Randomization in the tropics revisited: a theme and eleven variations

Angus Deaton

Woodrow Wilson School, Princeton University
Schaeffer Center for Health Policy and Economics, University of Southern California
National Bureau of Economic Research

October 2019

Written for Florent Bédécarrats, Isabelle Guérin and François Roubaud, *Randomized controlled trials in the field of development: a critical perspective*, Oxford University Press. For (generous, helpful, and amazingly rapid) comments on an earlier version, I am most grateful to Nancy Cartwright, Anne Case, Shoumitro Chatterjee, Nicolas Côté, Jean Drèze, William Easterly, Reetika Khera, Lant Pritchett, Dean Spears and Bastian Steuwer.

Development economists have been using randomized controlled trials (RCTs) for the best part of two decades[1], and economists working on welfare policies in the US have been doing so for much longer. The years of experience have made the discussions richer and more nuanced, and both proponents and critics have learned from one another, at least to an extent. In this essay, I do not attempt to reconstruct the full range of questions that I have written about elsewhere[2]. Instead, I focus on a few of the issues that are prominent in this volume of critical perspectives.

The RCT is a useful tool, but I think that is a mistake to put method ahead of substance. I have written papers using RCTs[3]. Like other methods of investigation, they are often useful, and, like other methods, they have dangers and drawbacks. Methodological prejudice can only tie our hands. Context is always important, and we must adapt our methods to the problem at hand. It is not true that an RCT, when feasible, will always do better than an observational study. This should not be controversial, but my reading of the rhetoric in the literature suggests that the following statements might still make some uncomfortable, particularly the second: (a) RCTs are affected by the same problems of inference and estimation that economists have faced using other methods, and (b) no RCT can ever legitimately claim to have established causality.

My *theme* is that RCTs have no special status, they have no exemption from the problems of inference that econometricians have always wrestled with, and there is nothing that they, and only they, can accomplish. Just as none of the strengths of RCTs are possessed by RCTs alone, none of their weaknesses are theirs alone, and I shall take pains to emphasize those facts. There is no gold standard. There are good studies and bad studies, and that is all. The most important things I have to say are about the ethical dangers of running RCTs in poor countries. I save those remarks for last.

1. Are RCTs the best way of learning, or of accumulating useful knowledge?

Sometimes. Sometimes not. It makes no sense to insist that any one method is best, provided only that it is feasible. It has always seemed to me to be a mistake for J-PAL to do only RCTs, and thus leave itself open to the charge that it is more (or as) interested in proselytizing for RCTs than it is in reducing poverty. Though as Tim Ogden notes[4], the *members* of J-PAL use a wide range of techniques in their own work, so perhaps J-PAL is just the RCT wing of a broader enterprise. Martin Ravallion is exactly right[5] when he argues that the best method is *always* the one that yields the most convincing and relevant answers in the context at hand. We all have our preferred methods that we think are underused. My own personal favorites are cross-tabulations and graphs that stay close to the data; the hard work lies in deciding what to put into them and how to process the data to learn something that we did not know before, or that changes minds. An appropriately constructed picture or cross-tabulation can undermine the credibility of a widely believed causal story, or enhance the credibility of a new one; such evidence is more informative about causes than a paper with the word "causal" in its title. The art is in knowing what to show. But I don't insist that others should work this way too.

The imposition of a hierarchy of evidence is both dangerous and unscientific. *Dangerous* because it automatically discards evidence that may need to be considered, evidence that might be critical. Evidence from an RCT gets counted even if when the population it covers is very different from the population where it is to be used, if it has only a handful of observations, if many subjects dropped out or refused to accept their assignments, or if there is no blinding and knowing you are in the experiment can be expected to change the outcome. Discounting trials for these flaws makes sense, but doesn't help if it excludes more informative non-randomized evidence. By the hierarchy, evidence without randomization is no evidence at all, or at least is not "rigorous" evidence. An

observational study is discarded even if it is well-designed, has no clear source of bias, and uses a very large sample of relevant people.

Hierarchies are *unscientific* because the profession is collectively absolved from reconciling results across studies; the observational study is wrong simply because there was no randomization. Such mindless neglect of useful knowledge is thankfully rare in economics, but there are many examples in other fields, such as medicine or education. Yet economists frequently do give special weight to evidence from RCTs based on methodology alone; such studies are taken to be "credible" without reference to the details of the study or consideration of alternatives.

Economics is an open subject in the sense that good studies that produce new, important, and convincing evidence are usually judged on their merits. But it is good to be careful that merit not be a cover for methodological prejudice. When I hear arguments that RCTs have proved their worth by producing good studies, I want to be reassured that the use of randomization is not itself a measure of worth and that the argument is not circular.

2. Statistical inference is simpler in RCTs than with other methods

This misunderstanding has been responsible for much mischief. There are two parts to the simplicity argument. First, randomization guarantees that the two groups, treatments and controls, are on average identical before treatment, so that any difference between them after treatment must be caused by the treatment. Second, statistical inference requires computing a *p*-value for the difference between two means, a simple procedure that is taught in elementary statistics classes.

Both parts of the argument are wrong.

R. A. Fisher understood from the beginning that randomization does *not* balance observations between treatments and controls, as anyone who actually runs an RCT will quickly discover. Ravallion[6], who has long observed RCTs in the World Bank and elsewhere argues that the

misunderstanding "is now embedded in much of the public narrative" in development. It is also common in everyday parlance.

