Benefits and limitations of randomized controlled trials[*]

Andrew Gelman[†]

28 Dec 2017

I agree with Deaton and Cartwright that randomized trials are often overrated. There is a strange form of reasoning we often see in science, which is the idea that a chain of reasoning is as strong as its strongest link. The social science and medical research literature is full of papers in which a randomized experiment is performed, a statistically significant comparison is found, and then story time begins, and continues, and continues—as if the rigor from the randomized experiment somehow suffuses through the entire analysis.

Here are some reasons why the results of a randomized trial cannot be taken as representing a general discovery:

1. Measurement. A causal effect on a surrogate endpoint does not necessarily map to an effect on the outcome of interest. Direct measurements also can be problematic if there is bias that is correlated with treatment assignment, as can occur in medical experiments without blinding or psychology experiments in which there is information leakage.

2. Missing data. Even a small proportion of dropout or nonresponse can bias the estimate of treatment effects, if missingness is correlated with outcome and treatment assignment.

3. Extrapolation. The participants in a controlled trial are typically not representative of the larger population of interest. This causes no problem if the treatment effect is constant but can leads to bias to the extent that treatment effects are nonlinear and have interactions. A related concern is realism: extrapolating from often-artificial experimental conditions to the real-world settings where the treatment might be applied.

4. Researcher degrees of freedom. The many options arising in data coding and analysis make it easy for researchers to obtain "statistically significant" p-values even in the absence of any large and consistent underlying effects. The fact that treatment assignment has been randomized does not protect researchers from this "garden of forking paths" (Simmons, Nelson, and Simonsohn, 2011, Gelman and Loken, 2014).

5. Type M (magnitude) errors. Selection on statistical significance leads to overestimates of treatment effects, this bias can be huge, and it can lead to a cascade of errors in the literature when exaggerated estimates in the literature are used in the design of overly optimistic future experiments (Gelman, 2018).

Each of these threats to validity is well known, but they often seem to be forgotten, or to be treated as minor irritants to be handled with some reassuring words or a robustness study, rather than as fundamental limitations on what can be learned from a particular dataset.

One way to get a sense of the limitations of controlled trials is to consider the conditions under which they *can* yield meaningful, repeatable inferences. The measurement needs to be relevant to the question being asked; missing data must be appropriately modeled; any relevant variables that differ between the sample and population must be included as potential treatment interactions; and

[*] Discussion of "Understanding and misunderstanding randomized controlled trials," by Angus Deaton and Nancy Cartwright, for *Social Science & Medicine*.

[†] Department of Statistics and Department of Political Science, Columbia University, New York

the underlying effect should be large. It is difficult to expect these conditions to be satisfied without good substantive understanding. As Deaton and Cartwright put it, "when little prior knowledge is available, no method is likely to yield well-supported conclusions." Much of the literature in statistics, econometrics, and epidemiology on causal identification misses this point, by focusing on the procedures of scientific investigation—in particular, tools such as randomization and p-values which are intended to enforce rigor—without recognizing that rigor is empty without something to be rigorous *about*.

This is not to say that existing investigations in social science, policy, and medicine are atheoretical or lacking in scientific content. Rather, there is a disconnect between the design of treatments (based on theory and the qualitative integration of the literature) and their evaluation (which, as noted above, implicitly assumes a direct relation between the measurements and the goals of the study, and between the participants in the experiment and the larger population of interest, and between the experimental conditions and the real world). As Deaton and Cartwright note, the traditional focus in statistics and econometrics on average treatment effects (and the corresponding focus in epidemiology on single-number summaries such as hazard ratios) is misleading for two reasons: first, we are often interested in variation of treatment effects, not just averages; second, an average is only defined relative to some population of participants and scenarios, so that the average of interest cannot be inferred from experimental data alone without some consideration of interactions and some effort to reweight to match the target population.

Where does this all leave us? Randomized controlled trials have problems, but the problem is not with the randomization and the control—which do give us causal identification, albeit subject to sampling variation and relative to a particular local treatment effect. So really we're saying at all empirical trials have problems, a point which has arisen many times in discussions of experiments and causal reasoning in political science; see Teele (2014). I agree with Deaton and Cartwright that the best way forward is to integrate subject-matter information into design, data collection, and data analysis, going beyond the sort of purely data-based reasoning of the sort we see in statistics and econometrics textbooks that is tuned to problems with large, stable, easily measured effects.

Once we recognize the importance of diverse sources of data, statistics can be helpful in making decisions and quantifying uncertainty. For example, Deaton and Cartwright discuss a hypothetical example of "two schools, St Joseph's and St Mary's, both of which were included in an RCT of a classroom innovation. The innovation is successful on average, but should the schools adopt it? Should St Mary's be influenced by a previous attempt in St Joseph's that was judged a failure?" Indeed, "if St Mary's is like St Joseph's, with a similar mix of pupils, a similar curriculum, and similar academic standing, might not St Joseph's experience be more relevant to what might happen at St Mary's than is the positive average from the RCT?" Some progress here should be possible using partial pooling of information. For example, in Weber et al. (2018), we fit a Bayesian multilevel model to the parameters in a pharmacometric model, allowing information obtained from an experiment on one drug to inform the inference for a new drug under development. In the example of St Joseph's and St Mary's, information can be shared using individual and school-level predictors (thus taking advantage of the postulated similarity in the pupils, curriculum, and academic standing), modulated by a distribution of the effects varying by school. The information included in such a model represents the sort of substantive knowledge that is needed to go beyond the limitations of the data. Along with this, it may be necessary to moderate our claims as well as our expectations, accepting that a study can be useful even if does not supply definitive evidence.

References

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* **44**, 16–23.

Gelman, A., and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102**, 460–465.

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **22**, 1359–1366.

Teele, D. L., ed. (2014). *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences.* Yale University Press.

Weber, S., Gelman, A., Lee, D., Betancourt, M., Vehtari, A., and Racine-Poon, A. (2018). Bayesian aggregation of average data: An application in drug development. *Annals of Applied Statistics.*