LAURA AND JOHN ARNOLD FOUNDATION

Why Government Needs More Randomized Controlled Trials: Refuting the Myths

Stuart Buck and Josh McGee

July 2015

I. Introduction

At the federal, state, and local level, there are literally thousands of social programs aimed at improving human welfare by increasing high school graduation rates, alleviating poverty and hunger, preventing teen pregnancy, and more. Yet, we know very little about whether these programs work as promised, let alone how to improve them. Today, governments regularly try new programs, while old programs remain in place without anyone ever asking how effective they are or whether there is a way to make them better. According to this line of thinking, good intentions are enough and evidence is beside the point.

We think that standard should be flipped. Evidence should matter *more* than intentions, not *less*. The government should routinely collect rigorous evidence. It should then use that evidence to improve programs and scale the ones that are most effective. At the same time, it should direct funds away from those that don't work.

The best way to learn whether a social program or policy works as intended is through a randomized controlled trial (RCT), in which people are randomly assigned to receive the program's services or to be part of a control group. RCTs produce the highest form of evidence because they make it possible to isolate the effect of a program from complicating factors, even those that are unseen. In the words of Guido Imbens, "Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top." Conducting more RCTs would help to ensure that we are spending public funds on programs that work and would tell us how to improve those programs as well.

Despite promising pilot efforts, RCTs remain vastly underused in the social sector.[1] Many skeptics have claimed that RCTs are inherently expensive, time-consuming, unethical, or not worth the trouble. Such objections are almost always overstated or false. This policy brief explains why RCTs are valuable, why they are often misunderstood, and why many common objections should be given little weight.

[1] More than a decade ago, the Institute of Education Sciences started preferentially funding RCTs in education. More recently, the Obama administration has pushed for RCTs in several important domestic policy areas, including job training programs and teen pregnancy programs.

II. The Value of RCTs — Two Examples

A. Medicaid

For decades, health policy analysts debated whether Medicaid actually helped its beneficiaries. Articles in top journals contended that Medicaid patients were worse off than patients with private insurance. This article, for example, claimed that "after adjusting for patient and hospital characteristics, patients with Medicaid were . . . 57% more likely to die postoperatively" than privately insured patients. Some experts even argued it might be better to be *uninsured* than to be on Medicaid (examples here and here).

The problem with all of these studies, however, is selection bias. People can *select* to apply or not to apply for Medicaid, and even those who can afford private insurance can *select* not to buy it. But people don't make selections at random. Of the people eligible for Medicaid, you would expect those who are elderly or are especially likely to become ill to sign up because they would be most worried about medical bills. By contrast, young people in the prime of their health might not bother signing up for Medicaid. The same goes for people who can afford private insurance—the ones who choose not to sign up even though they can afford it are probably healthier and younger.

Because of selection, the people on Medicaid and the uninsured are different in ways that we can't measure. Even if researchers adjust for observed differences like race and income, they won't ever be able to control for unobserved differences. For example, their datasets don't include variables indicating whether a person is a very conscientious individual who takes good care of his- or herself. Researchers comparing people on Medicaid to the uninsured would be unable to adequately distinguish between the effects of participating in Medicaid and the *unobserved* differences between the two groups of people. It is thus misleading to claim that being on Medicaid makes people less healthy than they would be if they had remained uninsured. The people who choose Medicaid *already* are very different from people who choose to remain uninsured, and Medicaid can't be blamed for differences in health outcomes down the road.

The same sort of bias will show up in any comparison between Medicaid recipients and people who have private insurance. Those two groups of people differ in all sorts of ways that can't be fully measured—the quality of their family doctor, their levels of health-consciousness, their ability to take a day off work, their involvement in a social network that provides advice and support, and more. Therefore, it is equally misleading to claim that private insurance makes people healthier than Medicaid when the difference is more likely due to all of the factors that make those two groups of people different from the outset.

Fortunately, there is a way around this problem of accounting for unmeasured differences when comparing two groups of people—an RCT. Health economists were fortunate when more than 90,000 people in Oregon signed up for Medicaid when access was expanded a few years ago and the state used a lottery to decide who would be offered coverage. The result was an RCT of Medicaid—the first RCT on health insurance in several decades.

