# *RCTs in Development Economics, Their Critics and Their Evolution*

June 2020

Timothy Ogden

Forthcoming in Florent Bédécarrats, Isabelle Guerin, and François Roubaud, *Randomized Control Trials in Development: the Gold Standard Revisited*. Oxford University Press.

## Abstract

The use of RCTs in development economics has attracted a consistent drumbeat of criticism, but relatively little response from so-called randomistas (other than a steadily increasing number of practitioners and papers). Here I systematize the critiques and discuss the difficulty in responding directly to them. Then I apply prominent RCT critic Lant Pritchett's PDIA framework to illustrate how the RCT movement has been responsive to the critiques if not to the critics through a steady evolution of practice. Finally, I assess the current state of the RCT movement in terms of impact and productivity.

* Thanks to Jonathan Morduch, David McKenzie, Lant Pritchett and Isabelle Guerin for helpful discussions in developing this chapter, as well as the participants in the AFD Workshop Randomized Control Trials in the Field of Development: The Gold Standard Revisited (Paris, March 2019). Also thanks to Cynthia Kinnan, Jessica Goldberg, Johannes Haushaufer, Cyrus Samii, and Bruce Wydick for helpful pointers, and the three editors of the book, Florent Bédécarrats, Isabelle Guérin, and François Roubaud, for valuable comments and discussions.

## Introduction

Pascaline Dupas simply wanted to help some of the poorest people in the world (this story is drawn from Ogden 2017). But she discovered, as unfortunately many do, and even more unfortunately don't, that the educational structures for training wealthy students in wealthy countries impart very few practical skills for living among or "helping" poor households in developing countries. Put another way, she couldn't get a job at an international NGO. Discouraged, she pursued her second best option: a fellowship at Harvard.

There she discovered an opportunity to move to Kenya to serve as a research assistant for a randomized field experiment. She abandoned the fellowship early to get literally closer to her original goal, helping poor people in developing countries. While living in Kenya she befriended a young mother—and watched as that woman struggled to afford needed medicine to treat her infant infected with malaria. Again, like many others, she began wrestling with the question of why. Why was it so hard for this woman to get medicine, to set aside a few dollars, to prevent her baby from contracting malaria in the first place? When friends back in France asked her what they could do to help, where they could give money to make a difference, she suggested buying bednets. When it became clear that there were few, if any, charities where it was possible to give toward buying bednets, she and a few friends (who also went on to become professional economists) set up a charity to do exactly that.

When that charity came under criticism for giving away bednets—criticism based on standard economic theories about people devaluing things that were free and free goods distorting markets—she didn't accept or reject the criticism. She decided the best thing

to do was to test whether the critics were right. So she set up a randomized trial to vary the price and the implementation of bednet subsidies, to discover the optimal way to distribute bednets. The experiment was very influential by quantitative and qualitative measures. The paper has been cited more than 570 times according to Google Scholar and is in the top 5% of all research articles according to Altmetric. It has heavily influenced policy recommendations on distribution of bednets. GiveWell, relying on this study particularly in terms of implementation of programs to distribute bednets, has channeled $100 million to free bednet distribution.[1] Overall, bednet distribution is estimated to have prevented 450 million cases of malaria and 4 million deaths. (Bhatt et al. 2015; Glennerster 2016). While in the 2000s there was a great deal of debate about the effects of subsidies and free distribution of health goods—debate that had raged on for years based on competing theory and claims from non-experimental studies—today such debates have largely disappeared.

This story encapsulates most of the debate about RCTs in development economics: motivations, impact, internal and external validity, scope, theories of change, ethics and hypocrisy. The so-called randomistas are famous for critiquing the use of anecdotes, for instance, and yet this is an anecdote. The randomistas demand evidence for causal claims, and yet there is certainly no randomized evidence to show the research in question caused anyone to change their mind or practices. Critics argue that there is "nothing special" about RCTs and yet evidence like this seems to have been more convincing to non-economists than claims based on theory or research produced

[1] Note: I am the Chairman of GiveWell; this data is sourced from interviews with GiveWell staff.

through other methods.[2] Critics claim that RCTs limit the questions that can be asked and answered to narrow unimportant ones and yet saving 4 million lives is "bigger" than many economists can claim as impact from their studies of "big" questions. Critics denounce the purposelessness of "external" evaluations that don't influence policy or practice and yet this RCT grew directly out of a practice question of one of the leaders of a charity.

The debate over RCTs is as old as RCTs. This volume is at minimum the fourth—after Cohen and Easterly (2010), Teele (2014) and Ogden (2016)—to bring together conflicting views on the role of randomized field experiments in social science/development economics/policy. A review of the history of the debates would lend evidence to the dictum that science advances one funeral at a time, since no one changes their minds. I approach the topic with trepidation. My (most likely forlorn) hope for a contribution is to attempt to summarize and systematize the most prominent critiques, examine the difficulties in making a meaningful response to those critiques and finally discuss how the practice of RCTs has evolved. Whether or not this evolution has been in response to critics is impossible to say, but the evolution *has* been responsive to the critiques. I conclude with a model for thinking about the evolution of RCTs and empirical and experimental methods, their current status and likely future.

## The Critiques of RCTs

[2] In Ogden 2017 (pp. 201-204), Frank DeGiovanni states that the Ford Foundation funded RCTs of "Targeting the Ultra Poor" programs because the foundation found that RCT evidence was more convincing to policymakers (another anecdote!)

Here I want to provide a brief overview of what I perceive as the main critiques of the use of RCTs in economics (and social science) to provide some semblance of order to responses. I group the critiques into seven categories:

1. The "Nothing Magic" critiques
2. The Black Box critiques
3. The External Validity critiques
4. The Trivial Significance critiques
5. The Policy Sausage critiques
6. The Ethical critiques
7. The "Too Much" critiques

There are other critiques and nuances of these critiques that I do not cover. I'm certain that some critics will object to the way I characterize their arguments. But one must start somewhere.

**The Nothing Magic Critiques**

This critique is so named because it's most direct expression is, "There is nothing magic about RCTs." This critique is a response to the idea that RCTs "sit atop a hierarchy of methods" (Ravallion 2020) for estimating causal impact.

The main version of the Nothing Magic critique is that randomization does not necessarily yield a less biased estimate of impact than other methods. Cartwright and Deaton (2018) is the most complete discussion of the main form of the Nothing Magic critique. Cook (2018) details 26 assumptions required to believe that an RCT in fact

yields an unbiased estimate. This critique often points back to or builds off of the debates around RCTs extending back to Student's (1938) critique of Fisher.

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—and could therefore be referred to as there is "nothing magic about development economics RCTs." The inability to run double-blind trials, or even blind trials, means that RCTs in social sciences generally don't meet the requirements to reduce one of the main sources of expected bias.

A third 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. Ioannidis (2018) summarizes the results of several reviews of RCTs that find significant evidence of bias in published RCTs in several disciplines. Young (2018) finds that the majority of published RCTs fail to correct for multiple testing and fail retroactive tests for significance. Kaplan and Irvin (2015) study the results of medical trials using RCTs and find 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. Furthermore RCTs are as vulnerable to false positives, false negatives and magnitude errors as any research method (Gelman 2018).

**The Black Box Critiques**

Closely related to the Nothing Magic critiques are a set of critiques positing what can be learned from most RCTs is limited to whether some intervention "worked" but not *why* it worked. An RCT does not necessarily illuminate the actual casual mechanism even when a causal relationship is convincingly established. A useful example is an interview of James J. Choi and Dean Karlan about an RCT they conducted where the two (and the implementer) disagree about the root cause of their results (Dubner 2018). RCTs that find no effect can be even worse. It many cases it is impossible to determine whether the null result is because of an ineffective treatment or an ineffective implementation of the treatment. Interpretation of null results is often unclear even among proponents of RCTs (Evans 2016).

A variation of the Black Box critique is the "theory-less" critique. An RCT that is not grounded in theory can be very difficult to interpret regardless of whether the outcome is distinguishable from zero or not. But if there is a well-grounded theory informing the RCT, the benefits of randomization may be quite limited. It is possible to conceive of alternative (and simpler to implement) approaches to test a clear theory.

