The limitations of randomised controlled trials

Angus Deaton and Nancy Cartwright

November, 2016

Randomized controlled trials have been sporadically used in economic research since the negative income tax experiments between 1968 and 1980, see Wise and Hausman (1985), and have been regularly used since then to evaluate labor market and welfare programs, Manski and Garfinkel (1992), Gueron and Rolston (2013). In recent years, they have spread widely in economics (and in other social sciences), perhaps most prominently in development and health economics. The "credibility revolution" in econometrics, Angrist and Pischke (2010), putatively frees empirical investigation from implausible and arbitrary theoretical and statistical assumptions and RCTs are seen as the most "credible" and "rigorous" of the credible methods; indeed credible non-RCT designs typically pattern themselves as closely as possible on RCTs. Imbens (2010) writes "Randomized experiments do occupy a special place in the hierarchy of evidence, namely at the very top."

In medicine, Pocock and Elbourne (2000) argue that only RCTs "can provide a reliably unbiased estimate of treatment effects," and without such estimates, they "see considerable dangers to clinical research and even to the well-being of patients." The link between bias and risk to patients is taken as obvious, with no attempt to show that an RCT experimental design does indeed minimizes the expected harm to patients. The World Bank has run many development related RCTs, and makes claims well beyond unbiasedness. Its implementation manual, Gertler et al (2011) states "we can be very confident that our estimated average impact" (given as the difference in means between the treatment and control group)

"constitute the true impact of the program, since by construction we have eliminated all observed and unobserved factors that might otherwise plausibly explain the differences in outcomes." High quality evidence indeed; the truth is surely the ultimate in credibility.

In Deaton and Cartwright (2016), we argue that some of the popularity of RCTs, among the public as well as some practitioners, rests on misunderstandings about what they are capable of accomplishing. Well-conducted RCTs could provide unbiased estimates of the average treatment effect (ATE) in the study population, provided no relevant differences between treatment and control are introduced post randomisation, which blinding of subjects, investigators, data collectors, and analysts serves to diminish. Unbiasedness says that, if we were to repeat the trial many times, we would be right on average. Yet we are almost never in such a situation, and with only one trial, as is virtually always the case, unbiasedness does nothing to prevent our single estimate from being very far away from the truth. If, as if often believed, randomization were to guarantee that the treatment and control groups are identical except for the treatment, then indeed, we would have a precise—indeed exact—estimate of the ATE. But randomization does nothing of the kind, even at baseline; in any given RCT, nothing ensures that other causal factors are balanced across the groups at the point of randomization. Investigators often test for balance on observable covariates, but unless the randomization device is faulty, or people systematically break their assignment, the null hypothesis underlying the test is true by construction, so that the test is not informative, and should not be carried out.

Of course, we know that the ATE from an RCT is only an estimate, not the infallible truth, and like other estimates, it has a standard error. If appropriately computed, the standard

error of the estimated ATE can give an indication of the importance of other factors. As was understood by Fisher from the very first agricultural trials, randomization, while doing nothing to guarantee balance on omitted factors, gives us a method for assessing their importance. Yet even here there are pitfalls. The *t*–statistics for estimated ATEs from RCTs do not in general follow the *t*–distribution. As recently documented by Young (2016), a large fraction of published studies have made spurious inferences because of this Fisher-Behrens problem, or because of the failure to deal appropriately with multiple-hypothesis testing. Although most of the published literature is problematic, these issues can be addressed by improvements in technique. Not so however, in cases where individual treatment effects are skewed—as in healthcare experiments, where a one or two individuals can account for a large share of spending (this was true in the Rand Health Experiment), or in microfinance, where a few subjects make money and most do not—where the *t*-distribution again breaks down. Once again, inferences are likely to be wrong, but here there is no clear fix. When there are outlying individual treatment effects, the estimate depends on whether the outliers are assigned to treatments or controls, causing massive reductions in the effective sample size. Trimming of outliers would fix the statistical problem, but only at the price of destroying the economic problem; for example, in healthcare, it is precisely the few outliers that make or break a program. In view of these difficulties, we suspect that a large fraction of the published results of RCTs in development and health economics are unreliable.