Imagine four units (villages, say), two of which are to be treated, and two not. One possibility is to let the village elders decide, for example by bidding to be included (or excluded), and then selecting for treatment the two villages who most want (least do not want) to be treated. This self-selection allocation of treatments and controls is clearly problematic. Yet many people seem to think that randomization fixes the self-selection. There are only six possible allocations, one of which is the self-selected allocation. We then have the absurdity that the *same* allocation is fine if it comes about randomly, but not if it is self-selected. With hundreds of villages, whether or not balance happens depends on how many factors have to be balanced, and nothing stops the actual allocation being the self-selected allocation that we would like to avoid. Nothing is *guaranteed* by randomization. Perhaps it is the idea that randomization is fair *ex ante* that confuses people into thinking that it is also fair *ex post*. Here, it is the *ex post* that matters.

Making the treatment and control groups look like one another is a good thing but requires information and deliberate allocation, both of which are scrambled by randomization. Fisher knew this and knew that there were more precise ways of estimating an average treatment effect by avoiding randomization, but understood that there was a difficulty in knowing what to think about the difference once measured; there will always be *some* difference even when the treatment has no effect for any unit. Randomization is a solution to this problem, because it provides the basis for making probabilistic statements about whether or not the difference arose by chance. Many years ago, the philosopher Patrick Suppes put it this way[7]. He imagined himself presented with an urn with fifty black and white balls; there are either (A) fifteen black and thirty-five white, or (B) thirty-five black and fifteen white. He is allowed to draw twelve balls, and must bet on A or B. He wrote "I find it hard to imagine a sophisticated bettor who would not insist on . . . randomization before

entering into the experiment." Randomization *does not* balance, but it *does* allow the calculation of odds, at least in simple cases like this where nothing else affects the outcomes. Calculating odds is useful and important, but it is not the same as balance.

Many people are surprised when they are told that inference about a mean—and therefore inference about the difference between two means—is an unsolved problem. One issue was stated long ago by Bahadur and Savage[8], who showed that without assumptions that limit skewness, the calculated *t*–value will generally not have the *t*–distribution. If we wrongly assume that it does, we will make mistakes, for example, by thinking that a large *t*-value indicates an effect of the treatment when, in fact, there is none. Skewness (a term that nowadays is often incorrectly used to mean bias) refers to the third moment, and in particular the presence of large outliers on one side of the distribution. Any experiment involving money is a likely example, and one can think of educational or microfinance experiments where one or two people are immensely talented, and the others not so much.

The RAND health experiment—one of the most famous RCTs in economics—had one participant who had a difficult and immensely expensive pregnancy. The outcome of the RCT then depends on whether the outlier(s) is among the treatments or among the controls, and with an extreme enough outlier, on little else. You may think you have hundreds or thousands of observations, but in fact you only have one. Wild answers look significant, because the use of the *t*-distribution is invalidated by the skew. Trimming of outliers, or transforming the outcome variable—e.g. by taking logs—will not always help. The million-dollar baby is what will break an actual insurance scheme, however much the insurers might wish to "trim" it. We need to measure profits in dollars, not in the logarithms of dollars, let alone trimmed dollars. Perhaps the median treatment effect might be more reliable but, once again, it is the mean that breaks the budget, not the median, and even when we would like to know the median treatment effect, it is not identified

from an RCT. If you are genuinely interested in the median, you will have to use a method other than an RCT.

The point is *not* that RCTs have unique difficulties here, the point is that they have no exemption from such troubles, no "get out of jail free" card. Ulrich Mueller has recently shown that the problem is widespread in contemporary applied economics, particularly when using clustered robust standard errors[9]. When clusters are of different sizes—as in much spatial work in applied econometrics—the *p*–values that come from STATA, for example, are not reliable. My guess is that Mueller's work, which also provides a repair, will lead to substantial revisions in how we work, and in what we think we know.

In work on a related disease of inference, Alwyn Young has demonstrated that many published papers using RCTs get their *p*-values wrong[10], so that many apparently significant results—sometimes quite startling results—are consistent with the operation of chance in the situation where the treatment has no effect. Young proposes that we return to Fisherian randomization as a way of calculating significance. If the treatment has no effect for anyone, and there is no post-randomization confounding, the estimated average treatment effect is a result only of the random allocation of subjects to treatments or controls. (Post-randomization confounding is anything other than the treatment that effects outcomes, such as "tells" in the treatment environment, or non-blinding of subjects, assessors, or analysts.) By looking at all possible random assignments in the actual data, we can tabulate the distribution of the differences in the two means, and calculate the probability of getting something as or more extreme than the actual difference. This "randomization inference" tests the hypothesis that the treatment has no effect for *any* individual. This hypothesis is often of interest, but it is not relevant to what we often want to know for policy, which is whether the *average* treatment effect is zero. While a zero effect for each observation means that the average must also be zero, the converse is not true, most notably so

when the treatment affects different individuals in opposite directions. A small daily dose of aspirin is an example; it saves some and kills others. In public policy, say in a teaching experiment, we might well want to know whether the new method increases test scores on average, not just whether it works for someone. (An additional complexity is that a statistical test can sometimes accept the hypothesis each of a group of estimates is zero, but reject the hypothesis that their average is zero.)

Because the calculated significance levels are unreliable in realistic situations, it is wise to be skeptical of many of the published conclusions from RCTs. *Poor Economics*[11] presents the findings of dozens of studies, many of which are interesting and important. But results that ought to be presented as estimates tend to be presented as if they are established facts. Indeed, the rhetoric of RCTs is that trials can establish the truth. They cannot. The surprising results that come out of RCTs are sometimes not results at all, and large *t*–values ought not to persuade us that they are.

