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

Christopher B. Barrett
Cornell University

Michael R. Carter
University of California, Davis

September 2010

Abstract

Development economists increasingly employ experimental methods in seeking to answer key questions that inform development policies and programs. Impact evaluation based on randomized controlled trials (RCTs) has in particular fundamentally altered the discourse of development economics over the past decade. While these methods have power, their use in development economics facing a number of intrinsic and extrinsic problems, including ethical dilemmas, faux exogeneity and what we term the ingénue effect. We illustrate these points with concrete examples concerning capital access and productivity growth, and argue for greater use of behavioral experiments that can help resolve the identification problems that motivate RCTs by eliciting otherwise unobservable, but fundamental economic characteristics such as rates of time preference and risk aversion. In the end, economic development and development economics will be better served if we can arrive at a balanced appreciation of the essential but limited role of RCTs and experiments.

1. Introduction

The challenge of doing development economics, and doing it well,[1] motivates the constant search for new methodological responses—economists, like most social scientists, are 'methods junkies.' 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 estimators), RCTs have engendered a tremendous amount of, often vitriolic, controversy. Our goal in this paper is to pick through this controversy and derive a balanced perspective on how development economics can continue to advance and contribute to the understanding of real world problems whose importance transcends that of methodological affinity.

Much of the controversy surrounding RCTs is undoubtedly an artifact of its advocates proclaiming the RCT as the "gold standard" of evidence, marking the apparent end of methodological history.[2] Arguments that the RCT evidence is of baser metal have been provoked by advocates' epistemological claims (*e.g.*, see Basu 2005, Ravallion 2009, Deaton 2010). Our goal here is not to rehash the arguments of Deaton (2010) and others concerning the validity of RCT impact estimates. Instead, starting from the perspective that exogenous variation

[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 understand better 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.

[2] While there are certainly devotees of the Heckman estimators, there did not emerge an identifiable group of 'Heckmanistas', nor of 'Panelistas,' as those who use these methods do not try to pretend that they decisively resolve identification problems.

in policy or control variables can certainly be statistically useful (as Imbens 2010 points out), we highlight some less discussed, but no less important, limitations of RCTs as applied to economic development problems.

We will argue that some of these limitations are intrinsic to the methodology when applied to economic problems, while others are extrinsic limitations that can be shed once RCTs are 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 literatures, we conclude with thoughts on a balanced, problem-centric approach to development economics. Among other points, we advocate greater utilization of behavioral economic experiments to help resolve the same fundamental identification problem that motivates reliance on RCTs.

To help frame the discussion that follows, consider the following stylized simultaneous equation model that empirical development economics confronts:

$$\begin{aligned} b &= g(y,p,s,\pi,\varepsilon_b) \\ y &= f(b,p,s,\pi,\varepsilon_y)\text{'} \end{aligned}$$

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. They 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, π , (*e.g.*, trustworthiness, ambition, time and risk preferences), and of course on classical measurement, sampling and specification errors, ε_b and ε_y .

While the fact that the above two equations are simultaneous and of unknown functional form clearly poses a challenge for statistical identification of 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 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 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 assumptions. RCTs attempt to circumvent these issues by randomizing the policy instrument, *p*, across the population in an effort to guarantee the orthogonality of *p* to π and *s*. The statistical advantage of such randomization is obvious and well-described by Banerjee and Duflo (2008). Perhaps less obvious are the limitations of RCTs. After briefly reviewing the well-known potential power of RCTs in Section 2, we will in Section 3 outline key intrinsic pitfalls in the use of RCTs in development economics, including ethical constraints and the problem of faux exogeneity. In Section 4, 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 (the *ingénue* effect). Section 5 will then briefly describe the potential power of behavioral experiments to aid reliable statistical identification. Sections 6 and 7 then illustrate the use and abuse of RCTs in two literatures—that on access to capital, and that on technological change in agriculture. Section 8 concludes the paper with thoughts on a methodological way forward.

2. **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 that has followed a parallel course, the laudable objective of RCTs is to resolve the serious econometric problems associated with program placement and selection effects – 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. IV 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 observed differences in the dependent variable, contaminating the resulting estimate of the "effect" of the program.

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, most particularly orthogonality to expected program benefits and to the error structure more generally.[3] RCTs have been deployed en masse by a range of scholars (including us), especially those associated with MIT's Poverty Action Lab, its sister organizations, and various donor organizations in answering a range of specific questions relating to education, health, finance, agriculture and other domains of interest to many development economists. A central tenet of the RCT movement is that economic theory is excessively limiting, that we must be open to surprises. Through clever experimental designs and this openness to letting the data lead the researcher, rather than vice versa, leaders of the RCT movement, 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 that has similarly championed healthy skepticism about many longstanding assumptions of neoclassical economic theory.

3. Intrinsic Pitfalls of Randomized Controlled Trials in Development Economics

RCTs in development economics are methodologically rooted in biomedical trials, as well as in basic scientific method as applied to 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 that responds endogenously following laws of nature.