The 90,000 people entered in the lottery had already made the same selection and gaining access to Medicaid depended on a random draw. This means that when they compared the treatment group (those who enrolled in Medicaid) to the control group (those who were not selected in the lottery), they didn't have to worry about unobserved differences between the two groups of people. The lottery took care of that. It used a random draw to automatically equalize the treatment and control groups across observed and unobserved differences from the outset, making it possible later on to say that differences in outcomes were really due to the treatment (in this case, the Medicaid offer), rather than to the treatment group having been different at the outset.

The results of the Oregon Medicaid randomized experiment refuted the earlier literature claiming that Medicaid made people less healthy. The experiment showed that people who gained access to Medicaid by way of the lottery, compared with those who wished to enroll but were not selected, were equally healthy in terms of blood pressure, cholesterol, and blood sugar. It also showed that their rates of depression went down by 30 percent and they were far better off financially. (For more on the results, see here.) Surprisingly, however, the Oregon Medicaid experiment also showed that people offered Medicaid were *substantially more likely* than the uninsured to visit the emergency room, contrary to claims that Medicaid recipients are less likely.

The Oregon Medicaid experiment thus showed how useful RCTs can be in demonstrating the true effects of a public program.

B. Hormone Therapy

In the 1980s and 1990s, researchers published numerous epidemiological studies showing that estrogen replacement therapy for menopausal women lowered their risk of heart attacks. A 1991 review of more than two dozen studies stated that estrogen's benefits were "unlikely to be explained by confounding factors."[2] Moreover, the protective effect of estrogen seemed to make perfect biological sense because estrogen also improved blood cholesterol by lowering LDL and increasing HDL. Overall, women who took estrogen therapy had a heart attack risk of only 56 percent the risk experienced by other women, according to the review.

In 1991, the Women's Health Initiative—one of the largest research projects ever—was launched. Among other things, it recruited nearly 11,000 menopausal women for an RCT on estrogen therapy. To everyone's surprise, the NIH had to stop the randomized trial early for ethical reasons. Not only did estrogen therapy *fail to prevent heart disease*, it also *raised the risk of stroke* by 39 percent. Even worse, a parallel RCT on hormone therapy also found that estrogen and progestin raised the risk of stroke as well as increased the risk of heart disease and breast cancer.

These negative results occurred even though the women on estrogen therapy still saw a significant reduction in LDL cholesterol and a significant increase in HDL cholesterol. In other words, estrogen therapy made cholesterol appear to move in the right direction, but it still increased the risk of stroke without preventing heart disease.

The Women's Health Initiative quickly became a cautionary tale in medicine. Based on observational and epidemiological studies, doctors had been treating millions of menopausal women with hormone therapy that

[2] An example of a confounding factor would be if women given estrogen therapy were wealthier on average than other women and had access to better health care in general.

made them *worse off in every possible way*. It cost the women money and time while actually putting them at far higher risk of conditions that cause death and disability.

Only RCTs were able to save millions more women from being mistreated in this way. Yet, despite their obvious benefits, RCTs continue to have many critics. In the next section, we address the most common objections to RCTs. These objections are usually overstated, if not altogether false.

III. Refuting Myths About RCTs

A. Myth 1: RCTs Are Expensive and Slow

In two typical examples of this critique, an essay in "Modern Healthcare" refers to "lengthy and expensive prospective RCTs," while a research scientist at a children's hospital claims, "RCTs are too expensive and too slow."

However, there is nothing inherently expensive about an RCT when compared to other forms of evaluation. If a program is over-subscribed (as was the case with Medicaid in Oregon), randomizing to decide who gets in creates a trivial additional cost—namely, the cost of implementing randomization rather than some other method for choosing program participants. Even for a treatment that is not over-subscribed, such as a math curriculum or a job training program, randomizing the recipients is not a substantial expense. Indeed, randomizing can even make a program substantially *less* expensive if the program initially serves only the treatment group, rather than everyone involved in the study.