**The External Validity Critiques**

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, 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 even running an exactly identical program carried out in a different context and time. An in-depth treatment of the External Validity critique can be found in Nancy Cartwright and Jeremy Hardie's book *Evidence-Based Policy*. Cartwright and Deaton also contribute to this critique in their papers (e.g. 2018), as do Pritchett (2020), and Ravallion (2020).

**The Trivial Significance Critiques**

I term this the Trivial Significance critique to differentiate it from the common use of the term "significance" in statistical discussions. Here I use it as a synonym of "material" in business and accounting vocabulary. The Trivial Significance critique is not about statistics or relative effect size but about (at times, truly) absolute effect size: whether the programs and policies the RCT movement is focused on matter.

The critiques 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, at anything approaching the scale of the problem of global poverty. These critiques can come from a macro perspective (the things that "really matter" are macroeconomic-level choices like trade policy which cannot be randomized) (Ravallion 2020), a systems perspective (an RCT on increasing vaccination rates doesn't improve the health system, and may in fact hinder system development) (Garchitorrena et al. 2020), or a political economy perspective (RCTs cannot answer whether investing in

transport infrastructure, health systems, or education systems is most likely to lead to growth) (Hammer 2014).

**The Policy Sausage Critiques**

The Policy Sausage critiques are primarily associated with Pritchett. 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. 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. Thus, the effort put into an RCT is likely wasted, as it will fail to have an effect on this complex process. Bédécarrats et al. (2020) note the very limited number of programs evaluated via an RCT that seem to have been scaled up.

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. Pritchett has specifically claimed that the randomistas model of policy adoption is “unbelievably Cro-Magnon.” (Ogden 2017).

## The Ethical Critiques

As long as RCTs have been conducted there have been critiques that the method is unethical. There are two main forms of this critique. One is that experimenting on human beings, particularly and especially on people in poor communities as is necessarily the case in development economics RCTs, is inherently unethical. Meyer et al. (2019) find this moral intuition that experimentation is wrong is widespread in the American population, at least.

Historically, the medical research community has dealt with moral aversion to experimentation and withholding of treatment via the concept of equipoise—only conducting experiments where there is reasonable uncertainty over the benefits (or relative benefits) of a treatment (Freedman 1987). In the context of development economics, Abramowicz and Szafarz (2020) argue that the concept of equipoise has been glossed over too quickly given the poverty and deprivation of many participants in RCT studies, and is frequently simply ignored by proponents of the method.

## The "Too Much" Critiques

The "Too Much" critique that even if there were advantages to RCTs over alternatives, those advantages do not justify the time, monetary, opportunity or "talent" costs they impose. Ravallion (2009a and 2020), for instance, argues that there are too many RCTs being conducted, that they are pushing out evaluations of programs that are not suited to randomization, and implies that RCTs occupy too much space in journals. Pritchett (Ogden 2017) worries that the "main founders of the movement are all geniuses" who should be leaving RCTs to "public health PhD students at Kansas State." Deaton suggests that they are better suited for consultants to governments to settle political

disputes than inform academic knowledge (Ogden 2017). Others decry that the time it takes to conduct and analyze an RCT are too high compared to other methods, even if they are imperfect. Alternatively the method generally requires organizations implementing a program to hold the intervention constant during the period of the evaluation, regardless of feedback, which imposes large opportunity costs (Whittle 2011).

## The Challenge of Responding to The Critiques

The proponents of RCTs have hardly been silent in the face of these critiques. As noted there are several volumes and many more standalone papers, blog posts, interviews, briefs, books and more that present the case for RCTs and responses to particular critiques or critics. But as time has passed, the responses have been less frequent. In recent years, if anything, the proponents of RCTs have seemingly begun to simply decline to engage with the critics (though, as I will argue later in the paper, they have engaged with the critiques, if indirectly).

Perhaps one reason is that the defenders are at something of a rhetorical disadvantage to the critics—and the situation recapitulates some of the early exchanges in the debates when it was the proponents of RCTs who were the critics of establishment methods. In brief, the critic needs simply find a few examples of a particular problem, while the response must defend an amorphous and evolving establishment.

### What is a randomista?

A specific example of this difficulty is that of the basic question of defining who (or what) is a randomista. Many of the critiques are founded on what the randomistas

believe—but there is certainly no manifesto or statement of beliefs or core principles that defines who is in "the club." Lant Pritchett has described the RCT movement as religious, and the emotions that are stirred up certainly seem to make that a valid comparison. To borrow Stackhouse's definition, a religion or movement is "a comprehensive worldview or 'metaphysical moral vision' that is accepted as binding because it is held to be in itself basically true and just even if all dimensions of it cannot be either fully confirmed or refuted." At first glance, particularly for anyone who has been in the audience for many public debates or discussions of the value of RCTs, this description may seem apt for both sides of the debate. But in depth discussions with the individuals quickly erodes the sense that there is a shared "comprehensive worldview" where it comes to the value, benefits and applicability of RCTs. The interviews I conducted for *Experimental Conversations* (2016), with 10 of the leading practitioners of RCTs provides ample evidence of the heterogeneity of beliefs among just this small sample of RCT proponents.

Absent a statement of the core beliefs of a randomista, critics tend to rely on a rhetorical construction that can be rendered, "Since Individual X said Y at t1, group A is wrong at t2." A particular favorite of the critics is the evergreen citation of Banerjee's 2006 statement: "Randomized trials…are the simplest and best way of assessing the impact of a program." To their credit, Deaton and Cartwright (2018) cite a number of statements in addition to the standard, and particularly point to statements from J-PAL's materials which are a much better starting place for a critique, but are still limited unless one is relying on collective guilt or guilt by association. The point is not that such statements are meaningless or that no one believes the statements in

particular, but that it is very difficult to identify who exactly would sign on to a specific statement and whether those who would or wouldn't could reasonably be defined as inside or outside the circle of randomistas.

The nearest attempt to define a randomista appears to be Ravallion (2020), who narrows in to define a randomista as someone who holds a belief that "RCTs sit atop a hierarchy of methods, who believe that RCTs are the "'gold standard' for impact evaluation—the most 'scientific' or 'rigorous' approach, promising to deliver largely atheoretical and assumption-free, yet reliable, I[mpact] E[valuation]." But even this statement leaves much to be desired in terms of the precision necessary to define a group. While there are those who like Imbens (2018) (though note that Imbens is primarily an econometrician and not a development economist) would sign on to the idea that there is a hierarchy with RCTs "at the top," is it sufficient to believe there is a hierarchy or does there need to be some specified amount of space between RCTs and other methods? How would that space be measured? Even the standard economics construction of "ceteris paribus, RCTs are a superior method for impact evaluation" leaves chasms of undefined terms where there would be significant heterogeneity between "randomistas" and quite possibly less space between a purported randomista and a critic. For instance, consider the following quotations and attempt to classify the statement as belonging to a randomista or to a critic. Some have been slightly modified to remove obvious "tells. The Appendix has the original unaltered quote (if you would like to judge if the alterations were fair) and for attribution.

1. What methods are best to use and in what combinations depends on the exact question at stake, the kind of background assumptions that can be acceptably employed, and what the costs are of different kinds of mistakes.
2. We should neither be encouraging or discouraging any particular tool just for the sake of the tool. We should be encouraging students to look for an interesting question and use the right tool to answer it. Period.
3. [V]ery good descriptive data that focuses people's attention on something they haven't focused on before has changed people's minds in policy as much as any experiment.
4. The novelty…drove the overselling of RCTs, like these silly statements that everything ought to be evaluated randomly, or the people who say they don't believe any observational evidence.
5. [No approach] makes all the problems disappear, and neither does an RCT. I don't think anybody [should think] that RCTs are magical.
6. [T]hat's part of how the randomized evaluation movement was sold to policy makers: "You're going to get answers." I don't think that's what we're going to get. My sense is that we're going to see evaluations that are all over the map. [i]f I had to choose, I …say we should pour more energy into the big stuff than the small stuff.
7. Organizations should be able to draw on different areas to answer the relevant questions…I see a lot of crossover between different forms of causal identification…I don't think you … focus on randomized evaluations. I don't think that makes sense.