The "credibility" of RCTs comes from their ability to get answers without the use of potentially contentious prior information about structure, such as specifying other causal factors, or detailing the mechanisms through which they operate. A skeptical lay audience is

often unwilling to accept prior economic knowledge and even within the profession, there are differences about appropriate assumptions or controls. Yet, as is always the case, the only route to precision is through prior information, and controlling for factors that are likely to be important, just as in a (non-randomized) laboratory experiment in physics, biology, or even economics, scientists seek accurate measurement by controlling for known confounders. Cumulative science happens when new results are built on top of old ones—or undermine them—and RCTs, with their refusal to use prior science, make this very difficult. And any RCT can be challenged *ex post* by examining the differences between treatments and controls as actually allocated, and showing that arguably important factors were unevenly distributed; prior information is excluded by randomization, but reappears in the interpretation of the results.

A well-conducted RCT can yield a credible estimate of an ATE in one specific population, namely the "study population" from which the treatments and controls were selected. Sometimes this is enough; if we are doing a post hoc program evaluation, if we are testing a hypothesis that is supposed to be generally true, if we want to demonstrate that the treatment can work somewhere, or if the study population is a randomly drawn sample from the population of interest whose ATE we are trying to measure. Yet the study population is often *not* the population that we are interested in, especially if subjects must volunteer to be in the experiment and have their own reasons for participating or not. A famous early example comes from Ashenfelter (1981), who found that people who volunteer for a training program tend to have seen a recent drop in their wages; similarly, people who take a drug may be those who have failed other forms of therapy. Indeed, many of the differences in results between

experimental and non-experimental studies can be traced, not to differences in methodology, but to differences in the populations to which they apply.

More generally, demonstrating that a treatment works in one situation is exceedingly weak evidence that it will work in the same way elsewhere; this is the "transportation" problem: what does it take to allow us to use the results in new contexts, whether policy contexts or in the development of theory? It can only be addressed by using previous knowledge and understanding; i.e. by interpreting the RCT within some structure, the structure that, somewhat paradoxically, the RCT gets its credibility from refusing to use. If we want to go from an RCT to policy, we need to build a bridge from the RCT to the policy; no matter how rigorous or careful the RCT, if the bridge is built by a hand-waving simile that the policy context is somehow similar to the experimental context, the rigor in the trial does nothing to support a policy; in any chain of evidence, it is the weakest link that determines the overall strength of the claim, not the strongest. Using the results of an RCT cannot simply be a matter of simple extrapolation from the experiment to another context. Causal effects depend on the settings in which they are derived, and often depend on factors that might be constant within the experimental setting but different elsewhere. Even the direction of causality can depend on the context. We have a better chance of transporting results if we recognize the issue when designing the experiment—which itself requires the commitment to some kind of structure—and try to investigate the effects of the factors that are likely to vary elsewhere. Without a structure, without an understanding of *why* the effects work, we not only cannot transport, but we cannot begin to do welfare economics; just because an intervention works, and because the investigator *thinks* the intervention makes people better off, is no guarantee that it actually

does so. Without knowing why things happen and why people do things, we run the risk of worthless casual (fairy story) causal theorizing, and we have given up on one of the central tasks of economics.

Citations:

Angrist, Joshua and Jörn-Steffen Pischke, 2010, "The credibility revolution in empirical economics: how better design is taking the con out of econometrics," *Journal of Economic Perspectives*, 24(2), 3–30.

Ashenfelter, Orley, 1978, "Estimating the effect of training programs on earnings," Review of Economics and Statistics, 60(1), 47–57.

Deaton, Angus, and Nancy Cartwright, 2016, "Understanding and misunderstanding randomized controlled trials," http://www.princeton.edu/~deaton/downloads/Deaton_Cartwright_RCTs_with_ABSTRACT_August_25.pdf

Garfinkel, Irwin, and Charles F. Manski, 1992, *Evaluating welfare and training programs*, Cambridge, MA. Harvard.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch, Impact evaluation in practice, Washington, DC. The World Bank.

Gueron, Judith M., and Howard Rolston, 2013, *Fighting for reliable evidence*, New York. Russell Sage.

Imbens, Guido, 2010, "Better late than nothing," *Journal of Economic Literature*, 48(2), 399–423.

Pocock, Stuart J., and Diana R. Elbourne, 2000, "Randomized trials or observational tribulations?" *New England Journal of Medicine*, 342, 1907–9.

Wise, David A., and Jerry A. Hausman, 1985, *Social Experimentation*, Chicago, Ill. Chicago University Press for NBER.

Young, Alwyn, 2016, "Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results," London School of Economics, Working Paper, Feb.