[3] Ironically, concerns about the misuse of IV estimation methods with observational data have fuelled increasing misuse of RCTs and experimental data in spite of the common pitfalls they share. 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.

3.1 Faux Exogeneity and other Pitfalls of Internal Validity

In retrospect, the seminal deworming study of 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 subject. However, when randomization is used for bigger, more 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. Treatments are likely non-randomly distributed among experimental subjects whose capacity to comprehend and to act vary in subtle but substantive ways. This problem can create a certain "faux exogeneity" in many experiments as 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. Thus the unobserved heterogeneity problem that one seeks to remedy through use of randomization can creep back in (Heckman et al. 2006). 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, as crucially distinct from the treatment the experimenter wishes to apply.

A somewhat similar problem can result from the use of side payments designed to bolster voluntary uptake of a new program within a treatment group. While such payments may be absolutely essential if an RCT is to achieve any measure of statistical power, in the presence of essential heterogeneity (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 the subsidy built-in to the experiment to encourage uptake of the treatment condition.

Note that this is a fundamentally different problem than that confronted by medical research, which typically employs payments to encourage participation. Participants in medical studies presumably have no idea whether their particular biological system will respond more or less favorably to a treatment than the system of the average person. We would not therefore expect that higher payments would bring in people who know that they will benefit less from the treatment. In contrast, many economic interventions (e.g., access to a new financial contract or technology) depend precisely on participants understanding and evaluating the returns to the new treatment. Mullally *et al.* (2010) illustrate this problem and the bias it imparts to estimated average treatment effects, using an encouragement design employed to evaluate an agricultural insurance program in Peru.

Faux exogeneity and encouragement bias in economic studies undercut the 'gold standard' claim that RCTs reliably identify the average treatment effect for the target population (i.e., that RCT estimates have internal validity). Just as the original gold standard depended on a range of strong assumptions – that 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 underlying identifying assumptions before accepting RCT results as internally valid.

Beyond these two problems, Heckman (1992) and Deaton (2010) discuss a variety of other statistical limitations to the internal validity of RCT estimates that merit mention. First, 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 too rarely confirm that control and treatment groups exhibit identical distributions of observable variables.

Given the likelihood of randomization bias, experimental approaches need to take special care to balance control and treatment groups based on observables. But there is no standard practice on how best to 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 randomization method used are commonly incorrect, leading researchers to incorrect inferences about treatment effects.

The attractive asymptotic properties of RCTs often disappear in practice, much like the asymptotic properties of other IV estimators. 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' aim to cleanly identify causal effects. This routinely happens in selecting survey respondents for observational studies as well, thus the problem has long been quietly accepted within the profession as an inevitable imperfection in data collection. But in the pre-RCT research environment, this was not a fatal flaw. The stakes are higher when the randomization is itself the source of identification. These dirty little 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.

In summary, experiments are invaluable tools for biophysical scientists, where the mechanisms involved are more, well, 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 and ecological literatures on experiments. Perceptions, comprehension, preferences, and subtle effects on material and non-material incentives heavily influence human response to experimental interventions. It is often unclear what is getting changed beyond the variable the research is intentionally randomizing. Hawthorne effects are but one well-known example. As a result, impacts and behaviors elicited experimentally are commonly endogenous to environment and structural conditions that do not vary in known ways within a necessarily highly-stylized experimental design.

3.2 External Validity of Randomized Controlled Trials

The most widespread critique of experimental evidence revolves around the external validity of results (see, for example, Rodrik 2008, Acemoglu 2009, Ravallion 2009, Deaton 2010). In brief, the problem is that unobservable and observable features inevitably vary at community level and cannot be controlled for in experimental design because 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 essential unobserved heterogeneity within sample (Heckman et al. 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. There is a nontrivial probability that there is no external population to whom the results of the experiment apply on average.

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 us 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 entirely be 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. While any micro 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.

3.3 Choosing Amongst Alternative Interventions

RCTs typically aim to establish the treatment effect of an intervention against the counterfactual that no intervention occurs. 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 then added to the theory and observational evidence so as to gradually generate a set of prescriptions.

Not so in economics, where we routinely teach undergraduates the fundamental importance of opportunity costs to explaining and informing choices, but routinely ignore our own lessons in trumpeting binary experimental results. While perhaps RCT methods in development economics will eventually build up a sufficient body of evidence to underwrite advocacy of one treatment versus another, that body of evidence does not yet exist and will not exist for many years, at best. Yet economists leap to make immodest policy and project recommendations from one-off results that lack credible assessments of the opportunity cost of an intervention.