8. An impact evaluation should help determine why something works, not merely *whether* it works. [RCTs] should not be undertaken [when they] provide no generalizable knowledge on the "why" question
9. There are many banal and useless examples of studies using every specific method.

While any one of those quotations could be set aside or disputed in some way, the fact remains that many of the purported randomistas repeatedly express sentiments that RCTs are a "tool in the toolbox" of modern economics, that there are many other useful tools, that RCTs are not appropriate for every worthy question and that other analytical tools are useful and credible. Moreover, most if not all of the randomistas use and publish other methodologies. As McKenzie (2019) has pointed out, the most cited papers of arguably the three most well-known randomistas—Banerjee, Kremer and Duflo—are not RCTs.

The second half of Ravallion's definition ("the most 'scientific' or 'rigorous' approach, promising to deliver largely atheoretical and assumption-free, yet reliable, I[mpact] E[valuation].") would be even more difficult to get meaningful numbers of economists who practice RCTs to sign on to. There is a decades-long debate within economics of the proper ordering of data analysis and theory that is as intractable as the debate over RCTs (Cherrier 2019). "Largely atheoretical and assumption-free" could (and has) inspired pages of debate on its own. It was this very fact that led to *Experimental Conversations*, where I decided the only meaningful thing to do was to interview many of the randomistas and near-randomistas to hear them explore the nuances of their

beliefs in terms of their actual practice in conducting and interpreting research, not just on the metaphysics of methodologies.

The intent of this discussion is not to end with the conclusion that the term randomista is so ill-defined and undefineable as to be practically useless, though I think that's true, but to illustrate why it is so hard to respond meaningfully to the critiques (and perhaps to explain why the randomistas have generally stopped responding directly to critics, as evidenced by the fact that none agreed to participate in this volume). Any response must necessarily be on behalf of an individual who likely will agree with at least some parts of any particular critique—and ultimately speaks only for themselves. At the same time, any person who, like I do in this paper, tries to defend RCTs must bear some burden to answer for any statement on behalf of RCTs no matter how off-the-cuff, uncareful, misguided or wrong. It's not surprising then that few relish, at this late stage of the ongoing discussions, the opportunity to respond to a critique of ill-defined randomistas in general with a specific statement of personal beliefs. What would that accomplish?

**The Argument Behind the Arguments**

Since I am not a randomista in the sense that I do not make a career out of running RCTs, I can hardly arrogate the authority to answer the critiques or critics on their behalf. That being said, in the next section I will use actions (both studies and other professional activities) of various nominal randomistas to illustrate how the movement has evolved in ways that at least blunt many of the critiques. Before doing so, I want to point out one other issue that makes productive direct responses to the critiques difficult.

The disputes between randomistas and their discontents can put too much emphasis on the particularities of methodology, and distract from the more important disagreement behind them. That more important disagreement is about theories of change. 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)?

There has always been wide disagreement within the economics profession about theories of change. One can think of most of the seminal texts in economics as manifestos about theories of change. Development economics has its own specific theory of change conflicts (e.g. Sachs vs. Easterly; Acemoglu's and Robinson's "institutions matter"; Rodrik's industrial policy; Deaton's anti-aid arguments).

While wary of reducing theories of change to short summaries or points on a chart, nevertheless I find it helpful in the context of this discussion, 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 individual versus collective actions via institutions.

The axes 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. 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—again the problem of definition of a randomista arises.

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 I've had with both randomistas and critics, my impression is that those in the RCT movement tend to believe that small changes can matter a great deal, 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.

This difference matters because it influences how one evaluates the quality and especially the utility of evidence. If you believe that the path to improving the world is through the accumulation of small changes then external validity is a much lower bar. You do not need to be convinced that a program will yield the exact impact or even close to the same impact in a different context to be worth transferring. You simply need to believe that it provides the starting place to run another small experiment in the different locale and adjust as you go. The distinction between a randomista with this theory of change and a "feedbackista" like Dennis Whittle is simply about the speed of iteration and how much to value different kinds of feedback. Implicit there, of course, is that this same randomista would look at an RCT that a critic calls atheoretical and be very confused—there is a theory, though perhaps not a structural model that allows informed estimation of cross-context impact. At the same time, a randomista with a

slightly different theory of change that puts more value on institutions will decline to defend small-scale RCTs with NGOs and advocate for much more experimentation at scale (e.g. Niehaus 2019).

Pritchett nods to the importance of the lack of common theories of change in some of his writing and speaking: "What I worry about development is that there are two ontologically different categories to which the word is commonly applied;" (Pritchett 2010a); " You can call that whatever you want, just don't call it development" (Ogden 2017). But in general the lack of a shared ontological universe is not given the attention it deserves in such debates. Perhaps that's because there is unlikely to be much value in such a debate—the discussants are using the same words to mean different things.

## The Evolution of the "Movement"

Among the more prominent critics of the RCT movement, such as it is, has been Lant Pritchett. In 2018, Pritchett began giving a talk he titled: "The Debate is Over. I won. They lost."[3] In this section, I'll argue that by examining the evolution of the use and practice of RCTs it is clear that many RCT critiques have in fact been acknowledged to be correct by virtue of changing practices among RCT practitioners and research centers. In that sense, his triumphalist title is correct. However, I'll also argue that the evolution of the "RCT movement" is best explained by a model that Pritchett himself (along with Matt Andrews and Michael Woolcock) introduced (Andrews, Pritchett and Woolcock 2012) arguing that this model was the best path to sustained development impact.

[3] The title slide reads "We won. They lost." but in his remarks he uses "I".

**Problem Driven Iterative Adaption**

In their original paper (which spawned a number of additional papers and ultimately a book, *Building State Capacity*) Andrews, Pritchett and Woolcock introduce the principles of problem driven iterative adaptation:

> We propose an approach, Problem-Driven Iterative Adaptation (PDIA), based on four core principles, each of which stands in sharp contrast with the standard approaches. First, PDIA focuses on solving locally nominated and defined problems in performance (as opposed to transplanting pre-conceived and packaged "best practice" solutions). Second, it seeks to create an 'authorizing environment' for decision-making that encourages 'positive deviance' and experimentation (as opposed to designing projects and programs and then requiring agents to implement them exactly as designed). Third, it embeds this experimentation in tight feedback loops that facilitate rapid experiential learning (as opposed to enduring long lag times in learning from ex post "evaluation"). Fourth, it actively engages broad sets of agents to ensure that reforms are viable, legitimate, relevant and supportable (as opposed to a narrow set of external experts promoting the "top down" diffusion of innovation).

These four principles are clearly seen in the evolution of the RCT movement.

***Principle One: Solving locally nominated and defined problems in performance***

Michael Kremer is generally credited with beginning the use of RCTs in development with a randomized experiment on the impact of textbooks on learning in Kenyan primary schools; the only controversy in this regard is that Santiago Levy was nearly simultaneously deploying an RCT to evaluate the impact of Progresa, a new conditional cash transfer program in Mexico. Regardless of which of the two is credited with initiating the use, both projects embody the first principle.

Levy's motivation was solving a particular local problem. At the time the program was created, it was clear that the then current government of Mexico was going to lose the next national election. Levy was concerned that the program would be canceled by the incoming government. He implemented an RCT to establish the impact of the program in order to protect it from political concerns.[4]

Kremer's first RCT was motivated by a discussion with a Kenyan friend, Paul Lipeyah, who had the job of picking seven primary schools to receive new textbooks from the NGO ICS. Kremer describes his thinking like this:

> In 1994 when I started the work in Kenya, I was very much influenced by the movement for better identification in labor economics and public finance…I also was not reacting to the critics of instrumental variables. Indeed, I think those working on instrumental variables and those of us working on RCTs were motivated by the same impulse, the concern that a lot of empirical work in economics at the time was potentially subject to confounders and required a lot of fairly strong assumptions. That being said, it's not like IV makes all the problems disappear, and neither does an RCT. I don't think anybody thinks that RCTs are magical, but they are a really useful tool for getting at causal impact. So I would say I was trying to get at causal impact in a way that was part of a broader movement in the economics profession to get better identification…My main impulse was practical—to get more believable answers to real world questions. I have always been mainly interested in the underlying questions of what policies can address poverty and I realized that RCTs were a tool that could be adapted to help answer this question. I was motivated to make RCTs a more flexible and useful tool. (Ogden 2017)

[4] Conversation with Santiago Levy, 2018

Kremer was clearly also trying to solve a locally nominated and defined problem in a double sense. First, he was trying to help the specifically locally defined problem of ICS of picking which seven schools would receive textbooks. Second he was addressing the locally defined and nominated problems among economists of improving causal identification.

