EXPERIMENTAL CONVERSATIONS

Perspectives on Randomized Trials in Development Economics

Timothy N. Ogden,
editor

Thankfully, I was nudged away from this thinking by a number of my interlocutors, particularly Jonathan Morduch and David Roodman, but also Bill Easterly, Michael Clemens, and Sendhil Mullainathan. They helped me push through my preconceptions and engage with the critics and their critiques, much to my benefit, and hopefully reflected in this book.

Trials, Controls, Randomization and Experiments

While I expect that if you have picked up this book, you are generally familiar with randomized control trials and field experiments, it behooves me to make sure no one is left behind.[3] I will strive to be brief and point to other resources with more complete explanations. In that spirit, for a more complete explanation of the econometrics of assessing causal impact and the various approaches, see Angrist and Pischke's *Mostly Harmless Econometrics* and *Mastering 'Metrics*. For a guide to how randomized control trials and field experiments are set up and run, see Glennerster and Takavarasha's *Running Randomized Evaluations*.

For those wanting a quick and simple introduction, let me begin by delineating some differences between some terms that are often casually used as synonyms or analogues: control trials, randomized control trials, and field experiments.

A control trial is part of the fabric of the scientific method. You assemble a sample and apply a treatment to one part of the sample and do nothing to the other part, the control group. The difference between the treatment group and the control group allows an assessment of the effect of the treatment. If we want to know the effect of the sun shining on a rock, we have to compare rocks that have been exposed to the sun and rocks that haven't. It is simply impossible to reliably assess the effect of a treatment without a control group. Unfortunately, a lot of research in the social sciences and humanities claims causal effects without a control group. To be clear, there is great value in descriptive research that helps us learn about the world without a control group; the problem is when such research makes a causal claim about the impact of a program or some other change.

There are multiple ways to put together a control group. Historically (and unfortunately, still today, evident if you spend any time reading business books or articles) in social science, economics, and otherwise, this was

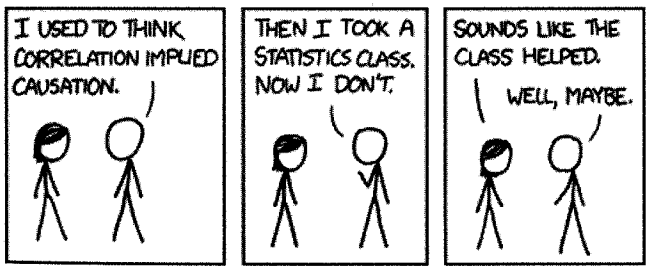
3. If nothing less, my explanations of RCTs, their origin, history, and modern application will further illuminate my own biases and blind spots for close readers.

often done by comparing people who had received a treatment—say, participating in an agricultural extension program—and people who didn't. When dealing with inert objects like rocks, such an approach may work. When it comes to assessing living things who have varying environments, histories, motivations, personalities, and innumerable other characteristics, particularly self-determination, that approach doesn't work. People who participate in an agricultural extension program may be quite different from people who don't—in age, experience, resources, knowledge, and so forth—and those differences may be responsible for different outcomes rather than the agricultural extension program. Thus the oft-repeated dictum, "Correlation does not equal causation."

This presents a major problem for people who want to understand the effect of a program. Social scientists have over the years devised a number of ways, primarily statistical techniques, to try to get around this problem.

One approach is to try to measure the factors that might affect an outcome and match up participants and nonparticipants who are most alike (often done through a technique called propensity score matching). This can even be done after the treatment is conducted. So, for instance, you could try to match up participants in the agricultural extension program with nonparticipants based on their characteristics (age, school grade completed, size of farm, crops grown, prior harvests, etc.). Then you can compare these very similar farmers to each other rather than comparing all participants to all nonparticipants.

Another popular approach is the use of what is known as regression discontinuity. Regression discontinuities occur when there is some external factor that changes for one part of an otherwise similar group, but doesn't plausibly influence the outcomes of interest except through participation. Again, a simple example is an agricultural extension program that becomes



available to farmers in Iowa. While it wouldn't make much sense to compare those Iowan participants to farmers in Saskatchewan (because the reason the program exists in Iowa and not in Saskatchewan has many causes and effects that are likely to be more important than the content of any education program), it is more plausible to compare Iowan farmers who now have access to an agricultural training program to a set of farmers just across the Missouri River in Nebraska. The approach depends on believing that the differences between a farm and a farmer on each side of a state line are negligible. A researcher might argue that these farms are at the same latitude, grow similar crops, share access to the river, and none of the potential participants had any influence on where exactly the river lies or the political process that led to the river being a state boundary. A researcher can then use the introduction of a new program available only to people on one side of the river to figuratively construct a treatment group and a control group in order to make comparisons.

