SOLVING
PUBLIC PROBLEMS

A Practical Guide to
Fix Our Government and
Change Our World

BETH SIMONE NOVECK

Yale
UNIVERSITY
PRESS

New Haven and London

how to get cost-effective programs, services and drugs to those who need them." It accomplishes this by providing "a synthesis of the best available research evidence on a given topic that has been prepared in a systematic and transparent way."[73]

While HSE primarily provides systematic reviews and evidence briefs for policy makers, it also contains a repository of economic evaluations and descriptions of health systems and reforms to them. Searches can be narrowed with a variety of filters based on the issue, the location of interest, or the type of document, among other variables. A "guided search" feature offers various tips on navigating the repository.

HSE's sister project, Social Systems Evidence, expands access to evidence reviews in twenty areas of government policy, including climate action, social services, economic development, education, housing, and transportation, in an effort to accelerate finding solutions to the problems outlined in the Sustainable Development Goals. Monash University, home of BehaviourWorks Australia, has partnered with McMaster to curate the SSE database. Monash University is the Australia, New Zealand, and Pacific Regional Centre of the UN Sustainable Development Solutions Network and therefore positioned to make this resource available to the broader sustainable development community. SSE identifies synthesized research by conducting weekly electronic searches of online bibliographic databases and manual searches of relevant websites.[74] As a result of such initiatives, the Minister of Health in Ontario, Canada, for example, now requires any policy memo proposing a new intervention to include a search of one of these two databases in order to demonstrate that the proposal is grounded in evidence. McMaster, like BehaviourWorks, also provides decision-makers with help finding evidence and creating rapid learning and improvement in health organizations, in particular. The university will prepare rapid evidence syntheses in three, ten, or thirty days and assist with conducting evaluations and training leaders in evidence review.[75]

RANDOMISTAS VERSUS CONTESTISTAS: THE LIMITS OF RCTS

In the hierarchy that has emerged in evidence-based policy making, the RCT is the gold standard and top of the evidentiary pyramid. Social scientists and social science methods experts who either run experiments or conduct systematic reviews—studies that gather known research on a particular topic and summarize both the findings and quality of the research—

as noted earlier,

tend to be fervent proponents for the value of RCTs. But that evidentiary hierarchy—what some people call the "RCT industrial complex"—may actually be dangerous and blind us to research challenges from which RCTs are not exempt, cause us to overlook important other sources of evidence, or lead us to discount workable solutions just because there is no accompanying RCT.

However, exclusive focus on RCTs can lead to other important sources of solutions being overlooked. A trawl of the solution space shows that successful interventions developed by entrepreneurs in business, philanthropy, civil society, social enterprise, or business schools who promote and study open innovation, often by developing and designing competitions to source ideas, often come from more varied places. Uncovering these exciting social innovations lays bare the limitations of confining a definition of what works only to RCTs.

Unfortunately, advocates of evidence evaluation, especially RCTs—what I nickname the *randomistas*—diverge from the *contestistas,* who focus their enthusiasm and energy on the use of incentive prizes to spur social and policy innovations. Both groups are growing but growing apart.

There are significant limitations to RCTs. For a start, systematic evidence reviews are quite slow. Not only do experiments take time to conduct and often over a year or more to publish in a peer-reviewed journal, but the curation and collection of a systematic review adds another year to the time frame—assuming that a standard for how to evaluate the evidence has already been agreed on. A Cochrane review takes upward of two years, and despite published standards for review, there is a lack of transparency.[76] Faster approaches are important.[77]

In addition, many solutions that have been tested with an RCT clearly do not work. Interestingly, the first RCT in an area tends to produce an inflated effect size. Many more entrepreneurial and innovative solutions are simply not tested with an RCT and are not the subject of academic study.

For example, MIT Solve describes itself as a marketplace for socially impactful innovation designed to identify lasting solutions from tech entrepreneurs for the world's most pressing problems. It catalogs hundreds of innovations in use around the world, like Faircap, a chemical-free water filter used in Syria and Mozambique, or WheeLog!, an application that enables individuals and local governments to share accessibility information in Tokyo. Every year, Solve issues four "Global Challenges" to attract teams

developing socially impactful projects. It then provides the winning teams with money and mentoring. There is little formal research about the fourteen hundred grassroots community innovation projects from which Solve selected thirty-two teams, yet these solutions are working in the field.[78] Solve is just one example of a prize-backed approach to find innovation.[79] You will recall from chapter 7 that Challenge.gov has over a thousand posted challenges with an enormous variety of solutions, not all of which are the subject of an academic article. As such challenges proliferate, and more programs in social innovation spring up, ever more promising solutions ~~are found.~~ emerge.