Similar stories accompany the entry into RCTs of other economists who are well-known implementers of RCTs. Pascaline Dupas' story, which opened this chapter, is a good example: the first RCT she conducted was in response to the locally defined and nominated problem of whether it was better to give away bednets or charge for them. David McKenzie's first RCT, a test of the returns to capital for microenterpreneurs in Sri Lanka, came about because he had already done similar non-experimental work in Mexico, "But people were not convinced by the nonexperimental results" (Ogden 2017). He was motivated by the locally defined and nominated problem of convincing economists and policymakers that microentrepreneurs could have significant returns to capital. Sometimes, as with Dupas's and Kremer's story, the locally defined and nominated problem was a question that was shared by the economist and an NGO or government agency.

### *Principle Two: create an 'authorizing environment' for decision-making that encourages 'positive deviance' and experimentation*

It's clear that the initiators of the use of RCTs created an authorizing environment that encouraged positive deviance and experimentation, the last quite literally. The level of innovation within the conduct of RCTs is quite impressive. From generally small-scale experiments on very simple interventions, e.g. textbook distribution to seven schools, the practitioners of RCTs have innovated and evolved consistently. From a methodological standpoint, the process of both randomization and analysis has become much more sophisticated to deal with highly varied contexts and potential confounders of prospective balance, multiple hypothesis testing and other potential biases. Early implementers of RCTs such as Ted Miguel have been leaders in research transparency, data and analysis disclosure and the use of pre-analysis plans. A "second wave" of randomistas have gone on to implement much more sophisticated experiments over

much longer timeframes (e.g. Blattman and Dercon's experiments comparing industrial jobs to microcredit) and much, much larger scales (e.g. Muralidharan and Niehaus's experiments with NREGA and Aadhar) on much more complex topics (e.g. Pomeranz's experiments on taxation schemes and Karlan's experiments on religious content in an intervention).

### *Principle Three: it embeds this experimentation in tight feedback loops that facilitate rapid experiential learning*

The only argument about the randomistas implementation of this principle is whether it is something they directly created or an extant feature of economics education that they exploited. I would argue it is both. The nature of economics education and practice means that each new generation of economists learns by doing the grunt work for the prior generations. It is hard to imagine a tighter feedback loop to and more rapid experiential learning than a PhD candidate (or even earlier) spending a year or two as a field research manager on a variety of experiments being overseen by existing practitioners. RCT implementers have also exploited the network of events that the economics profession has as an opportunity for feedback loops on innovations in field experiment set-up, management and analysis—conferences like the Northeastern Universities Development Consortium (NEUDC) and Pacific Development Consortium (PacDev) have, much to the chagrin of RCT critics, often become platforms for rapid learning in how to conduct, analyze and report RCTs.

### *Principle Four: it actively engages broad sets of agents to ensure that reforms are viable, legitimate, relevant and supportable*

The institutions to support the implementation of RCTs are the best examples of this principle in practice. Since the first RCTs, prominent users of RCTs have created organizations like J-PAL, IPA, and CEGA which easily match the description as broad

sets of agents that ensure that reforms are viable, legitimate, relevant and supportable. All of the organization are involved in ongoing projects to reduce the barriers to the conduct of and reporting of RCTs. These include books and courses about how to conduct RCTs, training of many students within and outside of universities, and creation of research organizations in developing countries (e.g. the Busara Center for Behavioral Economics, and permanent field staff in a number of countries).

**PDIA-driven Evolution and Critiques of RCTs**

True to Andrews, Pritchett, and Woolcock's promise, the implementation of PDIA has served to vastly improve the practice of RCTs, their relevance to development practice and to policymakers, and to institutionalize the process of conducting and reporting the results of randomized experiments. But two additional points are necessary.

First, the practice of RCTs has developed as practitioners confront the problems that they perceived. The "locally nominated" problems that the evolution has confronted are primarily the problems faced by RCT practitioners—publishing research and fulfilling personal career objectives. As I'll argue later, many of these career objectives are about improving the world and helping the world's poorest. But that does not obviate that the process of evolution is driven by motivations internal to the movement.

Second, is that this process of internally-driven improvement is, according to Andrews, Pritchet and Woolcock, the only reliable and sustainable way to build capacity. Pritchett often laments that many of the "locally nominated problems" that randomistas have been attempting to address were "entirely predictable" (Ogden 2017) and yet the randomistas ignored the critics. But as in the quotation introducing PDIA above says "transplanting pre-conceived and packaged 'best practice' solutions" and change based

on a “narrow set of external experts promoting the ‘top down’ diffusion of innovation” simply doesn’t work. The only way for the RCT movement to evolve into a sustainable and effective force for development was to develop capabilities and solutions internally.

In this section, I’ll briefly discuss how the PDIA process has led to the evolution of the practice of RCTs to at least in part address many of the critiques of the movement.

**Nothing Magic**

As noted in an earlier section, it is not clear how many RCT practitioners ever believed that RCTs were magic or were not subject to any biases. It is worth noting that Brodeur et al. (2018) find significantly less evidence of p-hacking and significance searching in RCT and RDD papers than in IV and Difference-in-Difference papers and Vivalt (2019) finds less significance inflation in RCTs than in papers using other methods. Perhaps even more important in relation to the present discussion, she finds that RCTs have “exhibited less significance inflation over time.”

It is however likely that many RCT practitioners did not appreciate the many sources of bias that remain in randomized experiments when they first began. If there were believers that RCTs solved all of the problems that critics point out, we would expect that RCT practitioners would resist innovations in implementation and analysis that better account for such sources of potential bias.

In fact, what we see is many economists who might be called randomistas are actively innovating to address concerns about bias, reliability and replicability. Several are worth specific note.

- Ted Miguel is one of the founders of the Open Science Framework and of the Berkeley Initiative for Transparency in Social Sciences. Both organizations encourage researchers to make all the data and code used in their work accessible for replication (whether RCT or not).
- Randomistas including Dean Karlan, Esther Duflo and Chris Blattman have been vocal advocates of pre-registration of studies, especially RCTs, and were involved in creating the AEA RCT Registry.
- J-PAL has started a replication service where "a graduate student attempts to replicate a complete paper from scratch, and can identify any error, omission, or questionable assumption." (Crepon et al. 2019).
- The World Bank's *Development Impact Blog* ("news, methods and insights about impact evaluation") frequently features critiques of papers using RCTs, advice on new statistical methods and techniques for improving RCT analysis, and on non-RCT methods.
- Guido Imbens, cited above as saying that RCTs do in fact sit on top of the hierarchy of methods, continues to work on methodological improvements on other methods, such as difference-in-differences and machine learning (e.g. Athey and Imbens 2018; Athey and Imbens 2019)

Each of these examples can be seen as examples of each of PDIA's principles, given that they address problems of performance in conducting and interpreting RCTs, and are conducted without any central authorizing authority or mandate. More importantly, these efforts illustrate that far from treating RCTs as magic, those within the loose borders of the RCT movement recognize sources of bias and error in RCTs and actively

invest in addressing them. Similarly, they continue to employ methods other than RCTs, and do not, in practice, ignore work on other methods.