A related statistical approach is the use of instrumental variables (IV). An instrument is, again, a factor that allows a researcher to distinguish a population but does not affect the outcome of the program in question. Distance is sometimes used as an instrument—in the example case, perhaps distance from the farm to the location of the extension course. For an IV approach to be useful, the instrument must only affect the outcome in one way. In the example case this would mean that distance from the location of the extension course does not affect anything about farmers' outcomes other than their participation in the course.

In each of these instances, statistical techniques are used to create a group for comparison in order to understand the impact of a program. But none of them fully solve the initial problem of ensuring that the only difference between the treated group and the comparison group is the program or change you are trying to study.

Matching depends entirely on whether you are able to determine and measure all of the characteristics that may have an effect on the outcome: a dubious proposition at best because so many likely factors are what is termed "unobservable." Unobservable characteristics,[4] in this context, might include motivation, relationships with agricultural suppliers or buyers, actual cultivation practices (unless you are going to follow the farmers around for the entirety of the study), or microclimatic differences. Regression discontinuity and IV approaches similarly depend on important, and

4. Note that "unobservable" does not mean impossible to observe in theory, but unobserved within the confines and budget of a study.

difficult to prove, assumptions. In the example, suppose that on the Iowa side of the river, the farms were on a bluff, whereas on the Nebraska side, farms were in lowlands subject to flooding. For someone standing in one of the fields, this may be obvious. For an economist[5] (or a later reader of an economics paper) who has never set foot in Iowa, it is completely unknown—which makes it difficult to assess the validity of the finding. Instrumental variables are, if anything, more problematic. For instance, the distance instrument essentially assumes that the location of the course was determined independently of any variables that affect farmers' outcomes. But it is likely that in the real world, courses are held in towns that grew up as hubs for the best-performing farms in the area. That land is therefore likely more productive and more valuable. As a result the farmers who own that land and are closest to the course are already better off than those further away. Of course, it's also possible that the farms closer to towns have been farmed more intensively for more years and the soil is of poorer quality, so the closer farmers are worse off. In just that way the validity of most any matching, regression discontinuity, or instrumental variable approach can be, and is, endlessly debated (at least among economists).

A randomized control trial is an experiment in which the treatment group and control group are determined using a random draw or lottery (or some similar process). Randomization, when done properly, can much more effectively isolate the effect of the treatment from other factors that may influence outcomes. Randomization avoids having to exhaustively measure and categorize every feature of the objects of study while usually ensuring that the treatment group and control group are similar enough for reasonable comparisons to be made.[6] While randomization has major

5. Many of the papers that use regression discontinuity or IV designs are done retrospectively—the person doing the analysis is separated by time and distance from the data. It would therefore not be surprising at all to find an economist working with data about agricultural extension programs in Iowa while never having visited the state.

6. Randomization does not guarantee that the treatment and control groups are similar, in the same way that flipping a coin twice does not guarantee that you will get one heads results and one tails result. It is always possible that the treatment and control group, even though randomized, are different from each other in important ways. In most RCT papers you will see an attempt to show that this is not the case by comparing the treatment and control group along observable characteristics gathered as part of a baseline survey and showing that the two groups are "balanced." While helpful, this still does not entirely do away with the possibility that unobserved factors are not balanced and will affect outcomes. This is one aspect of the "Nothing Magic" critique of RCTs discussed in detail later.

benefits in assessing impact, it isn't easy to do. Figuring out how to assemble the sample of interest and exactly how to randomize (e.g., by individual, by groups of individuals, by towns) can be quite complicated. For instance, continuing the agricultural extension example, you would want your sample of interest either to include all farmers and compel participation in the extension program for those who were randomly assigned to participate or to make your sample all farmers who wanted to participate but randomize which ones were allowed to do so. For these and other reasons, some of the statistical techniques described above are often used in conjunction with randomized control trials. There are many, many more considerations in setting up samples, assigning treatment status, and related factors that I won't go into here but, again, point you to *Mastering 'Metrics* and *Running Randomized Evaluations*.

In summary, it is difficult to convincingly assess causal impact because it requires comparison of groups that are identical in all important ways other than what you are trying to assess. Social scientists (and indeed other disciplines) have developed a variety of approaches to create reasonable comparison groups. None are perfect, but randomization is usually the most likely to yield the necessary conditions for comparison (though even this is not universally accepted). And this brings us to field experiments.

