through the creative combination of evidence generated by various methods deployed so as to exploit each method's comparative advantage and to minimize its overreach. In addition to RCTs, behavioral field experiments can play a role which is not yet fully appreciated. By eliciting credible estimates of otherwise unobservable parameters that matter to observable behavioral and welfare outcomes, the fruits of behavioral field experiments can accomplish much of what RCTs set out to do: render tolerable the unconfoundedness assumption upon which credible identification depends.

[19] *Maredia (2009) comes to similar conclusions in exploring the feasibility of using experimental designs for impact evaluation of investments in agricultural research by the CGIAR.*

Experimental design is extremely useful for generating exogenous variation in variables of interest. But often it can come at unacknowledged costs, including violations of basic principles of research ethics, faux exogeneity that compromises even the internal validity of estimates, and the failure to tap prior knowledge built using other, similarly-imperfect tools. All research methods, including those based on randomization under an experimental design, have shortcomings. There is no gold standard of perfect identification, no single best way to acquire knowledge. Indeed, the various conceptual, ethical, logistical and statistical concerns about experiments that we enumerate in this paper should remind us that all that glitters is not gold. Researchers need to beware of the blind pursuit of exogenous variation, lest it render development economics irrelevant in the face of pressing global challenges of poverty and growth.

## Acknowledgement

We thank Jorge Aguero, Marc Bellemare, Steve Boucher, John Hoddinott, Travis Lybbert, Annemie Maertens, Jordan Matsudaira, Craig McIntosh, Hope Michelson, Carl Nelson, Agnes Quisumbing, Kazushi Takahashi, Bruce Wydick and an anonymous reviewer for helpful comments on an earlier draft, and Ivi Demi and Ian Sheldon for editorial help. Any remaining errors are the authors' sole responsibility.

## References

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

Adams, Dale. 1988. The Conundrum of Successful Credit Projects in Floundering Rural Financial Markets. *Economic Development and Cultural Change* 8(2): 347–366.

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. 1990. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *American Economic Review* 80(3): 313–335.

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.

Ashraf, Nava, Xavier Giné, and Dean Karlan. 2009. Finding Missing Markets (and a Disturbing Epilogue): Evidence from an Export Crop Adoption and Marketing Intervention in Kenya. *American Journal of Agricultural Economics* 91(4): 973–990.

Banerjee, Abhijit, and Esther Duflo. 2008. The Experimental Approach to Development Economics. CEPR Discussion Paper No. DP7037.

——. 2010. Giving Credit Where Credit is Due. Unpublished manuscript, Massachusetts Institute of Technology.

Barrett, Christopher B., ed. 2005. *The Social Economics of Poverty: On Identities, Groups, Communities and Networks.* London, UK: Routledge.

——. 2008. Smallholder Market Participation: Concepts and Evidence from Eastern and Southern Africa. *Food Policy* 33(4): 299–317.

Barrett, Christopher B., Maren E. Bachke, Marc F. Bellemare, Hope C. Michelson, Sudha Narayanan, and Thomas F. Walker. 2010. Smallholder Market Participation in Agricultural Value Chains: Comparative Evidence from Three Continents. Unpublished manuscript, Cornell University.

Barrett, Christopher B., Michael R. Carter, and C. Peter Timmer. 2010. A Century-Long Perspective on Agricultural Development. *American Journal of Agricultural Economics* 92(2): 447–468.

Barrett, Christopher B., Christine M. Moser, Oloro V. McHugh, and Joeli Barison. 2004. Better Technology, Better Plots or Better Farmers? Identifying Changes In Productivity And Risk Among Malagasy Rice Farmers. *American Journal of Agricultural Economics* 86(4): 869–888.

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.

Bell, Clive, T.N. Srinivasan, and Christopher Udry. 1997. Rationing, Spillover, and Interlinking in Credit Markets: The Case of Rural Punjab. *Oxford Economic Papers* 49 (4): 557–585.

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.

Besley, Timothy, and Anne Case. 1993. Modeling technology adoption in developing countries. *American Economic Review Papers and Proceedings* 83(2): 396–402.

Binswanger, Hans P. 1980. Attitudes Toward Risk: Experimental Measurement in Rural India. *American Journal of Agricultural Economics* 62(3): 395–407.

——. 1981. Attitudes Toward Risk: Theoretical Implications of an Experiment in Rural India. *Economic Journal* 91(364): 867–890.