Deworming offers a shining example of this problem. Clever, heavily cited field experiments by talented development economists (Miguel and Kremer 2004) have been marketed into global solutions to literally "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 our completely unscientific 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 concerns worth investing in, and only two put it in the top five. Rather, the emphasis was solidly on early childhood (including prenatal) nutrition, immunization against infectious disease, and breastfeeding (in part as a combination of the prior two).[4]

3.4 Ethical Constraints

Even if experimental studies enjoyed complete internal and external validity and could be designed to address important questions with a sufficient range of alternative actions that they really informed choice among multiple alternatives, there would remain occasions when it would be unethical to employ experimental designs. We see three broad classes of ethical dilemmas associated with experiments conducted by development economists. These get distressingly little attention in graduate training and in the literature.

The first and most obvious class of ethical dilemma revolves around 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 by 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 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

[4] 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).

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, 2004). Taking its cue from John D. Rockefeller who refused to give money to Alcoholics Anonymous on the grounds that money would undercut the organization's effectiveness, this paper explicitly sets out to see if 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 making 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 the social gains; (2) that subjects be fully informed of the risks; and, (3) that compensation be paid to cover any damages done. It remains unclear if these rules were met in this study, which is somewhat chilling given that the study indeed confirms that poor women were damaged by the injection of cash into randomly selected women's groups.

A 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 humans 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. In order 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 an example well-known to development economists. Even when the randomization is public and transparent, cluster randomization maintains the exogeneity of the intervention but at an ethically questionable cost of sacrificing the well-accepted right of each individual participant to informed consent as well as the corresponding researcher obligation 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.

The third class of ethical dilemma arises from abrogating the targeting principle that undergirds most development interventions. Given very scarce resources and the 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 eschewing exploitation of private information held by study participants, randomized interventions routinely treat individuals known not to need 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 implementers' breach of the 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 latter 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. Extrinsic Pitfalls of Randomized Controlled Trials

The prior section considered limitations of RCTs that are intrinsic to experimental methods when applied to economic problems. A second set of pitfalls has emerged because in promulgating RCTs some influential scholars delegitimize other ways of learning about the phenomena under study. This has in turn induced the most loyal 'randomistas' to ignore both crucial questions that are not amenable to randomization and prior evidence based on non-experimental methods, even when the pre-RCT literature adds essential insights, balance and relevance to the development economics enterprise.

4.1 Distortion of the Research Agenda

A key 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 (Rodrik 2008, Collier 2010). Nor are infrastructure issues or any other meso- or macro-scale intervention that cannot be replicated in large numbers and the placement of which is necessarily and appropriately subject to significant political economy considerations (Ravallion 2009). As one moves from smaller, partial equilibrium questions – which type of contract generates a greater response from a MFI institution's clientele? What is the marginal effect of cash versus food transfers on recipients' nutritional status? – to more substantive, general equilibrium and political economy questions, the present fixation on experimental evidence in development economics becomes an impediment rather than a help.

There is indisputably a place for experiments in development economics research. But much of the low-hanging fruit has been harvested and many of those fruit have proved disappointingly small. The big questions in development are rarely amenable to the RCT approach alone, as sections 6 and 7 of this paper illustrate. Unfortunately, the fashion for randomized controlled trials (RCTs) and social experimentation has distorted research agendas to focus on a narrow subset of descriptive analysis – evaluation – implicitly delegitimizing other, equally important analytical functions.[5] One symptom of this problem is that too many excellent students (and faculty!) at the best universities fritter away their considerable talents in a quest for exogenous variation, often to prove points utterly obvious to laypersons who should be the key audience informed by our most important findings.

Even many of the questions that RCT proponents attempt to ask may not be as amenable to experimental inquiry as they seem at first, superficial glance. Given the extended lag times inevitably involved in human capital accumulation, the effective use of randomized experiments to study the impact of alternative human capital interventions is rare as experiments are typically designed to answer short-term questions uncomplicated by the long passage of time. The INCAP studies of early childhood nutrition in Guatemala (Hoddinott et al. 2008, Maluccio et al. 2009) are a valuable exception, but even these studies suffer massive attrition problems that

[5] Consider, for example, the launch in 2009 of a new Social Science Research Network (SSRN) eJournal on "Randomized Social Experiments" with "the primary objective … to produce internally valid impact estimates" as a sisterpublication to the SSRN eJournal on "Development Economics". The narrowness of the former alarms us.

compromise the internal validity on which the claims of RCT enthusiasts are founded. This is, of course, ironic given the extensive research that has been done on education in developing countries using experimental methods.

4.2 The LATE May Miss the Point

Further, even for questions that are amenable to experimental approaches, RCTs ultimately only identify the local average treatment effect (LATE, Deaton 2010). But much of what is interesting in development economics transcends the unconditional mean, revolving 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 scale-up of seemingly successful interventions than do estimated mean treatment effects.

RCT studies focus on generating consistent and unbiased estimates of the LATE. 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 misspecification. These concerns apply to all research but seem especially overlooked in the current RCT fashion.