It can be expensive, of course, to track the individuals or institutions that participate in an RCT in order to monitor their outcomes. But collecting data is an expense for *any* form of evaluation, not just RCTs. To the extent that critics complain about the cost of data collection, they are complaining about evaluation, not RCTs per se.

Moreover, it is easily possible to conduct RCTs of important social programs for as little as $100,000 or less, as the Coalition for Evidence-Based Policy has shown by funding such RCTs.[3] The reason some RCTs can be done so cheaply is because one or more agencies at the federal, state, and local level track key data *anyway,* such as school test scores, hospital admission rates, recidivism rates, or other outcomes of social importance. Researchers can then piggyback on that data collection without incurring the extra expense of collecting data from scratch. Likewise, medical researchers have recently begun doing registry-based clinical trials, which follow patients through national databases that collect electronic information on health outcomes. These trials can cost as little as 1 percent of a traditional RCT.

In any event, unduly worrying about expense seems beside the point. If we spent a mere 2 percent of the federal budget for social programs on RCTs, it would be an "expense" of a few billion dollars a year. However,

[3] The Laura and John Arnold Foundation (LJAF) funded the Coalition for Evidence-Based Policy's Low-Cost RCT Competition.

such unprecedented rigor would allow us to more effectively allocate many more billions of dollars that we now spend on untested programs that have not been proven to work. This is not to say that social programs should be cut; perhaps we should instead spend the same amount of money on *different* or *modified* social programs that are more evidence-based and actually accomplish their intended outcomes.

In addition, the idea that an RCT is “slow” is particularly odd. The real reason RCTs sometimes seem “slow” is because the outcome the evaluators chose to measure—such as high school graduation—unfolds over several years. But when the outcome takes that much time to unfold, so will *any* prospective evaluation—whether it is an RCT or not. If there is a high school program that aims to increase graduation rates over five years, leaving out an RCT will not make the five years elapse any more quickly. Instead, we will reach the end of the five years with no more knowledge than we had at the outset.

On top of that, RCTs are not limited to only measuring long-term outcomes. If there are shorter-term, intermediate outcomes that we care about in and of themselves, or that we are very sure are related to the longer-term outcomes, an RCT can be used to rigorously measure the effects of a program on those intermediate outcomes. For example, we might strongly believe that (1) college freshmen who are given help completing the Free Application for Federal Student Aid (FAFSA) are more likely to receive federal student aid, and (2) receiving federal student aid, in turn, makes them more likely to graduate from college. If it will take too long to measure graduation rates directly, we could do an RCT on the first question about the likelihood of recieving federal student aid and get a much quicker answer that could still affect graduation rates.

This is not to say that intermediate outcomes are always a reliable guide. For example, the Moving to Opportunity experiment from the 1990s, which gave low-income families housing vouchers to use in richer neighborhoods, was initially viewed as a bit of a disappointment because test scores for the children in those families did not significantly improve. But longer-term evidence from tax returns shows that, by the time the children reached their mid-20s, those who had been given a housing voucher before age 13 had incomes that were 31 percent higher than the control group—a substantial improvement.[4] The RCT allowed researchers to accurately measure the effect of the Moving to Opportunity treatment on both short- and long-term outcomes. Using another evaluation method would not have decreased the time required to measure these outcomes. It would simply have reduced our ability to tell whether the program worked or not.

B. Myth 2: RCTs Are Often Unethical

As an example of this common critique, the Education Writers Association informs its members that RCTs in education can be “unethical,” because “if researchers really believe an intervention will improve learning, for example, withholding it from a control group of students could be seen as unjustified.” An article in “The BMJ” spoofed the demand for RCTs by pointing out that: “As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evaluation by using randomised controlled trials. . . . We think that everyone might benefit if the most radical protagonists of evidence based medicine organised and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute.”

[4] LJAF provided funding for the Equality of Opportunity project from which this finding emerged.

Fair enough. If an intervention *really is* as obviously effective as a parachute, then it would be absurdly unethical to do an RCT in which the control group plummets to its death.