A field experiment is a trial conducted in a real world environment rather than a controlled environment like a laboratory. For instance, the majority of psychology experiments are conducted in labs. There are good reasons for this. Labs let you more closely control what a subject is exposed to. But such close control also has a downside. A person may behave differently in a lab environment than they would in the outside world. So the laboratory setting, rather than preventing an outside influence from unduly affecting the outcomes, can become the influence affecting the outcomes (keep this in mind; it will come up again later when we discuss critiques of RCTs). Running an experiment in the field is much more complicated and expensive, however. The possible differences between the behavior of people in lab experiments and real world situations have been a major area of contention in the development and acceptance of behavioral economics. When laboratory experiments showed people allowing irrelevant factors to influence their spending or investing choices, many traditional economists objected that such behavior in a lab experiment was not a reliable signal to how people would behave in the real world when they had to really live with the consequences of their decisions. See Richard Thaler's book *Misbehaving* for a terrific overview of how such issues have been debated.

In case it is not apparent, RCT is not a synonym of field experiment. That being said, when the phrase "field experiment" is used in this book, it means a field experiment using an RCT unless otherwise noted. Most of the studies discussed in the book are field experiments using RCTs in development economics, though lab experiments and natural experiments are also discussed.

Based on the description above, it's easy to think that randomized control field experiments are the gold standard for assessing the impact of any program and guiding decisions about what policies and programs to implement or expand and which to cancel. Whether that is the case is the heart of the contention over the explosive growth of RCTs in development economics. Before we take a look at the critiques of RCTs, I'll take a brief foray into the history of RCTs in development economics.

A Greatly Condensed History of RCTs in Development

Randomized control trials themselves have a long and somewhat disputed history. There was no breakthrough moment, or particular innovator who happened upon the concept of randomized control trials in a eureka moment. There were elements of RCTs in Louis Pasteur's public tests of his anthrax vaccine. Indeed there were elements of an RCT in one of the stories in the Old Testament book of Daniel. While there are various examples like these, often but not exclusively in the medical field, stretching back hundreds of years, most agree that the major steps in using RCTs to evaluate policies and programs took place in the 1920s. Particular praise is given to Roland Fisher, who studied agricultural practices, for making major advancements.[1] Julian Jamison documents that the use of randomized assignment to treatment or control was appearing in many disciplines through many channels around the same time.[2]

When it comes to the modern use of RCTs in assessing the impact of social programs, there are two main streams—one dealing with the evaluation of large-scale, mostly government programs in the United States and Western Europe, and the other with their use in developing countries by economists.

In the United States, the first RCT to evaluate a social program was an evaluation of a welfare program. That experiment carried a lot of controversy, but a group of social scientists at organizations like RAND, Mathematica and MDRC championed the use of large-scale randomized trials to assess impact. The story of how these social scientists convinced the federal government in particular to fund such trials and to use the evidence that

emerged from them is told by Judy Gueron (one of the interviewees) and Howard Rolston in their book *Fighting for Reliable Evidence*.

As the title of Gueron's book indicates, the motivation for the RCT movement in the United States and RCT movement in development economics 20 years later was the same: is it possible to accurately and reliably measure the impact of a program in order to determine whether it is worth the money spent on it? The proponents of RCTs were not the only ones proposing methodologies to provide more reliable estimates of program impact. Indeed there were raging battles from the 1970s into the present day of the relative value of different approaches to measuring impact. While few were arguing that RCTs didn't provide reliable measures of impact, the core question was the cost of arriving at the answer. Many economists argued that statistical techniques—like matching, regression discontinuities, and instrumental variables—provided reliable enough evidence without the expense and other complexities of setting up randomized trials.

It was the cost and complexity that caused many to assume that RCTs were not feasible outside of the world's wealthiest countries, where governments typically didn't have the capability, experience, or budget to fund and manage the kind of trials that RAND and MDRC were conducting in the United States. That changed in the early 1990s through two widely influential studies. The Mexican government ran a very sophisticated and large scale RCT to evaluate a new conditional cash transfer program called PROGRESA (now Opportunidades). Around the same time, Michael Kremer convinced a friend at a small Dutch NGO to conduct a randomized trial of the value of textbooks in Kenyan schools.