Indeed, much of the point in development economics is the essential heterogeneity of response. We want to understand the conditional nature of responses, not just the mean marginal response. 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 more refined form, and with adequate sample size, they may allow estimation of a LATE within quantiles or distinct subpopulation strata of interest. But often our interest is in the heterogeneous responses found within a population and the underlying structure that accounts for such heterogeneity. That is what drives the political economy of policymaking and the design of intervention strategies that require targeting.

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.   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 or project or technology.

Indeed, the endogenous processes that guide resource allocation by human agents (not subjects), whether by policymakers or individuals, can ultimately undermine the quest to eliminate endogeneity. The researcher who imposes exogenous allocations is not, in fact, replicating real human behavior. This is the crucial distinction made in epidemiology and public health between *efficacy* – the study of a treatment's capacity to have an effect, as established under fully controlled conditions – and *effectiveness* – the study of induced change under reallife conditions. Economists who seek to inform agents making real decisions in the uncontrolled real world ultimately need to be able to address questions of effectiveness, not merely efficacy. Overcorrection for endogeneity may, ironically, render findings consistently and unbiasedly irrelevant to the real-world questions concerning the intervention under study.

5. Behavioral Experiments

While RCTs have become the dominant method in contemporary development economics, another sort of experiment goes underexploited. Behavioral experiments seek to identify observable behaviors and elicitable characteristics through a research design that employs random assignment so as to identify causation via exogenous variation in the explanatory variable of interest. Lab experiments of the sort that historically predominate in behavioral economics are rare in development economics; within our subdiscipline 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).

A wide array of experiments have been conducted in developing countries for many years – dating back at least to Binswanger (1980, 1981) – in order to elicit defensible estimates of key risk and time preferences, indicators of trust, and other key behavioral parameters that give rise to the selection concerns that motivate so many RCTs. Behavioral experiments are especially important if we acknowledge the importance of social context, identity, norms, etc. (Barrett 2005, Cardenas 2009). Invoking the notation we used earlier, these sorts of experiments are useful where the ε_b and ε_y error terms are almost surely correlated with the s variables in the absence of proper control for the π behavioral parameters, leading to unobserved heterogeneity problems in standard econometric analysis based on observational survey data.

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 policymakers, 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 policymakers, 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 insecticide treated bednets (ITBs) as an inexpensive means to prevent malaria among the poor in rural Africa has sparked considerable debate about the effectiveness of free ITB distribution programs. Proponents deem it essential to give 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 et al. (2009) use an experimental auction mechanism to test the hypothesis that poor households keep and use free ITBs. Their experimental results offer strong evidence that the liquidity, income and endowment effects of ITB giveaways virtually eliminate hypothesized resale behavior, providing rigorous evidence to inform a key policy decision not readily addressed using observational data alone. These studies, among many others, demonstrate how and why behavioral experiments offer a valuable tool in development economics research. Unfortunately, this class of experiments remains underexploited to date.

6. Access to Capital and Randomized Controlled Trials: Power and Pitfalls

Building on the prior discussion, this and the next section will examine the role that RCTs have played and might play in the advancement of two longstanding areas of research in development economics. 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 were in turn 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 there is non-price rationing, 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 is 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.[6] 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 assumption, it is unclear whether any data point tells us something about the supply curve or the demand curve.

[6] Adams (1986) observed long ago, it is difficult to know if the observed higher performance of those with 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:

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

where the superscript *c* refers to the constrained (credit-rationed) regime, while the superscript *u* denotes the unconstrained.[7] As pointed out by Feder *et al.* (1991), 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 us wrote 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 in those and similar works was to control econometrically for both observed and unobserved differences between borrowers and non-borrowers and thereby 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.* (1991) and extending through such papers as Kochar (1997), Bell, Srinivasan and Udry (1999), Carter and Olinto (2003) and Boucher, Guirkinger and

[7] Recent theoretical work on ‘risk rationing’ (Boucher *et al.*, 2008) suggests that the regime structure is even more complex than that illustrated here.

Trivelli (2010), 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.

In this context, it is useful to ask what has been and what might be contributed to our understanding of capital access by behavioral 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 putative contributions of the RCT literature to date, this section will close with reflections on how behavioral experiments and randomization might be further utilized to help understand important problems of capital access.

6.1 Faux Exogeneity and the Extent of Non-price Rationing

As has been well-developed 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 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 see if 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 (forthcoming) 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, it is a bit hard to know what to make of these results. Had Karlan and Zinman 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 credible the announcement of a low price. But while a medical experiment can largely control the treatment (e.g., so many milligrams of a drug injected into the blood stream), manipulation of prices and other phenomenon is more complex. Although the price announcement was randomized, we really do not know what was effectively perceived by the treated subjects. Some may have reacted suspiciously to a seeming 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 faux randomization thus raises a serious issue of interpretation. Human agency often confounds the use of human subjects in experiments.