**Black Box**

Again there has been considerable evolution in the application of RCTs along PDIA principles to address the limitations of simple RCTs in exposing causal mechanisms. A few examples include:

- Alfonsi et al. (2017) study youth employment schemes in Uganda comparing vocational training programs to subsidies for on-the-job training. They not only establish that vocational training has a higher impact in terms of youth income, but that the mechanism is greater mobility between employers because of certifiable skills, ultimately leading to better matching of certified youth to higher productivity firms.
- Beaman et al. (2018a) study agricultural technology diffusion through social networks in Malawi and establish the nature of the "complex contagion" that leads to farmers adopting new methods, and use the mechanism to identify ways to cost-effectively improve targeting of agricultural extension programs.
- Cai and Szeidl (2017) experiment with Chinese business networks finding that networking meetings significantly improve firm performance, specifically through peer learning from better functioning firms on topics outside the intervention and in better supplier-client matching, and that the regularity of meetings matters.
- Campos et al. (2017) study small enterprise training programs in Togo, comparing traditional business training to personal initiative training. They

not only find that the personal initiative training is more effective in boosting firm profits, but the specific behaviors that changed including product diversification, innovation, and investment.

- Karing (2018) studies a program to encourage child vaccination in Sierra Leone and not only establishes the efficacy of public signaling through colored bracelets, but that the mechanism is social desirability (not attention) and that the effect of the bracelets varies based on the social desirability of specific vaccines.

The examples chosen here are not systematic but are deliberate to illustrate that an emphasis on addressing the "black box" critique is not limited to a few researchers, schools, contexts, or sectors. As more such studies are done, the implicit PDIA process operating within the RCT movement means that establishing mechanisms will increasingly be expected of new RCTs.

**External Validity**

The external validity critique would in general be more credible if some of its proponents were as vocal about the problems of external validity of all studies and not just of RCTs. As Pam Jakiela has noted in response to Cartwright and Deaton, "Nice insight by Deaton and Cartwright, but for some reason they keep spelling 'study' as R-C-T."[5] That being said, there are many instances of RCT proponents offering policy advice which assumes external validity.

[5] https://twitter.com/PJakiela/status/797053999925104640

Faced with the problem of proving external validity, RCT practitioners have evolved in several ways. First they have empirically studied whether the results of RCTs in one context predict results in another. For instance, Meager (2019), using Bayesian Hierarchical Modeling (another example of the application of PDIA to the "Nothing Magic" critique) shows that the variation between RCTs of microcredit is smaller than it appeared (Pritchett and Sandefur 2015) and therefore the results of the individual RCTs are reasonably predictive of the results in other locations. Alcott (2015) 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.

At the same time, RCT practitioners have put much more emphasis on replications and studies with multiple arms in multiple contexts. For instance Dupas, et al. (2018) test savings encouragements in Uganda, Malawi and Chile. Perhaps most famously, a wide variety of researchers collaborated to test Targeting the Ultra Poor programs in eight locations with both government and NGO implementations. This "replication" for external validity can also take place significantly ex-post. For instance, Bernhardt et al. (2018) reanalyze data from multiple experiments on differential returns to capital between male and female entrepreneurs to identify a previously unclear causal mechanism relating to household bargaining and optimization (which is also an example of addressing the Black Box critique). Such ex-post replications are becoming easier because of the efforts of other RCT implementers to ensure data and code for all experiments are available for replication.

Of course, there will always be questions of external validity in the application of any impact evaluation (RCT or otherwise) to predict outcomes in other contexts. But more systematic approaches are also evolving. As more RCTs address causal mechanisms, assumptions about external validity will become more explicit, and more studies will include structural models. This in turn, will allow more formal frameworks for assessing external validity and integrating results from multiple studies such as Dehejia et al. (2019) and Wilke and Humphreys (2019).

**Trivial Significance**

Earlier I noted that a key foundation of the trivial significance critique is differing theories of change between randomistas and critics. Little can be done to respond to a critique that the only changes that matter are macro-level policies. That, however is more a critique of applied microeconomics in general than RCTs. That being said, the RCT movement has a response to at least one of the critiques emanating from a different theory of change. For those that argue that institutions matter, the institution building prowess of the randomistas should be impressive. Aside from the obvious examples of IPA and J-PAL, the institutions build by the randomista movement directly and indirectly include the Global Innovation Fund, the Busara Center for Behavioral Research, 3ie, Evidence Action, Development Impact Ventures, AEJ: Applied, and many local survey firms.

Here though I want to focus on a variety of the trivial significance critique which is grounded in the original experiments that popularized the use of RCTs—textbooks in schools, getting teachers to show up for school, incentivizing vaccination, etc. These critiques focus on both the small nature of the intervention and the small measured

results even when statistically significant (see Harrison 2011 as one example). Another related variation laments that RCTs are not well suited for measuring long-term impact (see Ravallion 2020).

Confronting these issues has yielded an impressive amount of creativity in application of RCTs. The most direct response to the "too small" critique has been expanding the scale of RCTs. Muralidharan el al. (2016, 2018) offer the best example by studying a safety net program in Andhra Pradesh, India, with 19 million people in the experiment, with an estimated savings of $38.5 million per year. While the fact that both figures are in millions rather than billions will leave some unsatisfied, it's certainly marked progress over the early years of the RCT movement.

Other randomistas have pushed the boundaries of what can be studied via RCT in other ways:

- Dina Pomeranz, with a variety of co-authors, has conducted RCTs on a variety of tax policy questions.
- Chris Blattman, with a variety of co-authors, has randomized access to factory jobs in Ethiopia, policing strategies in Colombia, and anti-violence campaigns in Liberia.
- Bryan, Choi and Karlan even conduct an RCT on the impact of religious belief, randomizing Christian evangelism in the Philippines (finding support for the Protestant work-ethic hypothesis; though note the potential ethical controversy, noted later in this paper).

**Policy Sausage**

The translation of RCT results into policy changes has always been an explicit goal of RCT practitioners. Their stories of how and why they began running RCTs commonly include a pragmatic desire to influence policy in order to make a concrete difference in people's lives.[6] A quotation attributed to Michael Kremer by Karthik Muralidharan is illustrative: "Never apologize that your fundamental motivation is to improve the lives of hundreds of millions of people, and that economics is a tool to get there and not an end in itself."[7]

Their good intentions notwithstanding, the initial work of the randomistas in terms of policy influence could be described as Pritchett has on various occasions: naïve, Cro-Magnon. As Bédécarrats et al. (2019) note, less than five percent of RCT impact evaluations conducted by J-PAL have led to scaled-up policy changes.

But as time has passed the sophistication and intensity of efforts to affect policy has accelerated. Having identified this locally nominated problem of limited policy impact, the RCT practitioners rapidly iterated in an environment that encouraged positive deviance and provided rapid feedback within the group.

The initial assumption that evidence would generate policy change mechanically has fallen away in favor of focused efforts to influence policy. This includes the creation of policy-focused teams at both J-PAL (including a "government innovation" initiative at J-PAL that works specifically to support government agencies conducting policy implementation experiments) and IPA. But it also includes participation in the creation

[6] In Ogden (2016), see interviews with Michael Kremer, Esther Duflo, Dean Yang, Chris Blattman among others.

[7] https://twitter.com/karthik_econ/status/1102237584103600129

of standalone organizations to implement programs based on RCT evidence (namely, Evidence Action), organizations to encourage the creation of and use of evidence in policymaking (3ie), internal groups within existing policy and implementation organizations (Development Impact Ventures at USAID), close collaboration with research groups at NGOs (BRAC, Pratham), education programs for policy makers and implementers, and of course, training a huge number of masters and Ph.D. students in the methods and approaches, the vast majority of which will end up in policy-related jobs rather than in academia. Some practitioners have even taken on roles in the policymaking apparatus—Rachel Glennerster's role as Chief Economist at DfID and Andrew Leigh's role as a parliamentarian in Australia come to mind.

Put another way, the randomistas have engaged a broad set of agents to ensure the validity and continuity of the use of RCTs to influence policy. A new generation of RCT practitioners are going to be an integral part of policy-making institutions (if for no other reason than the shortage of jobs in academia).

**The Ethical Critiques**

There is much less to say on this topic. In part, that is because fundamentally randomistas clearly believe that experimentation with human beings is ethical, regardless of the moral intuitions of the majority of American public, an attitude of course shared by most scientists. The common refrain, which I am of course sympathetic too, is that there isn't a choice about whether to experiment (since every policy implementation in an experiment) there is only a choice of how much is learned from an experiment.