Boucher, Steve, Catherine Guirkinger, and Carolina Trivelli. 2009. Direct Elicitation of Credit Constraints: Conceptual and Practical Issues with an Application to Peruvian Agriculture. *Economic Development and Cultural Change* 57(4): 609–640.

Boucher, Steve, Michael R. Carter, and Catherine Guirkinger. 2008. Risk Rationing and Wealth Effects in Credit Markets: Theory and Implications for Agricultural Development. *American Journal of Agricultural Economics* 90(2): 409–423.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. Identification of peer effects through social networks. *Journal of Econometrics* 150(1): 41–55.

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.

Card, David. 1990. The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial and Labor Relations Review* 43(2): 245–257.

Cardenas, Juan C. 2009. Experiments in Environment and Development. *Annual Review of Resource Economics* 1(1): 157–183.

Carter, Michael R. 1989. The Impact of Credit on Peasant Productivity and Differentiation in Nicaragua. *Journal of Development Economics*. 31(1): 13–36.

Carter, Michael R., and Marco Castillo. 2005. Coping with Disaster: Morals, Markets and Mutual Insurance: Using Economic Experiments To Study Recovery From Hurricane Mitch. In Christopher B. Barrett, ed. *The Social Economics of Poverty: On Identities, Groups, Communities and Networks.* London, UK: Routledge.

Carter, Michael R., Lan Cheng, and Alexander Sarris. 2010. The Impact of Inter-linked Index Insurance and Credit Contracts on Financial Market Deepening and Small Farm Productivity. Unpublished manuscript, University of California, Davis.

Carter, Michael R., and Pedro Olinto. 2003. Getting Institutions Right for Whom? Credit Constraints and the Impact of Property Rights on the Quantity and

Composition of Investment. *American Journal of Agricultural Economics* 85(1): 173–186.
Conley, Timothy G., and Christopher R. Udry. 2010. Learning About a New Technology: Pineapple in Ghana. *American Economic Review* 100(1): 35–69.
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.
de Janvry, Alain, Elisabeth Sadoulet, and Craig McIntosh. 2010. The Supply and Demand Side Impacts of Credit Market Information. *Journal of Development Economics* 93(2): 173–188.
Delavande, Adeline, Xavier Giné, and David McKenzie. (forthcoming). Eliciting Probabilistic Expectations with Visual Aids in Developing Countries: How Sensitive are Answers to Variations in Elicitation Design? *Journal of Applied Econometrics.*
Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. Using Randomization in Development Economics Research: A Toolkit. In *Handbook of Development Economics*, T. P. Schultz, and J. Strauss, eds., 3895–3962 Amsterdam: Elsevier.
Duflo, Esther, Michael Kremer, and Jonathan Robinson. 2008. How High are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya. *American Economic Review Papers and Proceedings* 98(2): 482–488.
——. 2009. Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. Working paper 15131, National Bureau of Economic Research.
Engle-Warnick, Jim, Javier Escobar, and Sonia Laszlo. 2007. Ambiguity Aversion as a Predictor of Technology Choice: Experimental Evidence from Peru. Unpublished manuscript, McGill University.
Evenson, Robert E., and Douglas Gollin, eds. 2003. *Crop Variety Improvement and Its Effect on Productivity: The Impact of International Agricultural Research.* Wallingford, UK: CABI.
Feder, Gershon, Richard E. Just, and David Zilberman. 1985. Adoption of Agricultural Innovations in Developing Countries: A Survey. *Economic Development and Cultural Change* 33(2): 255–298.
Feder, Gershon, Lawrence Lau, Justin Lin, and Xiaopeng Luo. 1990. The Relationship Between Credit and Productivity in Chinese Agriculture: A Microeconomic Model of Disequilibrium. *American Journal of Agricultural Economics* 72(5): 1151–1157.
Giné, Xavier, Jessica Goldberg, and Dean Yang. 2010. Identification Strategy: A Field Experiment on Dynamic Incentives in Rural Credit Markets. Unpublished manuscript, University of Michigan.
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.
Harrison, Glenn W., and John A. List. 2004. Field Experiments. *Journal of Economic Literature* 42(4): 1009–1055.
Hayami, Yujiro, and Vernon W. Ruttan. 1985. *Agricultural Development: An International Perspective.* Baltimore, MD: Johns Hopkins University Press.
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.