But the vast majority of social interventions are nothing like parachutes. With a parachute, we can see the effect happening before our eyes, and we also can vividly imagine what the effect of gravity would be with no parachute to slow down a person's descent. It is quite easy to infer that parachutes cause users not to experience the negative effects of gravity. By contrast, most social interventions operate in a much more causally complex environment, where it is impossible to infer causation through simple intuition, observation, or even before/after comparisons—all of which can be very unreliable.

The history of "scared straight" programs provides a good example of how wrong proponents can be about a program's effectiveness. These programs, originally launched in the 1970s, took at-risk juveniles on field trips to a prison in order to "scare" them into being more law-abiding. The idea behind the scared straight programs seemed so intuitive that Illinois even passed a law in 2003 requiring the Chicago Public Schools system to administer such a program for its at-risk youth.

Just as with Medicaid and hormone therapy, however, randomized trials of scared straight programs told an entirely different story. Such programs were, if anything, harmful. An authoritative review by the Campbell Collaboration starkly concluded: "Results of this review indicate that not only does [the program] fail to deter crime but it actually leads to more offending behavior. Government officials permitting this program need to adopt rigorous evaluation to ensure that they are not causing more harm to the very citizens they pledge to protect."

The history of social interventions is filled with discrepancies between observational studies and randomized trials. Teen pregnancy prevention is another example.[5] A review of 30 studies on teen pregnancy prevention programs found that "observational studies yield systematically greater estimates of treatment effects than randomized trials." Indeed, in 1996, Congress began spending $50 million a year on abstinence-based education. But when those programs were evaluated in no fewer than four RCTs, it turned out that all that spending was pointless—there was no evidence that the programs had any effect on teen pregnancy outcomes.[6]

The moral is that it is rare that a social intervention is so undeniably effective that doing an RCT would be unethical. To the contrary, *not* conducting RCTs to learn how a program works is unethical. When a social program is enacted and funded without an RCT, the government is *still experimenting* on the people served by that program. It's just that the experiment is happening without the benefit of a control group that would allow us to know whether the program worked, had no effect, or even caused damage.

Without an RCT, doctors would still be giving dangerous hormone therapy to millions of women and state

[5] More broadly, a review of studies on psychology, health, and education found that, for some treatments, "nonrandom designs (relative to random) tend to strongly underestimate effects, and in others, they tend to strongly overestimate effects." To be sure, recent work by Tom Cook and others shows that other methods can sometimes deliver the same answers as an RCT, although these methods are not always applicable.

[6] Even so, Congress decided in April 2015 to raise spending on the program to $75 million annually.

legislatures would still be ordering juveniles to take field trips to prisons. It is entirely possible that many other social interventions are also doing damage, despite their proponents' best intentions. Moreover, if programs are ineffective, then we are wasting resources that could be redirected to more effective programs.

C. Myth 3: RCTs Are Limited to Narrow Contexts or Questions

"Randomization provides a high degree of proof for a very narrow set of facts: a particular program, under a particular set of conditions, for a particular population of people, at a particular time, made a difference," writes Jason Saul of Mission Measurement. Or, as Christopher Barrett and Michael Carter put it, "There is a nontrivial probability that no external population exists to whom the results of the experiment apply on average." And William Easterly said that the ability to generalize is the "single biggest concern about what RCTs teach us."

External validity is a real problem, and it may often be true that, as Lant Pritchett and Justin Sandefur said, observational studies "from the right context are, at present, a better guide to policy than experimental estimates from a different context."

This criticism ignores, however, that it is often quite possible to do replication RCTs in diverse settings to determine whether and how the findings generalize. The Nurse-Family Partnership, for example, has been subject to at least three RCTs across different contexts. Similarly, as was done with the Career Academies program, it is possible to do an initial RCT across multiple sites at the same time, along with careful measurement of implementation and other factors.

Even without multi-site replication efforts, we should still favor RCTs whenever we have the chance to prospectively evaluate a program. For example, if you want to know whether school vouchers will work in Chicago, an observational study on private schooling in Chicago would matter more than the best RCT on school vouchers in India. But an RCT on vouchers in Chicago would be more useful still. And if an RCT on vouchers in India was all the evidence we had, it would be more reliable than an *observational* study from India.