But clearly there remain many questions about the ethics of experimentation. During the Summer of 2019, a new working paper that randomized encouragement to participate in anti-authoritarian protests in Hong Kong (Bursztyn et al. 2019) attracted a huge amount of attention[8] specifically because many economists seem to believe the experiment was unethical. An oft-asked question, on Twitter at least, was how the experiment managed to apparently pass through several Institutional Review Boards. The paper and the subsequent discussion revealed[9] that there are yet no meaningful bounds or codes or even shared principles on where economists should draw an ethical line in terms of experimentation.

On the questions of equipoise, as noted above, this remains an area where the RCT movement has yet to significantly engage as best I can tell.

**Too Much: The Final Critique**

The final category of critique I identified falls outside of the PDIA framework as it is not a critique of what RCT practitioners in development economics do, but of how much they do it. I find this the least compelling of all critiques within the economics frame.

To begin with, as many of the Too Much critiques acknowledge, the emergence of RCTs in development economics is in no small part due to the conditions and structure of the market for academic economics. The use of RCTs gained popularity in the context of widespread questions about the credibility of other methods, in an environment that demanded of aspiring economists that they do work that was credible, novel and

[8] https://twitter.com/DurRobert/status/1148090885470654464

[9] https://twitter.com/arindube/status/1148807790787473410

publishable. RCTs promised—and delivered—work that was all three. Thus the criticism of Too Much should really be directed at the structures and incentives of the profession not at those who respond to the incentives the profession creates. This form of the critique is equivalent to criticizing market participants for doing the "wrong" thing, rather than addressing any market failures.

Second, the Too Much critique fails to articulate an objective measure of what the thresholds between "not enough", "just right", and "too much" might be. It is objectively true that the use of RCTs and the publication of papers using the method has increased greatly (Ravallion 2020, Bédécarrats et al. 2019), but this growth must be put in perspective. It's worth quoting McKenzie's (2019) look at the data on this question at length:

> despite the rapid growth, the majority of development economics papers published in even the top-five journals are not RCTs…[O]ut of the 454 development papers published in these 14 [economic development field] journals in 2015, only 44 are RCTs (9.7%). The consequence is that RCT-studies are only a small share of all development research taking place.
>
> The median [BREAD affiliate] researcher had published 9 papers, and the median share of their papers which were RCTs was 13 percent. Focusing on the subset of those who have published at least one RCT, the mean (median) percent of their published papers that are RCTs is 35 percent (30

percent), and the 10-90 range is 11 to 60 percent. So young researchers who publish RCTs also do write and publish papers that are not RCTs.

Third, the oft-repeated assertion that "enthusiasm for RCTs will fade" seems to me to be a hollow critique. Of course we should expect that methods will continue to improve, new innovations in all sorts of research designs will uncover heretofore unappreciated problems and improved approaches. At some point in the not too distant future I can confidently predict that someone will write an essay about "RCTs 2.0" and make a distinction of arguable difference between the "early days" of the RCT movement and the improved methods now in vogue. Perhaps this chapter falls into that category.

Susan Athey, reacting to Judea Pearl criticizing what he terms the naïve approach to causal inference in economics (*as a whole*, not the RCT movement), writes: "[I] think the most effective way to evangelize a new method is to demonstrate its effectiveness in a first-rate empirical application where the method clearly leads to a better quality and more credible result. Researchers will mimic a fully worked out, successful example."[10] That could serve as a short-hand history of the use of RCTs in development economics. Enthusiasm for the original practice of RCTs has already faded as "first-rate empirical applications" of more sophisticated experiments and analysis have emerged. And enthusiasm for current practice will surely fade as "first-rate empirical applications" of improved methods—with randomization at their core or not—are created. Until then, there isn't "too much."

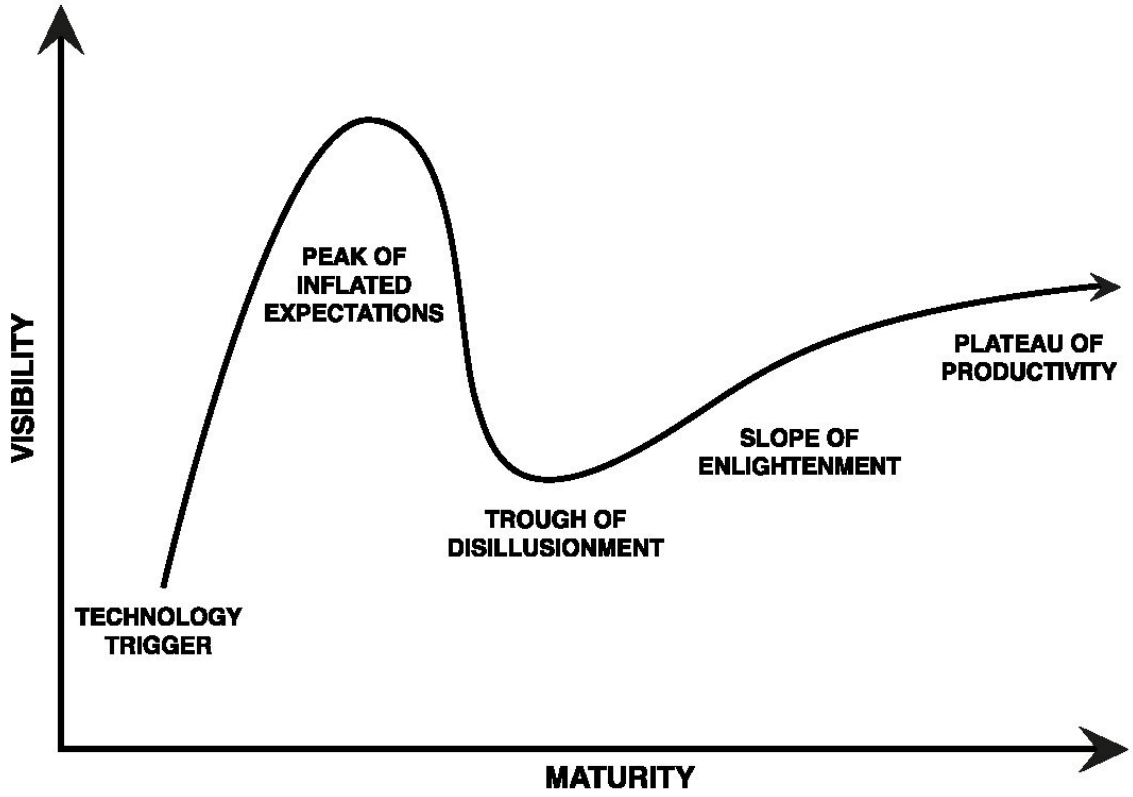
[10] https://twitter.com/Susan_Athey/status/1107422021753790464

Finally, there is the lament that the "brightest and best" economists are wasting their talents focused on RCTs. This critique makes the least logical sense of all. If the critics are right and the problems of RCTs are insurmountable, and there are clearly better alternatives, then that must indicate that those who continue to primarily use RCTs are not the brightest and best. This critique must explain why anyone should believe that the brightest and best are systematically wrong and yet still are the worthy of the moniker. And if they are not the brightest and best, why can't the actual brightest and best convince the next generation of students to abandon RCTs for other methods? The only plausible explanation that makes sense of this critique is that the entire profession of economics is broken, in which case the critics are wasting time on the symptoms and not the causes.

## Conclusion

To conclude, I want to provide a different framework for thinking about the evolution of the practice of RCTs and the various critiques and responses. Here, once again, Pritchett and I overlap. I spent the first 10 years of my career at the technology research firm Gartner. One of the organization's most widely known products is The Hype Cycle—a way of conceiving of the emergence, evolution and adoption of emerging technologies.

The Hype Cycle posits that as a breakthrough technology emerges it passes through 5 distinct phases, named colorfully enough that they require little additional expectation: "Innovation Trigger", "Peak of Inflated Expectations", "Trough of Disillusionment", "Slope of Enlightenment", and "Plateau of Productivity".




Pritchett stumbled across the Hype Cycle and applied it to RCTs in a 2013 essay. I agree that it is a useful model for thinking about RCTs—in fact, I would argue that RCTs are best thought of as an "emerging technology" in development economics rather than a movement.