3. RCTs are rigorous and scientific

This rhetoric is rarely if ever justified. The adjectives are used as code words for RCTs. Frequently so. The rhetoric appears to be successful, at least with funders. It is often coupled with an appeal to the importance of RCTs in medicine, but rarely coupled with a realistic reading of the successes and failings of RCTs in medicine. In the US, drugs require positive RCTs in order to be licensed, yet prescription opioids, such as OxyContin, have killed hundreds of thousands of Americans in the last twenty years. There are differences between how RCTs work in social science and in medicine, a topic on which more thinking could usefully be done. On one occasion, I discussed a series of development trials with a senior funding manager of a large foundation. He was happy to admit that the results were limited in applicability, and that some of the results were likely incorrect, but was unimpressed. RCTs, after all, he told me, are more rigorous than any other method and for him, that was enough. I think he had a notion that rigor meant that the results were generalizable, or could be

scaled up. Or perhaps he held the common belief that all other methods are worse. Being wrong did not appear to conflict with being rigorous.

4. External validity

"Finding out what works" is another common rhetorical slogan that, at least judged by its repetition, is effective among the public. Nothing works except in context, and finding out what works where and under what circumstances is a real scientific endeavor. What works depends also on for whom and for what purpose; finding out what works is also a matter of values. There is no experiment or series of experiments that can answer such questions unconditionally. That RCTs will identify what works to eliminate global poverty is a commendable but unfounded aspiration.

A result that is true in one place, at one time, and under one set of circumstances, will typically not be true in another place, another time, or under different circumstances. What works for you may work for me too, except that I don't like it. Once again, these things are true of all empirical findings, no matter what method is used. No one thinks that an estimate of the average income in America will be accurate a decade from now, yet an estimate of an average treatment effect, which is also a sampling-based estimate of a mean, is often treated as if it is likely to hold elsewhere, at least in the absence of evidence to the contrary.

The practice is perhaps not very different from a long-standing practice in economics to treat elasticities as constants, as in "the" elasticity of labor supply of prime age men, or "the" price elasticity of bread. My suspicion is that those elasticities are supported by strong intuitions about the nature of the goods concerned, that most men had little choice but to work, while, once upon a time, their wives had more, that staple foods are not easily substituted for, and that the demand for small luxuries is sensitive to their prices, intuitions that were supported by many studies in many places. But this is not where we are with development today. To take Lant Pritchett's example[12], I

see no reason to suppose that if chickens are better than money in Sierra Leone, they will be better than money in Laos or, for that matter in Trenton, New Jersey, nor why, if they were better in sixty trials in sixty different places, they would be better in the sixty-first. This matters. The Gates Foundation, the largest aid donor in many areas, sees scaling up as one of its central missions, and so has seized on one or two positive results in its African agriculture initiative as evidence that "it works," and extended "it" to other farms or other countries, without any theory of why it might or might not work elsewhere.[13] We have to face the truth that what works might be different from one farm to the next, something African farmers are likely to know, even if the experimenters do not.

It is a mistake to think of internal and external validity as twin properties that are ideally possessed by high quality studies. An RCT can be perfectly conducted using a large sample and hit the ATE on the nose. Whether it is externally valid is *not* a property of the *study* but a property of the *circumstances* in which it is to be used. There is nothing invalid about a study whose result does not apply elsewhere.

There is always a temptation to take an impressive study and push it beyond its original context. This too is true of observational and experimental studies alike. Raj Chetty and his coauthors have pioneered the use of merged administrative data to describe in extraordinary detail facts about the dynamics of inequality in the United States, and have so generated huge advances in knowledge. One important finding[14] is that, between 1989 and 2015, African Americans children were less likely than white children to move up the income distribution from their parents' position. Yet in many popular accounts in the press, "were" is replaced by "are," even though marriage and incarceration patterns have been changing in both groups. These are outstanding studies, among the very best in economics today, but they can make no more claim to external validity than can outstanding RCTs. Once again, the issue of external validity is general, and RCTs have no "get out of jail free" card. It may be that, without internal validity, a trial result is unlikely to hold elsewhere,

but it is certainly not true that internal validity implies external validity. I do not know of explicit claims to the contrary, but I have often been struck by the contrast between the care that goes into running an RCT and the carelessness that goes into advocating the use of its results. The phrase "primacy of internal validity," seems to justify such practices.

That the results of an RCT will be used in a context that is different from that in which it was done can inform the design of the trial to make it more useful. If we think that treatment effects are different in different subpopulations, then stratification by those subpopulations will not only improve the precision of the trial, but will also allow reweighting to a new situation. Scaling up will often affect potential variables that are constant across arms of the trial; for example, if an educational policy trains more students, wages are likely to fall, so that including a low wage arm of the trial might give useful information. The RCT can help provide the tools for modeling the policy consequences instead of simply leaping over or ignoring the gulf that lies between a trial and its implementation. But an RCT is unlikely to be enough by itself.

The fact that a given study replicates in different contexts in different countries—as in the study of graduation programs[15] in *Science*—is indeed surprising that the sign of the ATE is the same in all but one of the contexts—though it is unclear that the gains could be replicated by government workers facing realistic financial and political incentives, incentives that are quite different from those faced by highly-educated graduate assistants from abroad who want the project to succeed. Yet, in such a cross-country study it is not at all clear what replication means, what measure we want to be replicated, or what we can learn from replication. We might want something like the rate of return on investment, or perhaps the fraction of people lifted above some local or global poverty threshold per unit of international currency. Instead, the authors use the "effect size," which is the ATE standardized by the standard deviation of the treatment. In the words of Arthur Goldberger and Charles Manski[16], "standardization accomplishes nothing except to give the quantities in

noncomparable units the superficial appearance of being in comparable units. This accomplishment is worse than useless—it yields misleading inferences."

