A MOUNTAIN OF PEBBLES

EFFECTIVELY USING RCTS IN THE PUBLIC AND NONPROFIT SECTORS

JAMES MANZI

Randomized clinical trials have gained enormous currency as the most reliable way to measure the impact of social interventions, but their application has not reflected the dual issues of high failure rates and difficulty of generalization. This essay offers a short history of RCTs and suggests ways to make them more effective in predicting success.

Attempts to evaluate the effectiveness of interventions by applying the treatment to one group of patients ("test") and not to another group ("control") appear throughout recorded history. We see them in medicine from the biblical book of Daniel to Islamic and Chinese scholars in the eleventh century to James Lind's determination that citrus fruits prevent scurvy in 1747. An enormous challenge in this technique has always been how to hold all other factors constant between the test and control groups so we can know the difference in treatment must be the cause of the difference in outcomes.

The solution to this problem is to randomly assign participants to the test versus control group. The first randomized clinical trial that achieved modern standards of rigor was likely a 1938 U.S. Public Health Service trial of pertussis vaccine in Norfolk, Virginia.[1] One randomly chosen subset of Norfolk's population was selected for vaccination and another was selected to not receive the vaccine. The researchers could, thereby, conclude that any

subsequent differences in pertussis rates between the two groups were caused by the vaccine. This is precisely the method used in late 2020 and early 2021 to evaluate the safety and efficacy of COVID vaccines.

Social science researchers quickly observed that this approach could be applied to evaluate the effectiveness of programmatic social interventions, and the RCT has, appropriately, become the gold standard of evidence for the causal effects of social programs across fields including criminology, education, and social welfare.

Evaluation of several decades of these RCTs executed in the developed world leads to two important observations. First, the vast majority of tested social interventions do not produce measurable improvement in targeted outcomes. The second is the problem of generalization; programs that demonstrate gains in experiments often create these benefits only in specific contexts, such as types of recipients, environmental situations, or provider capabilities.

Consider the first observation—that most social interventions fail when tested. Criminologists at the University of Cambridge have done the yeoman's work of cataloging all known criminology RCTs between 1957 and 2004 with at least one hundred test subjects.[2] Twelve of the programs were tested in "multisite" RCTs: experiments in several cities, prisons, or court systems. Eleven of the twelve failed to produce positive results, and the small gains produced by the one successful program (which cost an immense $16,000 per participant) faded away within a few years. This is a 92 percent failure rate. The U.S. Department of Education's Institute of Education Sciences (IES) sponsored a series of RCTs that tested fourteen well-known preschool curricula and found only one curriculum that demonstrated some causal gains in performance that persisted only through kindergarten.[3] This is a 94 percent failure rate. And none of that considers whether either of the two successful programs is remotely cost-effective. We see this same pattern time and again for social interventions.

This high failure rate is not unique to social programs. A National Institutes of Health (NIH) evaluation of 798 drug development programs found that only approximately 6 percent of pre-clinical therapies complete a Phase III RCT successfully and are approved for use.[4] Google has reported that only about 10 percent of on-line changes tested in RCTs create business improvement.[5]

But unlike most medical interventions, even when we find a social intervention that proves impact in an RCT, the problem of generalization rears its head.

We can run a clinical trial in Norfolk, Virginia, and conclude with tolerable reliability that "*Vaccine X prevents disease Y.*" We cannot conclude that if literacy program X works in Norfolk then it will work everywhere. The real predictive rule is usually closer to something like "*Literacy program X is effective for children in urban areas, and who have the following range of incomes and prior test scores, when the following alternatives are not available in the school district, and the teachers have the following qualifications, and overall economic conditions in the district are within the following range.*"

In 1981–1982, Lawrence Sherman, a respected criminology professor at the University of Cambridge, led an extremely influential experiment that randomly assigned one of three responses to Minneapolis cops responding to misdemeanor domestic-violence incidents: they were required to either arrest the assailant, provide advice to both parties, or send the assailant away for eight hours.[6] The experiment showed a statistically significant lower rate of repeat calls for domestic violence for the mandatory-arrest group. The media and many politicians seized upon what seemed like a triumph for scientific knowledge, and mandatory arrest for domestic violence rapidly became a widespread practice in many large jurisdictions in the United States. But sophisticated experimentalists understood that, because of the issue's complexity, there would be hidden conditionals to the simple rule "mandatory-arrest policies reduce domestic violence." The only way to unearth these conditionals was to replicate the original experiment under a variety of conditions. Sherman's own analysis of the Minneapolis study called for such replications. So, researchers replicated the RCT six times in cities across the country. In three of those studies, the test groups exposed to the mandatory-arrest policy again experienced a lower rate of re-arrest than the control groups did. But in the other three, the test groups had a higher re-arrest rate.