——. 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.
Hoffmann, Vivian, Christopher B. Barrett, and David R. Just. 2009. Do Free Goods Stick to Poor Households? Experimental Evidence on Insecticide Treated Bednets. *World Development* 37(3): 607–617.
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.
Imbens, Guido W., and Jeff M. Wooldridge. 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* 47(1): 5–86.
Karlan, Dean. 2005. Using Experimental Economics to Measure Social Capital and Predict Real Financial Decisions. *American Economic Review* 95(5): 1688–1699.
Karlan, Dean, and Jonathan Zinman. 2008. Credit Elasticities in Less Developed Countries: Implications for Microfinance. *American Economic Review* 98(3): 1040–1068.
——. 2009a. Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts. *Review of Financial Studies* doi:10.1093/rfs/hhp092
——. 2009b. Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment. *Econometrica* 77(6): 1993–2008.
Kochar, Anjini. 1998. An Empirical Investigation of Rationing Constraints in Rural Markets in India. *Journal of Development Economics* 53(2): 339–372.
Krueger, Alan B. 1999. Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics* 114(2): 497–532.
LaLonde, Robert J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review* 76(4): 604–620.
Leamer, Edward E. 1983. Let's Take the Con Out Of Econometrics. *American Economic Review* 73(1): 31–43.
——. 2010. Tantalus on the Road to Asymptotia. *Journal of Economic Perspectives* 24(2): 31–46.
List, John A., Sally Sadoff, and Mathis Wagner. 2010. So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design. National Bureau for Economic Research Working Paper No. w15701.
Liu, Elaine M. 2010. Time to Change What to Sow: Risk Preferences and Technology Adoption Decisions of Cotton Farmers in China. Unpublished manuscript, University of Houston.
Luseno, Winnie K., John G. McPeak, Christopher B. Barrett, Getachew Gebru, and Peter D. Little. 2003. The Value of Climate Forecast Information for Pastoralists: Evidence from Southern Ethiopia and Northern Kenya. *World Development* 31(9): 1477–1494.
Lybbert, Travis J. 2005. Indian Farmers' Valuation of Yield Distributions: Will Poor Farmers Value 'Pro-poor' Seeds? *Food Policy* 31(5): 415–441.
Maertens, Annemie. 2010. Social Networks, Identity, and Economic Behavior: Empirical Evidence from India. PhD dissertation, Cornell University.
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.

Manski, Charles. 2004. Measuring Expectations. *Econometrica* 72(5): 1329–76.
Maredia, Mywish. 2009. The Scope and Feasibility of Using Experimental Designs in Evaluating Impacts of Investments in Agricultural Research for Development. Paper prepared for the CGIAR Standing Panel on Impact Assessment.
Marenya, Paswel P., and Christopher B. Barrett. 2009a. Soil Quality and Fertilizer Use Among Smallholder Farmers in Western Kenya. *Agricultural Economics* 40(5): 561–572.
——. 2009b. State-conditional Fertilizer Yield Response on Western Kenyan Farms. *American Journal of Agricultural Economics* 91(4): 991–1006.
Michelson, Hope. 2010. Welfare Effects of Supermarkets on Developing World Farmer Suppliers: Evidence from Nicaragua. Unpublished manuscript, Cornell University.
Miguel, Edward, and Michael Kremer. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72(1): 159–217.
Minten, Bart, and Christopher B. Barrett. 2008. Agricultural Technology, Productivity, and Poverty in Madagascar. *World Development* 36(5): 797–822.
Moser, Christine M., and Christopher B. Barrett. 2006. The Complex Dynamics of Smallholder Technology Adoption: The Case of SRI in Madagascar. *Agricultural Economics* 35(3): 373–388.
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.
Ruttan, Vernon W. 1997. Induced Innovation, Evolutionary Theory, and Path Dependence: Sources of Technical Change. *Economic Journal* 107(444): 1520–1529.
Santos, Paulo, and Christopher B. Barrett. (forthcoming). Identity, Interest, and Information Search in a Dynamic Rural Economy. *World Development*.
Schultz, T. Paul. 2004. School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics* 74(1): 199–250.
Schultz, Theodore W. 1964. *Transforming Traditional Agriculture*. New Haven, CT: Yale University Press.
Sial, Maqbool, and Michael R. Carter. 1996. Is Targeted Small Farm Credit Necessary? A Microeconometric Analysis of Capital Market Efficiency in the Punjab. *Journal of Development Studies* 32(5): 771–798.
Suri, Tavneet. (forthcoming). Selection and Comparative Advantage in Technology Adoption. *Econometrica*.