5. Pre-registration of trials

I unsuccessfully argued against the American Economic Association (AEA) requiring pre-registration of the trials whose results are to be published in its journals. I think it is a bad idea for the AEA to legislate on methods rather than assessing studies on their merits. In my experience as an economist and while serving on AEA committees, disagreements between economists that are, in truth, political or personal, are often presented as methodological differences. The AEA has, at least since the 1930s, been successful in avoiding schisms and has remained a broad church for economists of all stripes, and its presidents have ranged from Milton Friedman to Kenneth Galbraith, though I doubt they thought much of each other's methods. (Friedman tried unsuccessfully to block Galbraith's presidency.)

The problems of *p*-hacking, data mining, and specification searches are real enough. Funders who have spent large sums on an RCT often exert pressure to find at least one subgroup for which the treatment was effective. But, once again, such problems are not specific to RCTs. Some have indeed argued for preregistration for *all* studies, so that, before I start work on an observational study using the census, for example, I should notify the AEA—or perhaps the Census Bureau—of my data analysis plan. It is not clear where all this stops; must I report a conversation with a colleague or a finding that I read about in the newspaper that shapes my agenda or limits my choice of variables?

The findings of my own of which I am most proud have all had a large element of serendipity, though I was informed enough to know what I was looking *at*, even when I was looking *for* something else. None of these results would have appeared in a pre-analysis plan and would thus

not be publishable in the *Journal of Correctly Done Studies*. Bill Easterly has noted that Columbus could not have discovered America if he had been required to stick to a pre-analysis plan filed in a lockbox in Seville or Genoa[17]. I find it hard to believe that what Anne Case and I found on midlife mortality rates[18], results that were totally unexpected to us, came from data snooping. Though I can easily imagine a statistically-blinkered editor rejecting the paper because we could not produce the certificate of pre-registration that authorized our work on midlife mortality. The risk of stifling important but unexpected results is surely much worse than the risk of promoting fallacious ones.

6. *Experimentation: kick it and see*

I am all for experimentation.[19] But there is no logical connection between experimenting and randomization. Indeed, one might be wise, when directing one's kick, to be rather precise about one's aim; kicking at random is not advisable, and it might hurt. Randomization is about judging the significance of what has happened, not about designing a kick. The serious point here is that, in many cases, randomization is unhelpful for experimentation, it can turn a good experiment into a useless one. Information that we should be using to improve our study is scrambled.

The key laboratory experiments in economics did not use randomization[20]. The Industrial Revolution is often described as having come about by endless tinkering, not by randomization, which would have got in the way of purposeful trial and error. Another example I have used in the past[21] is the arcade video game, *Angry Birds*. The birds need to be fired at an angle from a catapult, and can sometimes be redirected, speeded up, or detonated in flight, the object being to kill the egg-stealing pigs that are hiding in inaccessible places. Given the immense number of combinations, a systematic set of RCTs would take unimaginably long, although a dexterous child can figure out the solution in minutes. There are many kinds of experiments where randomization is not required, or

would obscure the results. Randomization, after all, is *random* and searching for solutions at random is inefficient because it considers so many irrelevant possibilities, just as it did in Fisher's fields.

7. RCTs and other methods

In many discussions of RCTs, comparisons are drawn with other methods, typically instrumental variables (IV), regression discontinuity (RD) or difference in difference methods. But this is much too narrow a comparison. For someone who has lived with, used, and taught econometric methods for more than forty years, I watched the progression that led to RCTs. We used to run regressions of *y* on *x*, with much too little discussion of what generated the variation in *x*. We learned about differences in differences, instrumental variables and regression discontinuity as methods for purging unwanted variance from *x*, and creating two groups that were deemed to be identical apart from treatment. RCTs could be thought of as cleaner versions of IV, RD, or differences in differences, effectively reverting to regression but with a guaranteed assumption that *x* was randomly assigned. Given this history, we can see why an RCT seemed like the ultimate solution, as indeed it is when we think this way.

But as John Stuart Mill noted long ago[22], the "method of differences," which compares two groups, one treated, one not, is only one among many ways of making causal inference. Finding out the cause of a plane crash does not involve differences (or at least we might hope not), and the hypothetico-deductive method, which is how physicists say they work, does not involve differences, simply the making and checking of predictions. That is why graphs and cross-tabulations can be so powerful when they arrange data in a way that contradicts a mass of prior understanding about how the world works. More formally, the Cowles Commission developed a method of building causal models with careful attention to mechanisms, and with a language that emphasized causal structure and procedures for delineating which parts of the structure could or could not be estimated from

data. These models could be interrogated to test their predictions and the adequacy of the causal structure. Economists once used these methods more than they do today, and they comprised the main content of econometrics texts for many years, but my guess is that most graduate students in economics today would be hard pressed to define structural and reduced forms. Papers had a theory section, which developed checkable predictions, ideally predictions that are surprising and unique to the theory, which are checked out in the empirical section. Some of these methods can be interpreted as looking at differences between groups, but not all.

8. Small versus large

Lant Pritchett has provided a typically eloquent, funny, and passionate argument that it is growth that matters for poverty reduction, not "rigorous" (or not) project by project evaluation, whether of money or chickens[23]. In *Poor Economics*, Abhijit Banerjee and Esther Duflo argue the opposite, that it is only at the level of the "small" that we know what we are doing, so we must build knowledge trial by randomized trial.