6.2 Here Comes Santa Claus: Randomized Liquidity Injections

While RCT methods still have a way 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, namely the cost of liquidity constraints and the returns to relaxing them. In an ambitious project, de Mel, Woodruff and McKenzie (2009) did exactly that, randomly dropping liquidity gifts on Sri Lankan entrepreneurs. Employing the standard reduced form statistical methods found in the RCT literature to examine impacts on firms' capital stocks and profits, these authors contain mixed results—finding 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 *et al*. 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, 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—a finding relevant for a benefit-cost conscious Santa Claus, but not for profit-minded commercial lenders—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 who both and do not need it. By failing to take into account the economic structure of the problem, their results are not only unreliable predictor of what the relaxation of credit constraints might bring, they also are of dubious external validity as even a Santa Claus 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 *et al.* experimental data in which appropriate information and/or econometric methods were used to distinguish households based on their actual constraint regime. Had de Mel *et al*. 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 some retreat from purely RCT methods, but it would arguably be much more informative.

A recent effort by Karlan and Zinman (2009) to create exogenous variation in liquidity increments comes closer in principal 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 avoid some of the naiveté of the de Mel *et al.* and the early econometric literature.

While perhaps promising as an approach, difficulties with the Karlan-Zinman study illustrate several intrinsic difficulties of implementing RCTs with real economic institutions. First, unlike the de Mel *et al.* cash drops, the Karlan-Zinman created real debt for the randomly de-rationed, exposing them to not only 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, this raises real ethical concerns. Implementation of such experiments would thus seem from a human subjects protection perspective that requires full disclosure to the de-rationed and an ability to compensate 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 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 experiment is substantial non-compliance by the lender with the randomization scheme. In 47% of the cases, loan officers refused to de-ration applicants in accordance with the experimental protocol. De-rationing was thus anything but random. While one could imagine an incentive scheme to induce loan officers to follow a de-rationing protocol, the problems confronted by Karlan and Zinman illustrate the problems of working with real economic institutions and actors as opposed to contrived, completely controlled experiments.

In summary, if one thought we knew nothing about capital access for poor households and small firms in the developing world, because there were few if any prior experimental results on the topic, then we have learned something from the RCTs described here. However, 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 way 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.

6.3 Purging the error term with field experiments

While RCT methods try 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 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. 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.

6.4 Changing the Structure and Rules of Access to Capital

The discussion in sections 6.1 and 6.2 indicated some of the limitations encountered when trying to incrementally change prices (interest rates) and quantities (liquidity) in real economies. We have also seen some studies that more ambitiously try to change the markets, institutions and contracts that provide credit. A common feature of this work (beyond its complexity!) is that it relies on spatially randomized rollout of the innovations in an effort to more accurately gauge their impacts.

One example of this kind of work is that on credit reporting bureaus that link the MFI and conventional banking sectors (de Janvry, Sadoulet and McIntosh, forthcoming). This innovation recognizes that improved agricultural finance proceeds through three steps. First, the smallholder household establishes creditworthiness in the MFI sector, most likely using credit for non-agricultural purposes. Second, the credit bureau then establishes a credible, portable signal that the borrower is of good type. Third, armed with this signal, a lower wealth borrower should then be able to climb a lending ladder, moving from the more restricted purposes and term structures of MFI credit, to standard loan contracts from institutions able to bear the portfolio risk and term structures required for agricultural loans.[8] Utilizing a randomized rollout in the implementation and announcement of an MFI credit bureau in Guatemala, de Janvry *et al.* (forthcoming) shows that credit bureaus indeed fulfill the first two steps. Evidence on the third step is still lacking, but would seem to be a high area for research on access to capital.

[8] Note that from a theoretical perspective, credit bureaus work by offering an alternative solution to the problems of adverse selection that typically drive lenders to require large amounts of collateral, that in turn result in wealth-biased credit rationing.

A second example of recent work that employs randomization methods to explore structural changes in the information in real credit markets is the work reported by Giné, Goldberg and Yang (2010). Working with a lender in rural Malawi, Giné *et al.* 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. They find that fingerprinting indeed boosts loan repayment rates, but only for those borrowers who would ex ante be predicted to have lower loan repayment.

Finally, there are recent and newly initiated research projects that employ randomization methods to see if 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.[9] While the theoretical case for interlinking credit and insurance is strong (Carter, Cheng and Sarris, 2010), additional evidence is needed to see whether index-based insurance can crowd out risk rationing and crowd in new sources of agricultural credit supply to smallholders. Because they introduce novel (and complex) contracts into real markets, low uptake is a significant issue threatening the statistical power of RCT methods. While randomized encouragement designs may offset these problems,[10] it remains to be seen if indeed insurance innovations can structurally change access to capital.