Whether conducted in Chicago or India, an RCT will likely deliver better information than the alternative. Contrary to what some researchers have argued, there is no reason to think that observational research conducted on a large national database is more useful for extrapolation than an RCT in one city. This is because regression and similar methods end up weighting the observations in ways that may bear "little resemblance to the population of interest."

Another problem cited by critics is that the types of people or institutions who agree to sign up for an RCT are not necessarily representative of the broader population. For example, an important 2012 paper by Hunt Allcott and Sendhil Mullainathan looked at 14 experiments on energy conservation involving more than a half-million homes across the United States that were run by an electric company called Opower. Across these different sites, the treatment effect varied by a factor of two. The 14 sites that agreed to do RCTs differed from other sites served by Opower, and the researchers couldn't explain the different effects seen even by controlling for "seemingly good household-level demographics."

This is not an unanswerable problem, though. Allcott and Mullainathan suggest researchers should take pains to "clearly define the target site or population of interest," compare observable characteristics of the sample site versus the true population of interest, and conduct statistical tests of whether the treatment effect seems to be the same across the experimental sample. In any event, even if sites do not sign up for RCTs randomly, they argue that using observational evidence would be even worse and state, "Nonexperimental approaches . . . perform dramatically worse than experimental estimators in the same population."

Opponents also raise the issue of a possible Hawthorne effect in that people who know they are being observed in an experiment over perform in ways that would not occur in a more natural setting. This objection seems meritless. Apart from the fact that the original Hawthorne effect was "entirely fictional" (in the memorable words of Levitt and List), any widespread Hawthorne effect would flatly contradict Peter Rossi's oft-quoted "stainless steel law of evaluation," *i.e.*, "The better designed the impact assessment of a social program, the more likely is the resulting estimate of net impact to be zero." If Hawthorne effects were of any importance, then, instead of quoting Rossi's law, researchers would all solemnly observe that RCTs overwhelmingly find huge overestimates of the true treatment effects.

Yet another puported reason to be wary of extrapolating from RCTs is that the treatment may actually work better in an initial RCT than when scaled up or moved to other settings. This is because the original RCT is measuring the program as implemented by the original designers, who are both highly knowledgeable and highly motivated, whereas the program's effectiveness may be diluted when implemented by other less motivated and less skilled people elsewhere. Similarly, some argue that RCTs are fine for measuring small-scale programs, but when a program is scaled up to serve a larger population, the general equilibrium effects may be quite different. For example, charter schools may have a particular effect when they serve 3 percent of a city's students, but their effect might change if they were scaled up to serve 80 percent or 100 percent.

There are two answers to these related objections. First, as with many RCT myths, these arguments are not about RCTs at all. Instead, the real point is that we should be cautious when scaling up programs, regardless of how they might be evaluated.

Second, RCTs can be designed to measure *both* individual impacts and system-wide impacts. For example, a job training program might have one effect on its graduates, but its effect on the larger economy might be very different if the program's graduates displace other job-seekers. The researchers at the Jameel Poverty Action Lab at MIT did an experiment in France which used random assignment to select which communities and individuals would participate in the program. They were thus able to see whether individuals who completed the job training program were more likely to find jobs as well as whether communities that offered the job training programs experienced changes in overall employment.

In short, however difficult it might be to extrapolate from an RCT of a specific question in a specific context, it is not a reason to disfavor RCTs. It is at least as difficult to extrapolate from other forms of evaluation, and an RCT gives us better information about the original context.

D. Myth 4: RCTs Are a Black Box

According to some critics, RCTs are essentially a black box. All they can tell you is whether a treatment worked—not why. Yet, policymakers need to know the mechanisms by which policies affect human welfare. As Maria May put it in the "Stanford Social Innovation Review": "RCTs tell you only whether or not something works, and how well. Why it does or doesn't work, from the researcher's perspective, is up for interpretation, but for practitioners, it is critical to adopting a new practice."

As is often the case, this objection does not apply only to RCTs. For *any* quantitative form of research, the overall average effect will be a black box that doesn't necessarily tell us anything about the mechanisms that achieve that effect.