Through the lens of the Hype Cycle, in this paper I have argued that, 1) the peak of inflated expectations for RCTs was real, but was never as high as critics made it out to be, and in any event has passed, 2) initial enthusiasm for RCTs was quickly met by a range of valid critiques, leading if not to a trough of disillusionment at least to meaningful changes in the use and practice of RCTs, and 3) the current state is clearly in

the slope of enlightenment phase as evidenced by the statements and practices of advanced users of the technology.

It is worth noting specifically that the evolution of the practice of RCTs validates many of the critiques detailed here. The evolution that I have attempted to document is responsive to these critiques. The practitioners of RCTs are not evolving their practice to deal with novel issues that have not been raised by critics—regardless of any direct response to the critics, the randomistas have implicitly acknowledged many of the critiques by evolving in ways that take the teeth out of many of them.

That being said, I believe there is ample reason to believe the plateau of productivity for RCTs is higher than many critics seem to make it out to be, simply as a mechanical consequence of the way the world works. There are many more decisions about implementation than there are about what to implement. Implementation decisions are clearly within the scope of RCTs. Because there are many many more students being trained in development economics than will ever hold tenured jobs at R1 universities, a significant proportion of those students will end up in jobs where implementation questions rather than larger policy questions are their purview. The training they receive in experimentation and causal identification will be highly relevant and applicable to their ability to engage in PDIA in those jobs.

Beyond that, RCTs are a more useful tool for improving the world than most tools available to the *median development economist*, given the nature and requirements of the profession, and the difficulties of policy influence. The emergence of RCT technology and the supporting mechanisms around that technology are applicable to the vast

majority of the actual questions and discrete decisions about anti-poverty policies, programs and implementation. It is true that RCTs are unlikely to be a useful tool for evaluating exchange rate policies, the optimal level of public debt, or the consequences of wealth inequality (just as a few examples), the policies related to the answers to questions on such topics are far less susceptible to academic influence regardless of the methodology issued to answer them. As I write this (in the summer of 2019), the possibility of a massive global retreat from liberalized trading regimes is frighteningly real despite tens of thousands of macroeconomists' decades of policy effort. There is no reason to believe that the marginal impact of the average development economist studying one of these topics is greater than a precisely-estimated zero. The median development economist's comparative advantage would be in improving the implementation of a policy or program, even without any external validity or scale-up.

In closing, I would reiterate again that Lant Pritchett was right, and he won. The critiques of the RCT movement are generally valid if not objectively correct. However, many of those critiques have been addressed by the evolution in the practice of RCTs. I expect that the evolution will continue, and that eventually RCTs may be supplanted by some other methodology (already, of course, there are new "emerging technologies" in economics: big data, machine learning and artificial intelligence—and some of the debates over the use and applications of RCTs are being recapitulated). Until then, I expect that the plateau of productivity where RCTs currently reside will continue to yield benefits to the world.

## REFERENCES


Abramowicz, M. and A. Szafarz (2020). "Ethics of RCTs: Should Economists Care about Equipoise?," In Bédécarrats, F., Guérin, I., and F. Roubaud (eds.), Randomized Control Trials in the Field of Development: A Critical Perspective. New York and Oxford: Oxford University Press, Chapter 10.

Alcott, H. (2015). "Site Selection Bias in Program Evaluation," The Quarterly Journal of Economics, 130(3): 1117–65.

Alfonsi, L., Bandiera, O., Bassi, V., Burgess, R., Rasul, I., Sulaiman, M., and A. Vitali (2017)." Tackling Youth Unemployment: Evidence from a Labor Market Experiment in Uganda." Working Paper, Private Enterprise Development in Low Income Countries (PEDL) Programme, DfID.

Andrews, M., Pritchett, L., and M. Woolcock (2012). "Escaping Capability Traps through Problem-Driven Iterative Adaptation." Center for Global Development Working Paper 299.

Andrews, M., Pritchett, L., and M. Woolcock (2017). Building State Capability: Evidence, Analysis, Action. Oxford: Oxford University Press.

Athey, S. and G. W. Imbens (2018). "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." NBER Working Paper 24963.

Athey, S. and G. W. Imbens (2019). "Machine Learning Methods That Economists Should Know About," Annual Review of Economics, 11(1): 685–725.

Beaman, L., BenYishay, A., Magruder, J., and A. M. Mobarak (2018a). "Can Network-Theory Based Targeting Increase Technology Adoption?" NBER Working Paper No 24912.

Bédécarrats, F., Guérin, I., and F. Roubaud (2019). "All that Glitters Is Not Gold. The Political Economy of Randomized Evaluations in Development," Development and Change, 50 (3): 735–62.

Bernhardt, A., Field, E., Pande, R. and N. Rigol (2017). "Household Matters: Revisiting the Returns to Capital among Female Micro-entrepreneurs." NBER Working Paper 23358.

Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., and P. W. Gething (2015). "The Effect of Malaria Control on Plasmodium Falciparum in Africa between 2000 and 2015," Nature 526(7572): 207–11.

Brodeur, A., Cook, N., and A. Heyes (2018). "Methods Matter: P-Hacking and Causal Inference in Economics," IZA Working Paper 11796.

Bursztyn, L., Cantoni, D., Yang, D., Yuchtman, N., and Y. J. Zhang, (2019). "Persistent Political Engagement: Social Interactions and the Dynamics of Protest Movements," NBER Conference Paper F126621.

Cai, J. and A. Szeidl (2018). "Interfirm Relationships and Business Performance," The Quarterly Journal of Economics, 133(3): 1229–82.

Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H., McKenzie, D., and M. Mensmann (2017). "Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa," Science, 357(6357): 1287–90.

Cartwright, N. and J. Hardie (2012). Evidence-Based Policy: A Practical Guide to Doing It Better. Oxford, UK: Oxford University Press.

Cherrier, B. (2019). "Weekly Lecture Was on 'What Should Come First: Theory or Data?" So Here's Tweetstorm on the History of Quantitative Economics." Twitter, March 13, twitter.com/Undercoverhist/status/1105851715461570560.

Cohen, J. and W. Easterly (2010). What Works in Development? Thinking Big and Thinking Small. Washington, DC: Brookings Institution Press.

Cook, T. (2018). "Twenty-six assumptions that have to be met if single random assignment experiments are to warrant 'gold standard' status: A commentary on Deaton and Cartwright," Social Science & Medicine, 210: 37–40.

Crépon, B., Devoto, F., Duflo, E., and W. Parienté. (2019). "Verifying the Internal Validity of a Flagship RCT: A Review of Crépon, Devoto, Duflo and Parienté': A Rejoinder," DIAL Working Paper 2019–07A.

Deaton, A. and N. Cartwright (2018). "Understanding and Misunderstanding Randomized Controlled Trials," Social Science and Medicine, 210: 2–21.

Dehejia, R., Pop-Eleches, C., and Samii, C. (2019). "From Local to Global: External Validity in a Fertility Natural Experiment," NBER Working Paper 21459.

Dubner, S. J. (2018). Is the Protestant Work Ethic Real? Freakonomics Podcast, Episode 360. December 5, 2018. Online: http://freakonomics.com/podcast/religiosity/

Dupas, P., Karlan, D., Robinson, J., and D. Ubfal (2018). "Banking the Unbanked? Evidence from Three Countries," American Economic Journal: Applied Economics, 10(2): 257–97.

Evans, D. (2016). "That Zero Effect May Not Mean What You Think It Means, and Other Lessons from Recent Educational Research." Development Impact Blog, World Bank. January 21, 2016. Online: https://blogs.worldbank.org/impactevaluations/zero-effect-may-not-mean-what-you-think-it-means-and-other-lessons-recent-educational-research

Freedman, B. (1987). "Equipoise and the Ethics of Clinical Research," The New England Journal of Medicine, 317(3): 141–5.

Garchitorena, A., Murray, M., Hedt-Gauthier, B., Farmer, P., and M. Bonds (2019). "Reducing the Knowledge Gap in Global Health Delivery: Contributions and Limitations of Randomized Controlled Trials," In Bédécarrats, F., Guérin, I., and F. Roubaud (eds.), Randomized Control Trials in the Field of Development: A Critical Perspective. New York and Oxford: Oxford University Press, Chapter 5.