7. Productivity growth: markets and technologies

Significant, broad-based and sustained improvements in living conditions ultimately turn on productivity improvements, whether these arise through technological or institutional change,

[9] The BASIS Collaborative Research Support Program has launched an Index Insurance Innovation Initiative to support such projects. See http://i4.ucdavis.edu for details.
[10] However, see Mullally, Boucher and Carter (2010) for some caveats.

gains from market-based exchange, or some combination thereof. The study of productivity growth – What ignites innovation? Who adopts new innovations or enters new markets, when and why? 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 leading to faster productivity growth. Experimental approaches would seem to lend themselves well to obviating these problems. So how much are we now learning about productivity growth through experiments and how much might we learn through more innovative efforts to integrate experiments into development economists' research designs?

In 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 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.[11] Since relatively little innovation is undertaken or even funded by agencies that commission careful impact evaluation studies before the diffusion of a new

[11] 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 in his home plots as much to maintain his scientific skills as to uncover the dramatically yield-increasing suite of techniques he ultimately discovered, , according to colleagues at Tefy Saina, the group he founded.

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.[12]

In spite of the limited scope 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 policymakers, 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 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 intended beneficiaries what change they would most value. Behavioral and environmental economists have long employed a variety of methods for rigorously eliciting valuation of characteristics not (yet) available in the market. Experimental methods have become especially popular of late 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

[12] Ashraf et al. (2009) offers 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.

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.

7.1 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 1994). Partly, this is 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.[13]

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 by developing country farmers.[14] Interhousehold heterogeneity in transactions costs, social connections, wealth and other observable or elicitable characteristics likewise seems to explain much of the heterogeneity in smallholder market participation patterns (Barrett 2008, Barrett et al. 2010).

[13] Engle-Warnick et al. (2007) is a rare exception; they find ambiguity aversion helps explain adoption of modern crop varieties by Peruvian farmers.

[14] Feder et al. (1985) offer a thoughtful, albeit now dated, survey of the Green Revolution era evidence.

Just as with lending contracts, individuals' subjective perceptions of new markets and technologies can vary markedly, due to differing levels of confidence in or understanding of the product, interhousehold 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, elicitation of subjective distributions (Manski 2004, Delavande et al. forthcoming) and identifying social network effects (Bramoullé et al. 2009, Conley and Udry 2010, Santos and Barrett forthcoming). Combining these nonexperimental methods with experiments can contribute significantly to advancing rigorous inference about what drives uptake of productivity-enhancing innovations. As with the use of behavioral 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 experimental methods to help improve causal inference. But, rather as in the literature on capital access, experiments must build on what is already known 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 et al. (2008, 2009) ran a series of RCT experiments 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. Ultimately, they find that the 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 nonexperimental methods, however, establish convincingly that many farmers face low returns to fertilizer, even if the average returns significantly exceed the cost of purchase (Marenya and Barrett 2009b, Suri forthcoming).

Why the discrepancy? One likely reason is that Duflo et al. 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 ante soil conditions. Marenya and Barrett (2009a,b), studying similar western Kenyan maize farmers to those in the Duflo et al. 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 and purchase and application of fertilizer.

As this example illustrates, it matters less whether one uses an experimental design than whether one 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 studies), but not if methodological hubris induces researchers to rashly 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 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" in the biomedical literature). Beyond this impact evaluation benefit, however, encouragement designs based on discounts for commercially distributed innovations (e.g., fertilizer, index insurance) also permit 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 heavily affect uptake, encouragement designs are an example of an experimental method that generates credible identification not only of causal impacts but also of a key elicitable characteristic that matters for prescriptive analysis: how should agro-dealers price fertilizer or underwriters price insurance and, what, if any, subsidy level most effectively stimulated 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 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 are wasteful of scarce resources; 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 et al.'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 Kirinyaga District, a high agronomic potential area with good access to the international airport at Nairobi. Those estimates do not project to other areas, however, 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.

Unless they egregiously violate the targeting principle, experimental methods cannot overcome the placement problem intrinsic to technology adoption (Besley and Case 2003), 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.

7.2 Estimating the welfare effects of productivity growth

Most development economists' interest in productivity growth is instrumental not intrinsic; 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 for experimental methods 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, RCTs can do a reasonably good job of establishing the short-term, direct gains of, for example, technology adoption to adopting households as compared with those who do not adopt. 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 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 experimental methods so widely championed today for rigorous impact evaluation are intrinsically ill-suited to estimate 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. But there remains untapped power in 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 uptake of innovations, as well as use of encouragement designs to help evaluate and inform 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 preferential use of RCTs, as is currently favored by some donors and academics.[15]

8. Conclusions

Experiments 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 in development economics. Indeed, there are different sorts of experiments and the profession would do well to reallocate effort somewhat away from the current fashion of RCTs for impact evaluation and toward behavioral experiments that elicit credible estimates of otherwise-unobservable parameters that matter to observable behavioral and welfare outcomes. We must stop chasing the siren's song of "perfect identification" – no such thing exists – and work at more creative combinations of methods that can best help us generate useful new data and descriptive, predictive and prescriptive analysis to help answer the most pressing questions about improving the human condition. We need rigorous methods in service of answering pressing questions reliably rather than methods in search of questions they can answer, as is too often the case today.