The debate is (at least) as old as the World Bank. Here is a simplified history. The Bank started out with the small, doing projects, ports, roads, power plants, and the like. It became quickly obvious that evaluating projects using commercial criteria often did not improve people's lives, particularly in economies where prices were distorted by tariffs, marketing boards, rationing, or exchange controls. An early response by two groups of very distinguished economists was to develop shadow prices to replace the market prices. Partha Dasgupta, Stephen Marglin and Amartya Sen produced one set of methods for the United Nations[24], and Little and Mirrlees another for the OECD[25]. The latter was turned into a manual by Lyn Squire and Herman van der Tak for use in the World Bank[26]. Yet the calculations were sometimes elaborate, beyond the capabilities or inclinations of Bank lending officials whose own incentives were to move money quickly. And the rules must

have seemed incomprehensible to policymakers in the countries asked to implement them. As an example of the primitive state of project evaluation in much of the world, Lyn Squire later noted[27] that even the most elementary tool of project evaluation, the discounting of future benefits, was rarely used in borrowing countries left to themselves. If the economy was comprehensively distorted, there was surely little point in evaluating projects at market prices, and evaluation at shadow prices was not a feasible alternative.

The remedy was to switch from the small to the large, to fix the distortions first, and to get the macroeconomy right before doing project evaluation. Structural adjustment was the result.

In support of this, empirical analyses, like Pritchett's, showed that economic growth was the way to generate material poverty reduction. The great episodes of material poverty reduction in the world—particularly China and India—were driven by economic growth and by globalization. Aggregate growth came with growth in the small too, more jobs, more opportunities, more roads, more and better schools and clinics, but those were seen as springing up more or less spontaneously in an economy where rapid growth was ongoing. None of this explained *how* to stimulate economic growth. For this, cross-country regressions could help. These were widely criticized and are easily mocked, but yielded some useful knowledge, such as the importance of domestic investment—certainly a key in China, India, or Korea—of the provision of public goods, and that foreign aid, even at its best, was not likely to do much to stimulate growth by itself. They also systematized and disciplined the evidence, which was better than the country by country anecdotes (aka war stories) that had dominated much of the previous discussion. But we learned more about what slows growth than what speeds it up. All valuable, but hardly the keys to eliminating poverty through faster growth. No one, as far as I am aware, suggested that RCTs were the key to economic growth; it is hard to tell a story in which RCTs had any relevance for poverty reduction in China.[28]

The Bank was half right. The better macroeconomic management in many countries around the world, the better understanding of monetary policy and central banking, as well as of the costs of exchange rate undervaluation and commodity price taxation have all contributed to better growth and poverty reduction, especially given time to operate.[29] Economists today, adherents of the credibility revolution and of testing for causality, tend to dismiss such evidence on the grounds that, in their view, it is neither rigorous nor credible. (Yet they have no similar difficulty with the causal claim that RCTs are effective in reducing global poverty.)

Those who believe that external help can aid economic development need to square the circle. No one doubts the importance of the macro perspective, only that the tools to influence economic growth are limited. The micro level trials are often successful in themselves, but their role in diminishing poverty rates is largely a matter of faith. RCTs need a theory of implementation, or of scale up, that explains just how the results are to be used in practice. That has to include attention to unintended consequences—the effects of implementation on the actions of government and communities—which are usually not included in the end-points of the trial. General equilibrium effects need to be thought through; scaling up will change prices and behaviors that were held constant in the experiments. RCTs routinely make the assumption that spill-over effects do not exist (the SUTVA assumption), yet the assumption is routinely violated, for example in sanitation[30] or deworming projects. At the individual level, the treatment works and spill-overs on others are small and often cannot be (or are not) measured. Yet, at the aggregate level, the sum of the individually small spill-overs can negate or reverse the effect.

9. Models

There is a great attraction of being able to make policy recommendations without having to construct models. I understand the appeal, of allowing the data to speak, or of generating data that

speak for themselves, but I believe that attempts to do so are bound to fail. Interpreting an RCT always requires assumptions. We need to assume that it is only the treatment that matters, which is impossible to guarantee without careful policing of post-randomization confounding, just as it is impossible to be sure that the exclusion restrictions are valid for estimation using instrumental variables. People do not always accept their assignment, which can be handled by using intent to treat estimation, though the intent-to-treat average treatment effect is often not what we need to know. Or we can build models of why people do or do not accept their assignments, which is in itself potentially useful information, see Heckman and Pinto in this volume.[31] What happens if an RCT gives a positive effect when the outcome is measured in levels, but a zero effect when measured in logarithms? Such cases are easy to construct.[32]

As practitioners are aware, the use of prior information will improve precision.[33] In practice, average treatment effects are often estimated by running a regression that includes control variables. These have to be chosen, and it is not clear by what rules variables are included or excluded, or how many to use. Stratification can increase precision too, but only if the stratification uses valid prior information about differences in the average treatment effects across strata.

The *use* of trial results is where modelling becomes essential. We need some theory to tell us whether the results have relevance elsewhere, and if so, how to adapt them.

10. Causality

A well-designed RCT will tell us *something* about causality. Yet, once again, there are many assumptions that need to be made to get from the data to the conclusion. In any finite trial, and there are no others, the possibility that the result is due to chance can never be ruled out. The measurement of the outcomes may matter, as in the example of levels versus logarithms. To quote Alex Broadbent, Jan Vandenbroucke and Neil Pearce,[34] "Causal conclusions do not follow

deductively from data without a strong set of auxiliary assumptions, and these assumptions are themselves not deductive consequences of the data." In the same paper they write, "we suggest that it is good practice to refrain from calling any individual study's estimate `causal' even if it is a randomized trial. It is the totality of the evidence that leads to the verdict of causality. Causality is a scientific conclusion, a *theoretical* claim, and as such transcends any individual study." (italics added). Causality is in the mind, not the data, an idea that Heckman and Pinto trace back to Frisch and Haavelmo.[35] The triangulation of results, or learning about causal processes from many studies over time, is well-illustrated by the chapter on sanitation in this volume.[36]

It is worth noting that it is not just the results of an RCT that may fail to transport, but causality itself. Nancy Cartwright and Jeremy Hardie[37] illustrate with a Rube Goldberg machine in which opening a window leads, through a long chain of preposterous but effective causal connections, to a pencil being sharpened by a woodpecker. Yet opening windows does not usually sharpen pencils, and a causal chain in one setting may be quite different in another setting. My impression is that when economists put the word "causal" in the titles of their paper, they are claiming more than a single instance in a specific context. Beware of Rube Goldberg.