Gelman, A. (2018). "Benefits and Limitations of Randomized Controlled Trials: A Commentary on Deaton and Cartwright," Social Science & Medicine, 210: 48–9.

Glennerster, R. (2016). Not So Small. Running Randomized Evaluations, May 27, 2016, Online: http://runningres.com/blog/2016/5/27/not-so-small

Gugerty, M. K. and D. Karlan (2018). "Ten reasons not to measure impact—And what to do instead", Stanford Social Innovation Review, Summer, 1–18.

Hammer, J. (2014). The Chief Minister Posed Questions We Couldn't Answer. Building State Capacity Blog, Harvard University. Online: https://buildingstatecapability.com/2014/04/08/the-chief-minister-posed-questions-we-couldnt-answer/

Harrison, G. (2011). "Randomization and Its Discontents," Journal of African Economies, 20(4): 626–52. https://doi.org/10.1093/jae/ejr030.

Imbens, G. (2018). "Comments on Understanding and Misunderstanding Randomized Controlled Trials: A Commentary on Deaton and Cartwright," Social Science and Medicine, 210: 50–2.

Ioannidis, J. (2018). "Randomized Controlled Trials: Often Flawed, Mostly Useless, Clearly Indispensable: A Commentary on Deaton and Cartwright," Social Science & Medicine, 210: 53–6.

Kaplan, R. and V. Irvin (2015). "Likelihood of Null Effects in Large NHLBI Clinical Trials Has Increased over Time," PLoS One, 210(8): e0132382.

Karing, A. (2018). "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone," Working Paper. https://drive.google.com/file/d/1Gq59ismP9V6I2pUzuLr iMVC5t6y2MqX-/view

McKenzie, D. (2018). "Six Questions with Mark Rosenzweig," Development Impact Blog, World Bank. January 10, 2018. https://blogs.worldbank.org/impactevaluations/six- questions-mark-rosenzweig. Last accessed November 12, 2019.

McKenzie, D. (2019). "Discussant's Comments," in K. Basu, D. Rosenblatt, and C. P. Sepulveda (eds.), State of Economics, State of the World. Cambridge, Mass.: MIT Press, forthcoming.

Meager, R. (2019). "Understanding the Average Impact of Microcredit Expansion: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," American Economic Journal: Applied Economics, 11(1): 57–91.

Meyer, M., Heck, P., Holtzman, G., Anderson, S., Cai, W., Watts, D., and C. Chabris (2019). "Objecting to Experiments that Compare Two Unobjectionable Policies or Treatments," PNAS, 116 (22): 10723–8.

Muralidharan, K., Niehaus, P., and S. Sukhtankar (2016). "Building State Capacity: Evidence from Biometric Smartcards in India. American Economic Review," 106(10): 2895–2929.

Muralidharan, K., Niehaus, P., and S. Sukhtankar (2018). "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India." NBER Working Paper 23838.

Niehaus, P. (2019). "RCTs: Why Scale Matters," VoxDev video, https://youtu.be/fD6MgGM5jWI

Ogden, T. N. (2017). Experimental Conversations: Perspectives on Randomized Trials in Development Economics. Cambridge, Massachusetts: MIT Press.

Pritchett, L. (2010a) "Is Microfinance a Shumpeterian Dead End?," Center for Global Development Blog, March 10, 2010. http://blogs.cgdev.org/open_book/2010/05/is-microfinance-a-schumpeterian-dead-end.php. Last accessed August 15, 2019.

Pritchett, L. (2020). "Randomizing Development: Method or Madness?," In Bédécarrats, F., Guérin, I., and F. Roubaud (eds.), Randomized Control Trials in the Field of Development: A Critical Perspective. New York and Oxford: Oxford University Press, Chapter 2.

Pritchett, L. and J. Sandefur (2015). "Learning from Experiments when Context Matters," American Economic Review: Papers and Proceedings, 105(5): 471–5.

Ravallion, M. (2020). "Should the Randomistas (Continue to) Rule?," In Bédécarrats, F., Guérin, I., and F. Roubaud (eds.), Randomized Control Trials in the Field of Development: A Critical Perspective. New York and Oxford: Oxford University Press, Chapter 1.

Student (1938). "Comparison between Balanced and Random Arrangements of Field Plots," Biometrika, 29(3/4): 363–78.

Teele, D. L. (ed.) (2014). Field Experiments and Their Critics. Essays on the Uses and Abuses of Experimentation in the Social Sciences. New Haven & London: Yale University Press.

Ravallion, M. (2009a). "Should the Randomistas Rule?," Economists' Voice, 6(2): 1–5.

Ravallion, M. (2020). "Should the Randomistas (Continue to) Rule?," In Bédécarrats, F., Guérin, I., and F. Roubaud (eds.), Randomized Control Trials in the Field of Development: A Critical Perspective. New York and Oxford: Oxford University Press, Chapter 1.

Vivalt, E. (2019). “Specification Searching and Significance Inflation across Time, Methods and Disciplines,” Oxford Bulletin of Economics and Statistics, 81(4): 797–816.

Whittle, D. (2011). “If Not Randomized Trials, Then What?,” Pulling for the Underdog Blog, June 1, 2011. https://www.denniswhittle.com/2011/06/randomized-trials-not-silver-bullet.html. Last accessed November 12, 2019.

Wilke, A. and M. Humphreys (2019). “Field Experiments, Theory and External Validity,” Working Paper, https://www.dropbox.com/s/47s52xv0frrnvml/20190703_Wilke_ Humphreys.pdf ?dl=0

Young, A. (2019). “Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results,” Quarterly Journal of Economics, 134(2): 557–98.

**Appendix: Full Quotations**

1. What methods are best to use and in what combinations depends on the exact question at stake, the kind of background assumptions that can be acceptably employed, and what the costs are of different kinds of mistakes. –Nancy Cartwright and Angus Deaton (Cartwright and Deaton)

2. We should neither be encouraging or discouraging any particular tool just for the sake of the tool. We should be encouraging students to look for an interesting question and use the right tool to answer it. Period. –Dean Karlan

3. Often just very good descriptive data that focuses people's attention on something they haven't focused on before has changed people's minds in policy as much as any experiment. —David McKenzie

4. The novelty maybe drove the overselling of RCTs, like these silly statements that everything ought to be evaluated randomly, or the people who say they don't believe any observational evidence. – Chris Blattman

5. I think those working on instrumental variables and those of us working on RCTs were motivated by the same impulse, the concern that a lot of empirical work in economics at the time was potentially subject to confounders and required a lot of fairly strong assumptions. That being said, it's not like IV makes all the problems disappear, and neither does an RCT. I don't think anybody thinks that RCTs are magical. –Michael Kremer

6. …if I had to choose, I might even say we should pour more energy into the big stuff than the small stuff. I do think that's part of how the randomized evaluation movement was sold to policy makers: "You're going to get answers." I don't think that's what we're going to get. My sense is that we're going to see evaluations that are all over the map. —Chris Blattman

7. Organizations should be able to draw on different areas to answer the relevant questions…I see a lot of crossover between different forms of causal identification. So I think focusing, yes, but I don't think you just have to focus on randomized evaluations. I don't think that makes sense. –Rachel Glennerster

8. An impact evaluation should help determine why something works, not merely *whether* it works. Impact evaluations should not be undertaken if they will provide no generalizable knowledge on the "why" question— that is, if they are useful only to the implementing organization and only for that given implementation. This rule applies to programs with little possibility of scale, perhaps because the beneficiaries of a particular program are highly specialized or unusual, or because the program is rare and unlikely to be replicated or scaled. If evaluations have only a one-shot use, they are almost always not worth the cost.—Dean Karlan and Mary Kay Gugerty (Karlan and Gugerty 2018)

9. There are many banal and useless examples of studies using every specific method. —Mark Rozenzweig (McKenzie 2018)

All quotations are from Ogden 2017 unless otherwise specified: Cartwright and Deaton (Deaton and Cartwright 2018), Gugerty and Karlan (Gugerty and Karlan 2018) and Mark Rozenzweig (McKenzie 2018).