Research funding is also too limited (and too slow) for RCTs to assess every innovation in every domain. Many effective innovators do not have the time, resources, or know-how to partner with academic researchers to conduct a study, or they evaluate projects by some other means. Social innovation competitions, prize-backed challenges, and government contracts generally do not call for or provide funding to support associated academic research. Also, academic researchers cannot conduct experiments for which there is no grant funding, as they need such funding to hire the PhDs and postdoctoral students to conduct the work with them. Academic faculty are expected to provide such professional opportunities for students.

Moreover, limiting evidence to RCTs may also perpetuate systemic bias because of the underrepresentation of minority researchers and viewpoints in traditional academe and philanthropy. Limiting evidence, too, to that which is developed by academics, instead of by communities (see the discussion in chapter 10 about social auditing as an alternative), is biased. The Nobel Prize–winning economist Angus Deaton points out that RCTs create the risk of discrimination:

> How is informed consent handled when people do not even
> know they are part of an experiment? . . . 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 or their representatives)
> who had accepted, sometimes reluctantly, an obligation to

prevent the worst of poverty, and wanted to minimize the cost of doing so.[80]

Pushing decision-makers to rely only on RCTs could cause them to overlook important solutions developed by communities for communities. This is why organizations like BehaviourWorks in Australia and Project Evident and The GovLab in the United States help public partners find more diverse, nonacademic sources of evidence as well. Looking to nonacademic sources may debias one's sources of evidence.

And whereas RCTs answer the question "Did it work?" (and sometimes not even that), they do not explain how it worked or how satisfied people are. These are better measured using qualitative techniques such as structured or unstructured interviews. Deaton and the philosopher Nancy Cartwright write, "RCTs would be more useful if there were more realistic expectations of them and if their pitfalls were better recognized. . . . RCTs can play a role in building scientific knowledge and useful predictions but they can only do so as part of a cumulative program, combining with other methods, including conceptual and theoretical development, to discover not 'what works,' but 'why things work.'"[81] But qualitative approaches are often deemed inferior in the evaluative hierarchy.[82]

An intervention might even "fail" an RCT but still be promising. In the 1990s, the US federal government did an experiment in housing policy by giving some families a rental subsidy if they moved from a higher- to a lower-poverty neighborhood.[83] The results of the initial RCT found that "moving out of a disadvantaged, dangerous neighborhood into more affluent and safer areas does not have detectable impacts on economic outcomes four to seven years out."[84] The Moving to Opportunity project would have seemed to be a failure.

But the absence of economic results in the short to medium term did not mean they would not exist ten or twenty years later or that people would not realize other benefits such as reduced violence, lower stress, and improved physical and mental health. The evaluation measured only short-term changes in income. In fact, subsequent research a decade later reveals that these relocation programs had profound and positive economic benefits for the children of the families that moved.[85] A few extra years of data made all the difference. Yet the initial failed RCT might have wrongly reduced willingness to invest in social mobility programs like these.

Furthermore, RCTs test only very small and incremental experiments like whether to give away a kilogram of lentils. Sometimes we need to be radical and try bigger things than we can measure with an RCT. For example, of the thirteen thousand RCTs run by Google and Microsoft on their platforms, 80 to 90 percent have found no significant effects, largely because they are testing trivial interventions. Google tested forty-one shades of blue to determine which hyperlink color to use. Most of those tests were probably inconclusive.[86] RCTs intentionally do not look at broader context. They are designed to be narrow, to say "this" intervention is better than "that" and incrementally build a knowledge base through successive experiments. But the more significant research results may also not correspond to a more practical, palatable, and implementable solution for the real world.

Not all "solutions" lend themselves to testing with an RCT. Traditional RCTs have not previously lent themselves well to studying how institutions innovate. While many innovative solutions relate to the design of organizations, research into how such design influences performance has been slowed by the challenges of running RCTs on such complex questions. Professors Kevin Boudreau and Karim Lakhani, of London and Harvard Business School, respectively, have spent the past decade trying to develop experimental approaches to empirical measurement of how institutions make decisions and solve problems, areas where there is a dearth of empirical literature. Sometimes institutions do not want to randomize participants, especially if the goal is to measure, for example, what kind of innovators participate in open innovation competitions and what they do. As we saw in chapter 7, we need participants to self-select whether to join a prize-backed crowdsourcing challenge. Randomization interferes with the design of that incentive. Second, as Boudreau and Lakhani point out, it is rare that the most important dimensions of an institution can be manipulated in isolation.[87] Given the variety and diversity of new tools in the innovator's toolkit, including creative forms of institutional design, we do not want to rule out such solutions even if they are hard to test with a good RCT.