On the other hand, RCTs can be designed not only to measure the overall mean effect but also to test which mechanism is responsible. A factorial design, for example, creates several different treatment arms that each receive a different version of the treatment. By looking at which ones are more or less effective, we can get an idea of which mechanism leads to the effect. Indeed, RCTs are superior to other methods of evaluation because they allow researchers to look at distinct mechanisms.

In addition, RCTs can be designed to test mechanisms by themselves. As Ludwig, Kling, and Mullainathan write in their paper, "Mechanism Experiments and Policy Evaluations," the theory behind so-called "broken windows policing" in dangerous neighborhoods is that disarray is a marker that shows the neighborhood is ripe for further criminal action. The logic, in turn, is that broken windows policing leads to fewer broken windows and the like, which results in less criminal activity.

A traditional RCT would try different models of policing in different neighborhoods. But, Ludwig, Kling, and Mullainathan write, a less expensive "mechanism" RCT would go like this: "Buy a small fleet of used cars. Break the windows of half of them. Park the cars in a randomly selected subset of neighborhoods, and then measure whether more serious crimes increase in response." In other words, rather than run an RCT on the policy, run an RCT to "directly test the causal mechanism that underlies the broken windows policy." If the mechanism itself doesn't work, there is no need for a more expensive test of the policy.

Finally, as Ludwig, Kling, and Mullainathan note, as valuable as it is to have information about causal mechanisms, history is full of medical interventions that were known to work before the mechanism was fully understood. In their words, "Evidence that an intervention works, even if we don't understand why, is better than not having that intervention in our portfolio of policy options at all."

E. Myth 5: RCTs Are Not Suited to Complex, Fast-Changing Programs

Peter York writes in the "Stanford Social Innovation Review" that RCTs "do not lend themselves to the real-time learning and fast program adaptations demanded by the complex and tumultuous environment in which nonprofits operate today." Instead, "To continually refine their programs, nonprofit leaders need to know much more, including which members of the group benefited, which did not, why, and the explicit cause-and-effect relationships." Similarly, a team of authors from the Center for Medicare and Medicaid Services recently wrote: "When testing discrete, 'conceptually neat' interventions (*e.g.*, testing various forms of a letter to communicate information to beneficiaries), CMS uses randomized designs. However, for interventions that are constantly evolving or multimodal, randomization might not be feasible or appropriate." Instead, they prefer to evaluate Medicare programs using matching methods.

These arguments are illogical. Suppose there are 30 hospitals, half of which implement a complex payment program and half of which don't. If you want to evaluate that program, you can randomize which 15 hospitals will have the payment program, or you can let 15 select themselves and then try to find 15 hospitals somewhere that can be "matched." But the pace at which the program might evolve has *absolutely nothing to do* with whether you use randomization or matching. Using matching simply means that you have an inferior comparison group from the start, not that the program evolves any slower or that the comparison will be any easier to make. If a program is evolving too quickly for an RCT to be meaningful, then it is evolving too quickly for *any* method of evaluation.

As for "complex" or "multimodal" interventions, the argument apparently runs like this: It is challenging, at best, to get a school to agree to a multi-pronged strategy that requires changing curriculum, staffing rules, work schedules, tracking policies, and length of the school day, along with implementing new technology. The same goes for a hospital that is expected to change a dozen different things about its operations in order to become an accountable care organization under the Affordable Care Act. Thus, in the rare case that an institution's leaders have spent internal political capital to agree to such complex interventions, they will not be willing to scrap all of their reform efforts and be randomized into the control group. Thus, an RCT will not be feasible with those institutions.

Perhaps that is true in some cases, but there also are real-world examples of complex, multimodal interventions that have been evaluated in RCTs, such as the "Success for All" program for reforming an entire school's operations, or the 15 RCTs done on a complex Medicare care coordination program, or the recent six-part poverty intervention studied in multiple countries. Even in the face of political resistance, policymakers and funders can still press for other forms of randomization that may be more politically palatable. For example, researchers could use what is called a stepped wedge or randomized rollout design, in which some hospitals or schools implement the intervention this year, while others implement it at a later date. (With several treatment groups, each group could enter in a successive year.) With this kind of design, every school or hospital gets the intervention eventually, and the ones that implement in later years serve as a control group for the ones that implemented the intervention earlier.