Roughly 25 years later, when thousands of RCTs have been conducted in developing countries, it's easy to underestimate what a huge shift this was, particularly the work by Kremer. Kremer's innovation was not just in bringing a method into a new environment but seeing that while it was true that running large-scale experiments with government funding was usually impossible in developing contexts (Mexico may not be Sweden but it is a relatively wealthy country with a highly capable civil service by global standards), there was another path to conducting such trials: working with NGOs. NGOs had several advantages over local governments: (1) they were running a wider variety of programs, (2) they were typically more flexible and nimble in their ability to change operational procedures, and (3) they did not have to pretend to serve everyone—in fact their limited budgets meant that, in most cases, they knew they could not serve everyone they believed would benefit from their programs. The latter point is key to enabling randomization and overcoming ethical concerns: if you strongly

believe that the program you are running will benefit people, it would arguably be unethical to deny that program to some people in order to create a control group. But if you cannot serve everyone anyway, it is, again, arguably fairer to determine who is served via randomization than by some other method.

Running an RCT, however, is not just a question of convincing an NGO to randomize who it serves. A great deal of infrastructure is necessary to gather baseline information, implement the randomization, ensure the implementation follows the randomization plan (i.e., that the randomly selected treatment group gets the treatment and the control group does not), and follow-up with both the treatment and control group to see what happened. That infrastructure then makes it easier to conduct additional RCTs. So it was that Kremer's first RCT on textbooks soon led to what may be the most famous RCT in development economics: Michael Kremer and Ted Miguel's evaluation of the effect of deworming children, conducted in Kenya with the same NGO that had worked with Kremer on the textbook evaluation. Ultimately, an RCT nexus emerged in Kenya. Many of the economists interviewed here spent time in Busia, Kenya, learning to conduct RCTs and a number of the most well-known RCTs in development economics were conducted in and around Busia. Kremer later worked with Abhijit Banerjee and an NGO in India, Seva Mandir, which led to another RCT nexus in India. Kremer and Banerjee went on to work with Esther Duflo and Sedhil Mullainathan to create the Jameel-Poverty Action Lab (J-PAL) at MIT—a different type of nexus for RCTs—one focused on applying the findings of RCTs to policy and program design. Dean Karlan worked with Esther Duflo and Abhijit Banerjee on some early RCTs in India and Kenya and founded Innovations for Poverty Action, which now has developed an impressive infrastructure for conducting RCTs in many different countries. Ted Miguel, Kremer's partner in the deworming evaluation, went on to cofound the Center for Effective Global Action at UC-Berkeley, another center that enables RCTs and using their results to inform practice and policy.

While the earliest RCTs in development were focused on education and health, the use of RCTs to evaluate financial services, particularly microfinance, exploded. Just as RCTs were being proved feasible and useful in development economics, microfinance was emerging as a new and exciting approach to attacking poverty. RCTs and microfinance were well matched. Economists, of course, have a particular interest in finance and financial contracts. The contracts that govern microlending were particularly suited to being adjusted for the purposes of experiments. And microfinance institutions had the information, systems, and infrastructure to make many aspects of setting up an RCT easier than in other types of interventions.

While microfinance was providing a fertile ground for the growing use of RCTs, the situation in economics PhD programs was also helping make economists-in-training particularly receptive to the benefits of RCTs. A long-term evolution (and associated intellectual battle) of how to prove causal effects in economic research had come to a head in the late 1980s and early 1990s. The use of complex statistical techniques, such as instrumental variables, had become very popular in response to criticism of general use of regressions and comparisons without plausible controls. However, as noted, the credibility of conclusions based on such techniques is wholly dependent on whether someone believes the inherent assumptions (e.g., in an IV approach that the instrument is independent of the outcome)—assumptions that are nearly impossible to prove empirically. That meant that researchers had to withstand withering criticism from reviewers not just about the conclusions that they reached but about whether the statistical approach used was valid and reliable. Meanwhile econometricians, particularly Joshua Angrist and Guido Imbens, were pointing out that most analyses using these techniques were reporting average treatment effects across a sample, assuming that the average was a useful measure when there was good reason to be suspicious that averages were themselves biased. As a result there were even more challenges for those graduate students to overcome in getting papers published and dissertations approved. In that environment the appeal of a method that seemed to provide surer ground in identifying causal effects is obvious.