While successful RCTs, especially as evaluated through a third-party systematic evidence review, are one way to evaluate the quality of evidence, they are not the only way. Evidence can be evaluated using other criteria that place the value on different measures. Take the example of York, England. York measures the effectiveness of its programs using a method called Social Impact Return on Investments, through which it analyzes the

anticipated return on investment for every dollar spent with regard to its stated social values and goals. In 2016, York's People Helping People partnership came together to reimagine the approach to volunteering and to explore a people-led approach, embedding social action across the city. Among its initiatives, York enlists neighbors to visit and help the elderly to combat loneliness and social isolation. GoodGym is one such example of thinking locally, acting personally: neighbors form running clubs to get exercise while they run to visit an isolated senior and run back. In the city's people-helping-people approach, it places value on—and hence measures success by—mobilizing the talent and energy of citizens and civil-society organizations and creating a more collaborative and compassionate public service, not only on decreasing loneliness.[88]

Finally, modern machine learning may offer approaches superior to an RCT. It might make it possible to assess more complex context and variables than can be assessed in an RCT. The Dutch data scientist Arjan Haring writes, "We can utilize novel breakthroughs in our understanding of causality; as it turns out, it is strictly not necessary to resort to uniform random allocation as is done in the clinical trial. Rather, as long as we can compute and store the probability of receiving a treatment conditional on the patient characteristics, we can use the collected data to estimate causal effects."[89] The methods by which machine learning can complement or substitute for RCTs are still in their infancy (and beyond the scope of this book or the competence of its author to explain). What is important to note is that big data is leading to new ways of measuring causality and running experiments to measure, and even predict, the success of interventions.

In summary, the existence of so much bottom-up innovation means that we cannot be limited to choosing only solutions that have been evaluated using an RCT. This approach is only one way to spot solutions, and it comes with inherent biases. While some philanthropists demand RCTs, others are eager to reward and celebrate community activism and creativity may see academic research as cumbersome and out of keeping with their desire to reward the agile innovator. For example, the Unorthodox Prize rewards unorthodox ideas that are not attracting sufficient attention and resources elsewhere.[90] If an innovation has been the subject of much study and publication, it does not meet the prize organizers' definition of unorthodox and undiscovered. Whereas they want to know how the idea is achieving impact and can be scaled, they impose no evidentiary standard or

delete ?

requirement. Nonetheless, such competitions attract potentially effective solutions with their promise of millions of dollars in prize money.

BEYOND THE RCT: FINDING EVIDENCE IN EXPERIENCE

Despite the strengths of RCTs, they are not enough to ensure that policy is informed by evidence and practice, as will be discussed in "Evaluating Solutions" later in this chapter. Solutions that are not backed up by an RCT or even academic research must also be sought and found. Public entrepreneurs need a strategy for complementing the kinds of reports available from HSE, Campbell, or Cochrane with their own research, using both documents and people, all without drowning in a sea of information. Again, we need to take advantage of collective intelligence to crowdsource knowledge of old ideas: What else is out there? What else has been tried?

Let us explore some of the methods and tools for conducting a rapid field scan that goes beyond systematic reviews of RCTs to include experience-based learning from both documents and people.

Cataloging Generic Solutions

While public problem solvers are seeking creative ideas to solve a public problem, their approaches, however unique, are also likely to fit a familiar paradigm. Reviewing generic responses to typical policy problems can help to improve the search for the solution by zeroing in on those that seem most promising. For each generic solution, it can be helpful to ask, Might this be a useful approach for me? Is there a version of this approach that will respond to my problem?[91]

Generic mechanisms for solving problems include those typical policy responses such as legislation, budgets, grants, contracts, loans, and economic and social regulation.[92] Lester Salamon's *Tools of Government* and Eugene Bardach's *Practical Guide for Policy Analysis* both list these usual instruments of governance. For each, there is a range of common options for what to do, including adding something new, getting rid of something old, changing how people do things, and so on. These generic lists capture the standard solutions—the acceptable Overton Window of options—in response to frequent problems of service delivery. Here are a few examples of typical plays in the policy playbook that may be useful for public entrepreneurs or public problem solvers to consider.

Governments levy taxes to raise revenue or to create a disincentive for