The danger of drawing conclusions based on a single RCT on a social policy topic is obvious in this example. Suppose Sherman had happened to run the original experiment in Memphis (one of the cities where the replication failed). Would we then have been justified in concluding that mandatory arrest doesn't work? Based on this set of replications, whether it works in any given city is roughly equivalent to a coin flip. It is important to keep this in mind when presented with the gold-standard evidence of any one well-designed RCT. The obvious question is whether anything about the situations in which mandatory arrest worked distinguishes them from

situations where it did not. If we knew this, we could apply the program only where it is effective.

In 1992, Sherman surveyed the replications and concluded that in stable communities with high rates of employment, arrest shamed the perpetrators, who then became less likely to reoffend, while in less stable communities with low rates of employment, arrest tended to anger the perpetrators, who would, therefore, be likely to become more violent.[7] The problem with this kind of conclusion, though, is that because it is not itself the outcome of an experiment, it is subject to the same uncertainty as any other pattern-finding exercise. How do we know whether it is right? We do so by running an experiment to test it—that is, by conducting still more RCTs in both kinds of communities and seeing whether they bear out this conclusion.

Confronting this difficult reality directly can help us be much more effective in evaluating interventions. RCTs have gained enormous currency as the most reliable way to measure the impact of social interventions, but their application has not reflected the dual issues of high failure rates and difficulty of generalization. I believe public agencies and nonprofit organizations attempting to use RCTs should embrace three simple principles:

1. *Kiss a lot of frogs to find a prince.* Based on experience, we should expect to try at least ten very promising intervention ideas before we find one that actually will improve any targeted outcome. This means building the capacity to run many tests at low cost per test. This, in turn, requires using administrative data, semi-automated test design and analysis, and organization and procedures that lower the hard dollar and organization friction costs of running a test.
2. *Build a mountain of pebbles.* There are no silver bullets for social problems out there waiting to be found through RCTs. Agencies and nonprofits should be looking for lots of small wins through testing, not transformational moonshots. Foundations that fund nonprofits would be better off requesting *"Show me the number of tactical RCTs you have done and the results,"* than *"Show me the impact of your overall program according to an RCT."*

(*continued*)

3. *Bottom-up not top-down.* Use of RCTs in social program evaluation often proceeds from observations of their successful use in medicine, but this analogy is far from perfect because the problem of generalization is so much more severe for social interventions. Rather than an image of experts who develop theory-dependent program ideas that are then rigorously tested to find "what works," we should, instead, think of a continuing flow of localized, tactical ideas that emerge from practitioners who then have the capacity (expertise and resource) embedded in their organization to rapidly test these potential innovations and implement the small fraction that create improvement.

NOTES

1. Iain Chalmers, "Joseph Asbury Bell and the Birth of Randomized Trials," *Journal of the Royal Society of Medicine* 100, No. 6 (June 2007): 287–93, https://journals.sagepub.com/doi/pdf/10.1177/014107680710000616.

2. David Farrington and Brandon Welsh, "Randomized Experiments in Criminology: What Have We Learned in the Last Two Decades?," *Journal of Experimental Criminology* 1, (April 2005): 9–38, https://doi.org/10.1007/s11292-004-6460-0.

3. Preschool Curriculum Evaluation Research Consortium, "Effects of Preschool Curriculum Programs on School Readiness (NCER 2008–2009)," National Center for Education Research, Institute of Education Sciences, U.S. Department of Education, http://ies.ed.gov/ncer/pubs/20082009/pdf/20082009_rev.pdf.

4. Tohru Takebe, Ryoka Imai, and Shunsuke Ono, "The Current Status of Drug Discovery and Development as Originated in United States Academia: The Influence of Industrial and Academic Collaboration on Drug Discovery and Development," *Clinical and Translational Science* 11, no. 6 (July 2018): 597-606, doi: 10.1111/cts.12577. Epub 2018 Jul 30. PMID: 29940695; PMCID: PMC6226120.

5. Stefan Thomke, "Building a Culture of Experimentation," *Harvard Business Review*, March–April 2020, https://hbr.org/2020/03/building-a-culture-of-experimentation.

6. Lawrence Sherman and Ellen Cohn, "The Impact of Research on Legal Policy: The Minneapolis Domestic Violence Experiment," *Law & Society Review* 23, no. 1 (1989): 117–44, https://doi.org/10.2307/3053883.

7. Lawrence W. Sherman, Janell D. Schmidt, Dennis P. Rogan, Douglas A. Smith, "The Variable Effects of Arrest on Criminal Careers: The Milwaukee Domestic Violence Experiment," *Journal of Criminal Law & Criminology* 83, no. 137 (1992–1993).