Several other factors also seem to have played a role in the rapid growth of the RCT movement in development economics, particularly among younger economists. One was the seeming exhaustion of existing data sets. In the early 1990s there were very few reliable long-term data sets available for development economists to use. So many, many economists were using the same data sets over and over again. Those data sets were not only limited but confined to a few countries, mostly in Asia and Latin America (Morten Jerven has a recent book about the paucity of reliable data for African countries, even today). Young economists, to make their mark, had to do something novel—and there was little novel data to be had. Creating your own data, as a field experiment does, was one of the few paths to doing something new.[7]

7. It wasn't just young economists conducting field experiments taking this path though. For example, Robert Townsend, an economist at the University of Chicago at the time and now at MIT, began visiting a group of villages in northern Thailand regularly over more than a decade and created a novel data set. See http://cier.uchicago.edu/ for more details on what is known as the Townsend Thai Project.

The path to creating your own data was made more palatable with the decreasing costs of travel to and communications with developing countries. Of course, many things were still difficult, expensive, and unreliable, but it was much easier to do field work in the 1990s that it had been in the 1960s, and even easier today. Rapid advancement in technology also made it easier to collect, enter, and analyze data. Whereas earlier eras of development economics had been biased toward macroeconomics—because macro-level data was all that could be reasonably collected—these technological advancements made it much easier to study microeconomics and the development of regions, towns, even individual households.

The infrastructure (via organizations like J-PAL and IPA) that the pioneers of the movement set up first in Kenya, then India, and then in other countries has continued to make conducting field experiments easier. Thus we moved from a novel way of evaluating the causal impact of textbooks in Kenya to thousands of RCTs in dozens of countries in 20 years.

Still, many weren't and aren't convinced.

An Overview of Critiques of the RCT Movement

Here I want to provide a brief overview of what I perceive as the main critiques[8] of the RCT movement in development economics:

1. The "Nothing Magic" critique
2. The External Validity critique
3. The Policy Sausage critique
4. The Trivial Significance critique

There are obviously other critiques—and nuances of the critiques I do cover—than it is possible to cover here. For another perspective of the critiques, and in-depth essays from proponents and resisters of the RCT movement, see *Thinking Big and Thinking Small: What Works in Development*, edited by Cohen and Easterly.

These are not critiques of randomized control trials at their essence—few, even among the staunchest critics of the RCT movement, would argue that the method does not help us learn something about the world. For instance, James Heckman who was one of the main critics of the very large RCTs in the United States, arguing that other statistical methods allowed

8. I don't address the critique of randomization or experimentation on the ethical grounds because I believe that critique to have been soundly answered. The critiques presented below are ongoing debates where many thoughtful and well-informed people disagree.

reaching conclusions as reliable as RCTs at much lower cost, has participated in an RCT of early childhood interventions.[3] The question is about what exactly RCTs teach us about the world and the value of that knowledge for making future choices. The critiques are about the claims made for the knowledge RCTs produce and how that knowledge is applied.

The Nothing Magic Critique

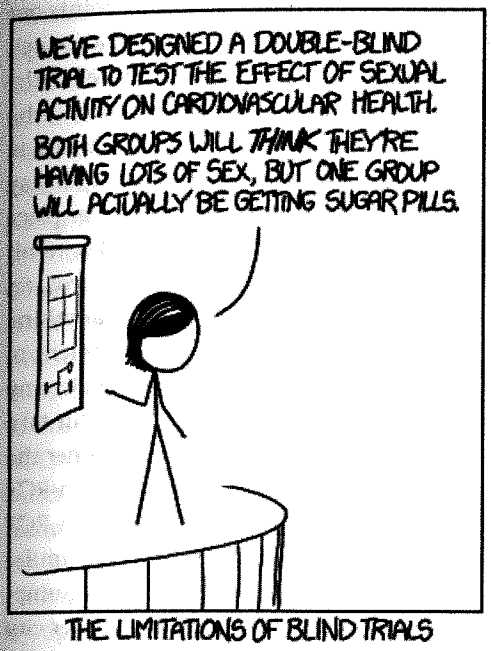
This critique is so named because you often will hear something like, "There is nothing magic about RCTs." This critique is a response to the oft-repeated assertion among RCT advocates (and you will see this in many of the interviews) that RCTs are the most reliable measure of causal impact because they offer the most understandable and certain way to compare treatment to controls. In addition, advocates of RCTs often argue that the RCT method limits the degrees of freedom of researchers to cherry-pick samples or finesse results to find evidence for their preferred theory, and therefore provide more reliable answers to impact questions than any other method.