F. Myth 6: RCTs Can Still Be Biased

William Easterly says RCTs can be "manipulated to get the 'right' results," in that "one could search among many outcome variables and many slices of the sample for results" and also search for settings that were "more likely to give good results." Likewise, Angus Deaton points to "practical problems that undermine any claims to statistical or epistemic superiority," such as attrition or missing data.

True enough. The mere fact that an evaluation has the label "RCT" does not mean its findings are perfect. There are plenty of opportunities for researchers to affect the outcomes—consciously or unconsciously—by exercising too much flexibility in how they treat the data. There are also plenty of things that can go wrong with an RCT.

But these criticisms don't go very far. As for manipulation, there are two responses. First, other types of studies are *far more susceptible* to such manipulation. An RCT can be evaluated merely by comparing the average outcome across two groups of people. But observational studies with complicated econometric models leave much more room for manipulation.

Second, we can prevent most of the opportunities to manipulate RCT analysis by having researchers establish pre-analysis plans and pre-register them publicly with well-known sites such as ClinicalTrials.gov, SocialScienceRegistry.org, or the Open Science Framework. Indeed, it would suffice in most cases to have a one-page statement listing the primary and secondary outcomes of the RCT, the sample size calculation, sources of data, rules for excluding outliers, and the intended statistical model.

Pre-registration became common in medicine when medical journals started requiring it over a decade ago and even more so after a 2007 federal law mandated registration of certain clinical trials done in support of a new FDA drug approval. Journals in social science could adopt similar mandates or could even adopt the new format known as "registered reports," in which peer review and the provisional acceptance of an article occurs based on the statistical analysis plan itself, before the experiment or analysis even occurs. What makes this a good idea? For one, the peer reviewers and editors cannot be biased by whether the results were positive, null, or negative, but instead make their decision based on the quality of the experimental design and data collection methods. The researchers are, in turn, less motivated to skew the findings in a way that makes them more publishable and are more likely to report the findings however they occur.

As for issues such as attrition, missing data, crossover in treatment status, and the like, an RCT is still likely to provide better information than the alternative. In the words of Guido Imbens: "It is true that violations of assignment protocols, missing data, and other practical problems can create complications in the analyses of data from randomized experiments. There is no evidence, however, that giving up control of the assignment mechanism and conducting an observational study improves these matters. Moreover, the suggestion that any complication, such as a violation of the assignment protocol, leads to analyses that lose all credibility accorded to randomized experiments is wrong." For example, researchers can and usually should use intent-to-treat analysis, in which all experimental subjects are assumed to be part of their group (treatment or control) as initially assigned. Then, even if there is selection in who drops out or crosses over to the other group, intent-to-treat analysis will still find the overall effect of having the program.

G. Myth 7: RCTs Are Too Limited

William Easterly claims that RCTs are "limited to small questions" but are infeasible for important questions about topics like "the economy-wide effects of good institutions or good macroeconomic policies."

It is true that RCTs are often infeasible when it comes to state- or nation-wide policies, such as the federal minimum wage or the availability of a mortgage income-tax deduction. But no one has ever said RCTs are the perfect study design that must be used for all questions, however inapt, or that anything other than an RCT is to be utterly disregarded. It is entirely consistent for proponents to say that RCTs could be used an order of magnitude more often, even though other methods must still be used for questions where RCTs are infeasible.

Moreover, as discussed above, RCTs can be designed to look at the specific mechanisms by which a broader state- or nation-wide policy is thought to operate. Such mechanism RCTs could be much more informative than other types of evaluation.

IV. Conclusion

RCTs are an invaluable means of determining the effects of social programs and policies. The arguments against RCTs are usually nihilist, as they would discredit any other form of evaluation even more than RCTs. Moreover, the feasibility arguments against RCTs are often overblown. Policymakers and funders should establish an expectation that at least 2 percent of any program budget should go toward funding not only new and creative RCTs on many questions about program effectiveness, but also toward funding RCTs about program components, geography and context, what sorts of people benefit the most, and more. RCTs can help to ensure that government policies actually help the people who need it the most