Experimental design is extremely useful for generating exogenous variation in variables of interest. But it often comes at a cost unacknowledged by RCTs' most ardent champions, especially when methodological cheerleading leads to rash dismissal – or willful ignorance – of prior knowledge built using other tools and to flagrant violations of basic principles of research ethics. All research methods, including ones based on randomization under an experimental design, have shortcomings. Despite the repeated assertions of RCT fans, there is no 'gold

[15] 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.

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 crucify development economics on a cross of golden irrelevance and hubris.

References

Acemoglu, D. (2009), “Theory, General Equilibrium, Political Economy and Empirics in Development Economics,” *Journal of Economic Perspectives*, forthcoming.

Adams, D. (1988). “The Conundrum of Successful Credit Projects in Flundering Rural Financial Markets,” *Economic Development and Cultural Change* 8:347-366.

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

Ashraf, N., X. Giné and D. 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, A. and E. Duflo (2008), “The Experimental Approach to Development Economics”, CEPR Discussion Paper No. DP7037.

Banerjee, A. and E. Duflo (2010) “Giving Credit where Credit is Due,” unpublished manuscript, Massachusetts Institute of Technology.

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

Barrett, C.B. (2008) “Smallholder Market Participation: Concepts and Evidence from Eastern and Southern Africa,” *Food Policy* 33(4): 299-317.

Barrett, C.B., M.E. Bachke, M.F. Bellemare, H.C. Michelson, S. Narayanan and T.F. Walker (2010), “Smallholder Market Participation in Agricultural Value Chains: Comparative Evidence from Three Continents,” Cornell University working paper.

Barrett, C.B., M.R. Carter and C. P. Timmer (2010), “A Century-Long Perspective on Agricultural Development,” *American Journal of Agricultural Economics* 92(2): 447-468.

Barrett , C.B., C.M. Moser, O.V. McHugh and J. 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, K. (2005), "The New Empirical Development Economics: Remarks on Its Philosophical Foundations", *Economic and Political Weekly* XL(40): 4336–4339.

Bell, C., T.N. Srinivasan and C. Udry (1997). “Rationing, Spillover and Interlinking in Credit Markets: The Case of Rural Punjab,” *Oxford Economic Papers* 49: 557-585.

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

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

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

Binswanger, H.P. *(*1981), "Attitudes Toward Risk, Theoretical Implications of an Experiment in Rural India." *Economic Journal 91:867-890.*

Boucher, S., C. Guirkinger and C. 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. S., M.R. Carter and C. 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.

Bruhn, M. and D. McKenzie (2009), "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics* 1(4): 200-232.

Cardenas, J.C. (2009), "Experiments in Environment and Development," Annual Review of Resource Economics 1(1): 157-183.

Carter, M.R. (1989). "The Impact of Credit on Peasant Productivity and Differentiation in Nicargaua," *Journal of Development Economics.*

Carter, M.R. and M. Castillo (2005), "Coping with Disaster: Morals, Markets and Mutual Insurance: Using Economic Experiments To Study Recovery From Hurricane Mitch," in C.B. Barrett, ed., *The Social Economics of Poverty: On Identities, Groups, Communities and Networks* (London: Routledge).

Carter, M.R., L. Cheng and A. 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, M.R. and P. 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, T.G. and C.R. Udry (2010), "Learning About a New Technology: Pineapple in Ghana". *American Economic Review* 100(1): 35–69.

Conning, J., and Kevane, M. (2002), "Community-based targeting mechanisms for social safety nets: A critical review," *World Development 30*(3): 375-94.

Deaton, A. (2010), "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature* 48(2): 424-455.

de Mel, Susresh, D. McKenzie and C. Woodruff (2008). "Returns to capital in microfinance: Evidence from field experiments," *Quarterly Journal of Economics* 123(4): 1329-1372.

de Janvry, A., C. McIntosh and E. Sadoulet (2007). "The Supply and Demand Side Effects of Credit Market Information," unpublished manuscript, University of California, Berkeley.

de Janvry, A., E. Sadoulet and C. McIntosh (forthcoming). The Supply and Demand Side Impacts of Credit Market Information," *Journal of Development Economics*

Delavande, A., X. Giné and D. 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, E., M. Kremer and J. 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.

Duflo, E., M. Kremer and J. Robinson (2009), "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya," NBER working paper number 15131.

Engle-Warnick, J., J. Escobar and S. Laszlo (2007), "Ambiguity aversion as a predictor of technology choice: Experimental evidence from Peru," McGill University working paper.

Evenson, R.E. and D.Gollin,eds. (2003), *Crop Variety Improvement and Its Effect on Productivity: The Impact of International Agricultural Research (*Wallingford, UK: CABI).

Feder, G., Just, R.E., and Zilberman, D. (1985), "Adoption of agricultural innovations in developing countries: A survey," *Economic Development and Cultural Change* 33(2): 255-298.