One version of the Nothing Magic critique is simply stating that other methods can be as, or even more, reliable than RCTs. This critique often points back to or builds off of the debates around RCTs in the 1970s that were never definitively resolved. Glenn Harrison and Angus Deaton, in particular, have consistently expressed this version of the Nothing Magic critique.[9][4]

Another version of the Nothing Magic critique is that field experiments in economics do not conform to the double-blind standard of RCTs in medical practice. Ideally, neither the participants in the experiment nor the researchers themselves know whether an individual is receiving the treatment or is in the control group. This is why drug trials often feature a placebo—so that both the control group and the treatment group are taking something. This is particularly important because it is well established that taking placebos does have a positive impact on many health conditions. It is not unreasonable to believe that people who know they are in a field experiment and know they are receiving a treatment—improved seeds, an infusion of cash into their business, a textbook—behave differently than the otherwise similar people who know they didn't receive the treatment.

9. Harrison's perspective is notable because he does not dismiss the value of RCTs but strongly disputes their primacy. Deaton typically takes what might be called a more pessimistic view, emphasizing that there are substantial problems that come with any methodology, including RCTs.

Some critics argue that the inability to run double-blind trials, or even blind trials, means that field experiments don't provide the better answers that RCT proponents claim.



Another version of the critique says that even if RCTs do limit degrees of freedom, nothing is eliminated. Therefore RCTs have to be as carefully scrutinized as other methods. Recent work examining the results of medical trials using RCTs bolsters this critique; it found that the number of "no-effect" results increased markedly when researchers had to file a pre-analysis plan documenting exactly how they would assess the data gathered before the experiment was conducted.[5] Furthermore RCTs are as vulnerable to inadvertent false positives and false negatives as any research method.

As a result there is a limit to how much is gained from RCTs, particularly as running RCTs is generally more expensive than alternative evaluation approaches. While proponents advocate for increasing the use of RCTs, the Nothing Magic" critique says that there may already be little value for money in the number of RCTs being conducted.

The External Validity Critique

Clearly the goal of the RCT movement is to shape policy and programs around the world so that they are more effective. This involves not only the specific program studied in a particular trial, but applying what is learned from that trial to other situations. The External Validity critique points out that each RCT is anchored in a highly specific context. This includes such things as the implementer carrying out an intervention, often an NGO, the personnel hired by that NGO, local and regional culture and customs, the survey technique, the specific way questions are asked, even the weather. Thus the critique points out, while the results from a particular RCT may tell you a lot about the impact of a particular program in a particular place during a particular point in time, it doesn't tell you much about the result of a similar program carried out in a different context. In other words, an RCT of microcredit in urban India does not necessarily tell you anything about the impact of microcredit in rural Kenya. An in-depth treatment of the External Validity critique can be found in Nancy Cartwright and Jeremy Hardie's book *Evidence-Based Policy*. Lant Pritchett and Justin Sandefur take it in on with specific reference to the RCTs of microcredit and whether the results of one help predict the results of another, finding that a non-RCT from a local context does a better job predicting outcomes than an RCT from a different context.[6] Conversely, Hunt Allcott does something similar comparing the ability of an RCT of reminders to reduce energy consumption in one city to predict the effect of the same campaign in another city finding that RCTs don't do a great job, but a better one than other methods in common use.[7]

It's important to note that the External Validity critique doesn't just apply to RCTs. It applies to studies or experiments using any methodology. Every study is conducted in a specific context and is not necessarily valid in other contexts—at least until it is replicated in multiple contexts with similar results. David McKenzie has pointed out that there seems to be a double standard in the application of the External Validity critique to field experiments using RCTs.[8]

That being said, many published RCTs don't do enough to explain the context in which the study takes place to allow a reader to form a judgment about external validity. As Jonathan Morduch points out in his interview, the journal publishing process is biased toward broader claims of validity, which gives teeth to the external validity critique.

The Policy Sausage Critique

The Policy Sausage critique is primarily associated with Lant Pritchett—and we discuss it in his interview. The simplified version is that policies

(whether policies of government or of NGOs) are created through complex and opaque actions influenced by politics, capability, capacity, resource constraints, history and many other factors. In other words, policy making is like sausage making. Impact evaluation, and independent academic research in general, plays only a small role in the policy sausage, especially if it is impact evaluation that comes from outside the organization. That may seem irrelevant to the use of RCTs in development economics, but the RCT movement is far from just an academic exercise. Many of the lead practitioners advocate RCTs not just as a better way of estimating causal impact but as an essential guide to program design and policy making.