That there are other ways of building causal models is well-known to economics students brought up in the Cowles tradition, or to readers of Judea Pearl.[38] Pearl argues that we have to *start* with a causal model and then use it to confront the data and to test its structure and, like the Cowles Commission before him, offers a series of tools and methods to do so. The wisdom of Austin Bradford-Hill's discussion[39] of the many ways to detect causality seems to be little referred to in economics; Bradford-Hill was the pioneer in randomized clinical trials seventy years ago and it sometimes seems as if we are losing knowledge, not gaining it.

11. Ethics

It is good that economists should think about the ethics of experimentation. I have very little to add to the discussions about equipoise and informed consent that are covered elsewhere in this volume.[40] Yet some of the development RCTs seem to pose challenges to the most basic rules. How is informed consent handled when people do not even know they are part of an experiment? Beneficence is one of the basic requirements of experimentation on human subjects. But beneficence for whom? Foreign experimenters or even local government officials are often poor judges of what people want. Knowing what is good for other people is not an appropriate basis for beneficence.

Ethics also require us to be realistic about what RCTs can and cannot do. Ethical lapses are more easily justified for those who subscribe to the hierarchy view, that the only evidence that counts is evidence from RCTs, thus ruling out options that might pose fewer risks to subjects and might lead to better conclusions. Telling developing country policymakers that RCTs are the only way of gathering evidence for policy is unethical, because it can cause them to ignore important information. The previously discussed issues of getting *p*-values right is relevant here too. An underpowered trial that cannot possibly establish its aims is also unethical when it imposes burdens on subjects.

My main concern is broader. Even in the US, nearly all RCTs on the welfare system are RCTs done *by* better-heeled, better-educated and paler people *on* lower income, less-educated and darker people. My reading of the literature is that a large majority of American experiments were not done in the interests of the poor people who were their subjects, but in the interests of rich people (or at least taxpayers) who had accepted, sometimes reluctantly, an obligation to prevent the worst of poverty, and wanted to minimize the cost of doing so.[41] That is bad enough, but at least the domestic poor get to vote, and are part of the society in which taxpayers live and welfare operates,

so that there is a feedback from them to their benefactors. Not so in economic development, where those being aided have no influence over the donors. Some of the RCTs done by western economists on extremely poor people in India, and that were vetted by American institutional review boards, appear unethical, sometimes even bordering on illegality, and likely could not have been done on American subjects.[42] It is particularly worrying if the research addresses questions in economics that appear to have no potential benefit for the subjects. Institutional review boards in the US have special protection for prisoners, whose autonomy is compromised; there appears to be no similar protection for some of the poorest people in the world. There is an uncomfortable parallel here with the debates about pharmaceutical countries testing drugs in Africa.

I see RCTs as part of what Bill Easterly calls the "technocratic illusion,"[43] that is the original sin of economic development, an aspect of what James Scott[44] has called "high modernism," that technical knowledge, even in the absence of full democratic participation, can solve social problems. According to this doctrine, which seems especially prevalent in Silicon Valley, among foundations, and in the effective altruism movement, global poverty will yield to the right technical fixes, one of which is the adoption of RCTs as the basis for evidence-based policy. Ignoring politics is seen as a virtue, not the vice that it is. Foundations and altruists often "know" what is good for poor people, and have the best intentions, but provide little evidence that poor people agree with their assessments or value their remedies, so that their interests can easily come to conflict with those they are trying to help. The technocrats believe that they can develop other people's countries from the outside, because they know how to find out what works. In this, at least, there is no great difference between designing a gadget and designing social policy. Both are exercises for engineers.

Engineering poverty reduction is at best hopeless, and at worst disastrous. Development agencies today use the word "partnership" a great deal, but there is no genuine partnership when all

the money is on one side. Nor can there be genuine informed consent in an RCT when aid money is at stake.

Finding out what works is not the same thing as finding out what is desirable. Good intentions by donors are no guarantee of desirability. Jean Drèze has provided an excellent discussion of the issues of going from evidence for policy.[45] One of his examples is the provision of eggs to schoolchildren in India, a country where many children are inadequately nourished. An RCT could be used to establish that children provided with eggs come to school more often, learn more, and are better nourished. For many donors and RCT advocates, that would be enough to push for a "school eggs" policy. But policy depends on many other things; there is a powerful vegetarian lobby that will oppose it, there is a poultry industry that will lobby, and another group that will claim that their powdered eggs—or even their patented egg substitute—will do better still. Dealing with such questions is not the territory of the experimenters, but of politicians, and of the many others with expertise in policy administration. Social plumbing should be left to social plumbers, not experimental economists who have no special knowledge, and no legitimacy at all.[46]

Working to benefit the citizens of other countries is fraught with difficulties. In countries ruled by regimes that do not care about the welfare of their citizens—extractive regimes that see their citizens as source of plunder—the regime, if it has complete control, will necessarily be the beneficiary of aid from abroad. This is most obvious in war zones where it is impossible to deliver aid without paying off the warmongers and prolonging or worsening the suffering.[47] The dilemma extends to peacetime too. In authoritarian regimes with full control, it is only possible for outsiders to help when it is in the government's interest to accept that help. Development agencies then find themselves in the situation of being "allowed" to help the poor, or to help provide health services, while providing political cover for the "enlightened" despot who is thereby free to persecute or eliminate his opponents.[48] Similar issues arise in democracies too, though less sharply; the step from

evidence to policy is never ethically neutral but is less fraught when the poor have a voice and some political power.