Feder, G., L. Lau, J. Lin and X. Luo. 1990. "The Relationship between Credit and Productivity in Chinese Agriculture: A Microeconomic Model of Disequilibrium," *American Journal of Agricultural Economics* 72: 1151-1157.

Giné, X., J. Goldberg and D. Yang (2010). "Identification Strategy: A Field Experiment on Dynamic Incentives in Rural Credit Markets," unpublished manuscript, University of Michigan.

Harrison, G.W. and J. List (2004), "Field experiments," *Journal of Economic Literature* 42 (4): 1009-1055.

Hayami, Y. and V. Ruttan (1985), *Agricultural Development: An International Perspective* (Baltimore: Johns Hopkins University Press).

Heckman, J.J. (1992), "Randomization and Social Policy Evaluation," in C.F. Manski and I. Garfinkel, eds, *Evaluating Welfare and Training Programs (*Cambridge, MA: Harvard University Press).

Heckman, J.J., S. Urzua and E. Vytlacil (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics* 88 (3): 389-432.

Hoddinott, J., J. Maluccio, J. Behrman, R. Flores and R. Martorell (2008), "Effect of a nutrition intervention during early childhood on economic productivity in Guatemalan adults," *The Lancet*, 371 (9610): 411-416.

Hoffmann, V., C.B. Barrett and D.R. Just (2009), ""Do Free Goods Stick to Poor Households? Experimental Evidence on Insecticide Treated Bednets," *World Development* 37(3): 607-617.

Hutton, J.L. (2001),“Are distinctive ethical principles required for cluster randomized controlled trials?” *Statistics in Medicine* 20(3): 473 – 488**.**

Imbens, G.W. (2010), “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009),” *Journal of Economic Literature* 48(2): 399-423.

Karlan, Dean (2005). “Using Experimental Economics to Measure Social Capital and Predict Real Financial Decisions,” *American Economic Review*, Volume 95(5), pp. 1688-1699.

Karlan, Dean and Jonathan Zinman (2008). “Credit Elasticities in Less Developed Countries: Implications for Microfinance,” *American Economic Review*, Volume 98(3), pp.1040-1068.

Karlan, D. and J. Zinman (2009). “Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts,” *Review of Financial Studies*.

Karlan, Dean and Jonathan Zinman (forthcoming). “Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment,” *Econometrica*.

Kochar, A. (1998). “An Empirical Investigation of Rationing Constraints in Rural Markets in India,” *Journal of Development Economics* 53(2): 339-372.

Luseno, W.K.. J.G. McPeak, C.B. Barrett, G.Gebru and P.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. T.J. (2005), “Indian farmers’ valuation of yield distributions: Will poor farmers value ‘pro-poor’ seeds?” *Food Policy* 31(5): 415-441.

Maertens, A. (2010), *Social Networks, Identity and Economic Behavior: Empirical Evidence from India*, Cornell University Ph.D. dissertation.

Maluccio, J.A., J. Hoddinott, J.R. Behrman, R. Martorell, A.R. Quisumbing and A.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, M. (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, P.P. and C.B. Barrett (2009a), "Soil Quality and Fertilizer Use Among Smallholder Farmers in Western Kenya," *Agricultural Economics* 40(5): 561-572.

Marenya, P.P. and C.B. Barrett (2009b), "State-conditional Fertilizer Yield Response on Western Kenyan Farms," *American Journal of Agricultural Economics* 91(4): 991-1006.

Michelson, H. (2010). "Welfare effects of supermarkets on developing world farmer suppliers: evidence from Nicaragua," Cornell University working paper.

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

Minten, B. and C.B. Barrett (2008), "Agricultural Technology, Productivity and Poverty in Madagascar," *World Development* 36: 797-822.

Moser, C.M. and C. B. Barrett (2006), "The Complex Dynamics of Smallholder Technology Adoption: The Case of SRI in Madagascar," *Agricultural Economics* 35(3): 373-388.

Mullally, C., S. Boucher and M.R. Carter (2010). "Perceptions and Participation: Mistaken Beliefs, Encouragement Designs and Demand for Index Insurance," unpublished manuscript, University of California, Davis.

Ravallion, M. (2009), "Should the Randomistas Rule?" *BE Press Economists' Voice*.

Rodrik, D. (2008),"The New Development Economics: We Shall Experiment, but How Shall We Learn?" unpublished manuscript, Harvard University.

Ruttan, V.W. (1997), "Induced Innovation, Evolutionary Theory and Path Dependence: Sources of Technical Change," *Economic Journal* 107: 1520-1529.

Schultz, T.W. (1964), *Transforming Traditional Agriculture* (New Haven: Yale University Press).

Sial, M. and M.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, 1996.

Suri, T. (forthcoming),"Selection and Comparative Advantage in Technology Adoption", *Econometrica*.