Pritchett and others argue that the process of policy change or organizational change is completely separate from the process of knowledge creation. The bridge between the two is not built on policy briefs but on painstaking work inside bureaucracies, political machines, and organizations. External evaluation when imposed from above or outside, according to this critique, usually hampers that work rather than accelerates it. Where RCTs have influenced policy significantly, it is in areas like PROGRESA's conditional cash transfers that were conducted within the policy-making realm and because they support the existing political goals of policy makers, not because those policy makers change their minds as the result of evidence. The Policy Sausage critique argues that the RCT movement, while trying to influence program and policy, does not have a reasonable path to actually affecting policies and programs.

The Trivial Significance Critique

I term this the Trivial Significance critique to differentiate it from the common use of the term "significance" in statistical discussions, which refers to the likelihood of a measured effect arising from chance but also confusingly about the total size of the effect (in this second sense, it is a synonym of "material" in business and accounting vocabulary). The Trivial Significance critique is not about statistics or relative effect size but about absolute effect size: whether the programs and policies the RCT movement is focused on matter.

The critique can take several different guises, but all share the basic point that the programs and projects measured and measurable by RCTs yield changes, even when "successful," that are not big enough to make a difference between poverty and prosperity, even for a single family. One version might be phrased, "Yes, the program you evaluated increased the average time spent in school by a full year, but there are still no jobs available for

those kids."[10] Another version is that the things that "really matter" are macroeconomic-level choices like trade policy—and those macroeconomic choices cannot be randomized. A third version is that what "really matters" is the allocation of funds (or effort) across a variety of policies or goals: Should a government spend on roads or sanitation or trade promotion?[11] While RCTs may be able to say something about what approaches are most effective in encouraging hand-washing, it is hard to imagine an experiment that could compare the impact on economic growth of spending on infrastructure to the effect on employment of providing tax credits for exporters to the effect on health of increasing pay for community health workers.

The Argument Behind the Arguments

The RCT movement has voluminous responses to each of the critiques I've just laid out in simplified form. I do not think much is to be gained by providing grossly simplified versions of the responses to the grossly simplified critiques. The critiques and the responses can put too much emphasis on the particularities of methodology, and distract from the more important disagreement behind them. That more important argument (most apparent in the Trivial Significance critique) is about theories of change; it only occasionally bursts into view, most often in books authored by development economists and reviews of those books by other development economists.[12]

Argument over theories of change—ideas about how the world changes—are hardly unique to the present moment in development economics. Indeed, it is the foundation of development economics (and much of other social sciences): how is it that poor countries become richer (or, why is that poor countries stay poor)? Obviously, this is not a mechanical process or an outcome of natural law. Poor countries become richer or stay poor because of human action. But which human actions? And what is the process for changing those actions? That is what theories of change are all about.

10. See Lant Pritchett's post "Is Your Impact Evaluation Asking Questions That Matter?" at the Center for Global Development's blog for a particularly pointed exemplar of this version of the critique: http://www.cgdev.org/blog/your-impact-evaluation-asking-questions-matter-four-part-smell-test

11. An example of this version of the critique can be found in a post on the World Bank's Future Development blog by Jeffrey Hammer, titled "The Chief Minister Posed Questions We Couldn't Answer": http://blogs.worldbank.org/futuredevelopment/chief-minister-posed-questions-we-couldn-t-answer

12. The book *What Works in Development*, mentioned earlier, also helps illuminate differences in the theories of change among the various contributors.

There has always been wide disagreement within the economics profession about theories of change. The disagreement over the effectiveness of efforts to intervene in markets, or the benefits of reducing intervention in markets is a good example; to put a more personal face on it, think Keynes versus Hayek. Development economics has its own versions of theory of change conflicts. To outsiders, the most visible disagreement on theories of change among development economists in recent years has been between Jeff Sachs and Bill Easterly. While Sachs has promoted large-scale, precisely planned technocratic interventions (emphasizing the need to intervene), Easterly advocates for political and economic rights of individuals and the value of local knowledge, contrasting "Searchers" (enabled by free markets and effective) from "Planners" (interveners in markets who are ineffective and often harmful). Abhijit Banerjee and Esther Duflo explicitly couch their book, *Poor Economics*, as a contrasting vision between the Sachs and Easterly poles. They make a case for the value of technocratic knowledge and planned interventions, but not of the size and scale advocated by Sachs. Meanwhile, more of today's prominent development economists have staked out their own theories of change in their own books: Daron Acemoglu and James Robinson argue, most simply, that "institutions matter"; Angus Deaton that development aid does more harm than good by undermining political rights and accountability.[9]

While wary of reducing theories of change to short summaries or points on a chart, nevertheless I find it helpful in the context of the RCT movement in development economics, to think about the competing theories of change along three main axes:

- the value of small versus big changes;
- the value of local knowledge versus technocratic expertise;
- the role of individuals versus institutions.