What does this have to do with RCTs? Irrelevance for one. It makes no sense to spend resources randomizing schools or medicines when the President, facing an election, is imprisoning his foes or inciting violence against his tribal and political enemies.[49] As larger numbers of the world's poor come to live in nominally democratic states with populist autocratic leaders, more and more ethical dilemmas will confront trialists. Why are agencies funding aid, or RCTs to support aid, in countries whose leaders do not accept the liberal democratic beliefs of the donors, or experimenters? I am not claiming there are no answers to this question, only that donors need to know what they are.

There have already been protests[50] about the Bill and Melinda Gates Foundation's award of one of its Global Goal Awards to Narendra Modi for building toilets in India, at a time when Modi is depriving Kashmiris of their rights, is removing citizenship from millions of Assamese, and is threatening to do the same to Muslims and other non-Hindus. The Foundation argues that the reward recognizes only Modi's achievements in sanitation; this is surely a perfect example of limitations and dangers of technocratic aid. It empowers despotism and intolerance. Modi has received other prestigious awards from development agencies, including the United Nations. And much worse has happened repeatedly in Africa.

Aid agencies are turning a blind eye to political repression so long as the oppressors help check off one the Sustainable Development Goals, preferably as demonstrated by randomized controlled trials. The RCT is in itself a neutral statistical tool but as Dean Spears notes[51], "RCTs provide a ready and high-status language" that allows "mutual legitimization among funders, researchers, and governments." When the RCT methodology is used as a tool for "finding out what

works," in a way that does not include freedom in its definition of what works, then it risks supporting oppression.

---

[1] The Nobel Prize to Abhijit Banerjee, Esther Duflo and Michael Kremer was announced as this essay was being revised. As it already has done, the Prize will raise the visibility of the debate about the pros and cons of conducting RCTs directed towards economic development. The extensive press discussion has revealed substantive concerns, especially about ethics. It also reveals widespread misperceptions, among both critics and defenders, about how RCTs actually work, particularly highlighting the widespread but false beliefs that randomization guarantees that the treatment and control groups are similar prior to treatment, and that an RCT can demonstrates causality.

[2] Angus Deaton, 2007, "Instruments, randomization, and learning about development," *Journal of Economic Literature*, 48 (2), 424–55. Angus Deaton and Nancy Cartwright, 2018, "Understanding and misunderstanding randomized controlled trials," *Social Science and Medicine*, https://doi.org/10.1016/j.socscimed.2017.12.005. Angus Deaton, 2010, "Understanding the mechanisms of economic development," *Journal of Economic Perspectives*, 24(3): 3–16.

[3] Angus Deaton, 2013, "The financial crisis and the wellbeing of Americans," *Oxford Economic Papers*, 64)1, 1–26. Angus Deaton and Arthur A. Stone, 2016, "Understanding context effects for a measure of life evaluation: how responses matter," *Oxford Economic Papers*, doi: 10.1093/oep/gpw022

[4] Timothy Ogden, 2020, "RCTs, the hype cycle, and a high plateau," Chapter 5, this volume

[5] Martin Ravallion, 2020, "Should the randomistas (continue) to rule?" Chapter 2, this volume.

[6] Ravallion, *op cit*

[7] Patrick Suppes, 1982, "Arguments for randomizing," *PSA: proceedings of the biennial meeting of the philosophy of science association*, Vol. 1982, Volume 2, Symposia and invited papers, 464–75.
http://www.jstor.org/stable/192437

[8] R. R. Bahadur and Leonard Savage, 1956, "The nonexistence of certain statistical procedures in nonparametric problems," *Annals of Mathematical Statistics*, 27(4), 1115–22.

[9] Ulrich Mueller, 2019, "Inference to the mean," in preparation

[10] Alwyn Young, 2019, "Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results," *The Quarterly Journal of Economics*, Volume 134, Issue 2, May 2019, Pages 557–598, https://doi.org/10.1093/qje/qjy029

[11] Abhijit Banerjee and Esther Duflo, 2011, *Poor Economics: a radical rethinking of the way to fight global poverty*, Public Affairs.

[12] Lant Pritchett, 2020, "Getting random right: a guide for the perplexed practitioner," Chapter 3, this volume.

[13] Rachel Schurman, 2018, "Micro(soft) managing a `green revolution' for Africa: the new donor culture and international agricultural development," *World Development*. 112, 180–92.
https://doi.org/10.1016/j.worlddev.2018.08.003

[14] Chetty, Raj, Nathaniel Hendren, Maggie R. Jones and Sonya R Porter, 2019, "Race and economic opportunity in the United States: an intergenerational perspective," NBER Working Paper 24441, (June)

[15] Abhijit Banerjee, Esther Duflo, Nathaneal Goldberg, Dean Karlan, Robert Osei, William Parenté, Jeremy Shapiro, Bram Thuysbaert and Christopher Udry, 2015, "A multifaceted program causes lasting progress for the very poor: evidence from six countries," *Science*, Vol. 348, Issue 6236, 1260799 doi: 10.1126/science.1260799

[16] Arthur S Goldberger and Charles F Manski, 1995, "Review article: the Bell Curve by Herrnstein and Murray," *Journal of Economic Literature*, 33(2), 769.

[17] William Easterly 2012,, https://nyudri.wordpress.com/2012/10/15/if-christopher-columbus-had-been-funded-by-gates/

[18] Anne Case and Angus Deaton, 2015, "Rising mortality and morbidity among midlife white non-Hispanics in 21st century America," *Proceedings of the National Academy of Sciences of the USA,* December 8, 2015 112 (49) 15078-15083; first published November 2, https://doi.org/10.1073/pnas.1518393112

[19] Jonathan Morduch, 2020, "RCTs are usefully disruptive," Chapter 4, this volume.

[20] Andrej Svorenčik, 2015, *The experimental turn in economics: a history of experimental economics*, https://dspace.library.uu.nl/bitstream/handle/1874/302983/Svorencik.pdf?sequence=1&isAllowed=y

[21] Angus Deaton, 2012, "Searching for answers with randomized experiments," Development Research Institute, NYU, video presentation https://www.youtube.com/watch?v=yiqbmiEalRU (March 22, 2102)

[22] John Stuart Mill, 1843, *A System of Logic, ratiocinative and deductive, being a connected view of the principles of evidence and the methods of scientific evidence*, Eight edition available at https://books.google.com/books/about/A_System_of_Logic_Ratiocinative_and_Indu.html?id=ndMLAAAAIAAJ&printsec=frontcover&source=kp_read_button#v=onepage&q&f=false
[23] Pritchett, *op cit*
[24] Partha Dasgupta, Stephen Marglin, and Amartya Sen, 1972, *Guidelines for project evaluation*, United Nations Industrial Development Organization.
[25] Ian Little and James Mirrlees, 1974, *Project appraisal and planning for developing countries*, Basic Books.
[26] Lyn Squire and Herman van der Tak, 1975, *Economic analysis of projects*, Johns Hopkins University Press for World Bank.
[27] Lyn Squire, "Project evaluation in theory and practice," in Chapter 21 in Hollis Chenery and T. N. Srinivasan, *Handbook of Development Economics*, Volume 2, pages 1126–27.
[28] Yao Yang, 2019, "The open secret of development economics," *Project Syndicate*, Oct 22.
[29] William Easterly, 2019, "In search of reforms for growth: new stylized facts on policy and growth outcomes," NBER Working Paper 26318, September.
[30] Dean Spears. Radu Ban and Oliver Cumming, 2020. "Trials and tribulations: the rise and fall of the sanitation RCT," Chapter 7, this volume
[31] James Heckman and Rodrigo Pinto, 2020, "Exploiting noncompliance to enhance causal inference of randomized controlled trials," Chapter 13, this volume.
[32] Consider a small binary treatment that changes log income by an amount *a* that varies over units, but averages to zero. The effect on individual income is *a.y*, where *y* is income. The mean of *a.y* depends on the correlation of income and the individual treatment effect, which can be positive, negative, or zero.
[33] Eva Vivalt, 2020, "Using priors in experimental designs: how much are we leaving on the table?" Chapter 12, this volume.
[34] Broadbent Alex, Jan P Vandenbroucke and Neil Pearce, 2017, "Formalism or pluralism? A reply to commentaries on `causality and causal inference in epidemiology," *International Journal of Epidemiology*, 1–12.
[35] James J Heckman and Rodrigo Pinto, 2015. "Causal analysis after Haavelmo," *Econometric Theory*, 31(1), 115-151.
[36] Dean Spears. Radu Ban and Oliver Cumming, op cit.
[37] Nancy Cartwright and Jeremy Hardie, 2012, *Evidence based policy: a practical guide to doing it better*, Oxford.
[38] Judea Pearl and Dana Mackenzie, 2018, *The book of why: the new science of cause and effect*, Basic Books.
[39] Bradford-Hill, Austin, 1965, "The environment and disease association or causation" *Proceedings of the Royal Society of Medicine*, 58, 295–300.
[40] Michel Abramowicz and Ariane Szafarz, 2020, "Ethics of RCTs: Should Economists Care about Equipoise?", Chapter 11 in this volume.
[41] Judith Gueron and Howard Rolston, 2013, *Fighting for reliable evidence*, Russell Sage.
[42] Anonymous Shrew (Ankur Sarin), 2019, "Indecent proposals in economics," https://docs.google.com/document/d/11FS4l8hKxPPgo7Qdtham4EkbLxW1uW9gXgxZuyNtTHQ/edit
[43] William Easterly, 2013, *The Tyranny of Experts: economists, dictators, and the forgotten rights of the poor*, Basic Books.
[44] James C. Scott, 1998, *Seeing like a state: how certain schemes for improving the human condition have failed, Yale.*
[45] Jean Drèze, 2018, "Evidence, policy, and politics," *Ideas for India*, August 3. https://www.ideasforindia.in/topics/miscellany/evidence-policy-and-politics.html
[46] Esther Duflo, 2017, "The economist as plumber," *American Economic Review*, 107(5), 1-26.
[47] Alex de Waal, 1997, *Famine crimes: politics and te disaster relief industry in Africa*, Currey.
[48] Angus Deaton, 2015, "The logic of effective altruism," *Boston Review*. http://bostonreview.net/forum/logic-effective-altruism/angus-deaton-response-effective-altruism
[49] Michela Wrong, 2009*, It's our turn to eat: the story of a Kenyan whistleblower*, Harper.
[50] Sabah Hamid, 2019, "Why I resigned from the Gates Foundation," *New York Times*, September 26. "Dismay at Gates Foundation prize for Narendra Modi," *The Guardian*, Letter, 23 September, 2019, "Bill and Melinda Gates Foundation under fire for award to Narendra Modi," *The Guardian*, 12 September 2019.
[51] Dean Spears, personal communication, October 14, 2019. Quoted with permission.