I have not created a three-dimensional chart (I tried, but failed to produce something comprehensible) to capture these axes because they are not completely independent of each other. Someone who believes strongly in the value of big changes is obviously also very likely to place more value on technocratic expertise and the role of institutions.

There is significant variation in the theories of change of RCT advocates and critics of the movement on these axes. The general view is that the randomistas have a theory of change that, with apologies to Margaret Mead, could be stated as, "Never doubt that a committed group of small tweaks can change the world." In practice, there is significant variation within the RCT movement and between the critics, such that in some cases there is more in common between a particular RCT advocate and a

particular critic than there is between two different critics. A good example is the Targeting the Ultra-Poor (TUP, sometimes also referred to as the Graduation Model) programs evaluated by teams including Banerjee, Duflo and Karlan.[10] The program is a package of interventions designed to lift people out of extreme poverty (defined as living on under $1.25 per day). The TUP program shares with Sachs' Millennium Villages Project the concept that a package of interventions is necessary to make a difference for the extreme poor. The TUP program, though, was created by BRAC, a Bangladeshi NGO, based on their long experience serving ultra-poor populations in Bangladesh in a process that bears more resemblance to Easterly's Searchers. The impact evaluations found that the TUP program was quite effective and so the *randomistas* are now encouraging the scale-up and adoption of TUP programs in more countries. Indeed, that was the plan all along (discussed in the interviews with Dean Karlan and Frank DeGiovanni), which again bears more resemblance to Sachs' model.

Underneath each of the critiques of RCTs noted above is a theory of change that differs from that of RCT advocates along at least one of the three axes. After the many conversations collected in this book, my impression is that those in the RCT movement tend to believe that small changes can matter a great deal,[13] that technocratic expertise is highly valuable, and that individuals within institutions matter as much as the institutions themselves. Those critics who invoke the Trivial Significance critique, in contrast, usually agree on the value of technocratic expertise, but disagree about the value of small changes and the role of institutions. Because differing theories of change are so foundational in RCT debates, I explicitly ask each interviewee about their theory of change.

How to Read This Book

Rather than delving further into the metaphysics of the arguments over RCTs and perhaps confusing the issues more than illuminating them, let me suggest an approach to reading this book and then get out of the way. The intent, after all, is to let you hear directly from the advocates, critics and others.

13. Rachel Glennerster, one of the interviewees, has a post that illustrates this point well by looking at how the "small" change of free bed net distribution based on RCTs can be traced to averting 450 million cases of malaria and 4 million deaths—as she puts it, "that's anything but small." See http://runningres.com/blog/2016/5/27/not-so-small

The approach I'd suggest is to keep the critiques and the axes of theories of change in mind[14] as you read the interviews:

1. Look for examples of the proponents of RCTs treating the results of RCTs as "magic" or universally better than alternative methods.
2. Look for examples of claims to external validity, and caution about external validity. More important, look for examples of RCTs causing someone to think differently about an issue, to alter her or his beliefs about how the world works.
3. Pay special attention to the discussions with each of the interviewees about their theory of change.
4. Think how the work of participants in the RCT movement is likely to materially change the lives of individuals, communities, and countries for the better. Think about alternative ways the resources—money and brain power—going into that work could be deployed to greater effect.
5. And finally, think about your own theory of change—the role of programs like those being evaluated, be they large or small, the role of individuals and institutions, the role of technocrats in setting policies and creating programs, and the role of evidence in policy and program development and evolution.

With that, I'll let what I intended to happen all along begin: letting you hear directly from the people in this book.

Acknowledgments

Well, one last thing before that. There are a variety of people who deserve specific recognition in the long, slow process of getting this book done. Among the interviewees I specifically want to thank Abhijit Banerjee and Esther Duflo for consenting to an interview back in 2008 (which doesn't appear in the book) but which kicked off this whole process; David McKenzie who spent more time with me than anyone else, tolerating my endless questions about the Sri Lanka experiments; and Jonathan Morduch, who helped tremendously on the introduction and conclusion as well as constantly nudging me to finish. Others who have my deep gratitude: Erin Graham for originally signing the book for UPenn Press;

14. A helpful mnemonic: PENS; Policy sausage, External validity, Nothing magic, trivial Significance.