The RCTs-Only Doctrine: Brakes on the Acquisition of Knowledge?

E. Jane Davidson
Davidson Consulting Ltd., Aotearoa/New Zealand

The debate about how to rigorously gauge the outcomes of policies, programs, and other initiatives continues to rage around the globe. One camp pushes the notion that randomized controlled trials (RCTs) are always the preferred way to go, using scoring schemes such as the "Maryland Scale," which yields a Scientific Methods Score that places RCTs at the top for rigor (Sherman et al., 1998). The other camp in the debate argues that methods should be matched to the situation.

Debates like this have always been part of academic life in many fields. But what is unique in this case is that the preferential allocation of funding for evaluations that use certain designs is being used in some instances to effectively dictate practice to practitioners. The best known example is the US federal government's use of these scoring systems that effectively make it impossible to get funding for evaluations of certain programs unless they conform to the so-called "Gold Standard" (randomized controlled trials) or at least the "silver" alternative (quasi-experimental designs).

Here in New Zealand we are also starting to see this kind of thinking creeping in in some quarters. In a recent meeting when I shared some mixed method evaluation tools I had been helping develop with one group in a particular government agency, I was stunned to get a comment along the lines of "This is just the kind of work we are trying to put a stop to [in this agency]. We need to get people to move from *opinion-based* to *evidence-based* decision making."[1] In this case the "opinions" were specialists' expert judgments of performance on key outcomes for quite small numbers of individuals, triangulated with behavioral observations from other sources. Other [quantitative] forms of evidence were considered, but in this case the options would have yielded inferior (i.e., less valid and useful) data and at a cost that was completely unfeasible.

I certainly don't want to be dismissive of the work that Sherman and colleagues (as well as others) have done to develop ways of evaluating and concisely rating the quality of evaluations. The literature is awash with studies of widely varying quality, and the sheer volume of studies in certain areas is so daunting that it drastically inhibits the utilization of findings. Somehow there needs to be a way to sort the wheat from the chaff and to draw conclusions about, as Sherman et al. (1998) put it, "what works, what doesn't, and what's promising." In other words, in principle

[1] Thankfully, the views expressed in this case were not those held at more senior levels of the organization.

I support the development and use of a concise, straight-to-the-point scoring system for evaluating the quality of evaluative evidence.

The problem I have with the Maryland Scale and similar scoring systems is that somewhere along the way the word "evidence" has been unilaterally redefined to mean "quantitative data" – specifically, quantitative data derived from an experimental or at least a quasi-experimental design. If it doesn't meet this standard, it's not worth considering as evidence, according to proponents of RCTs.

There has already been much debate on the validity of the arguments on either side, so, rather than revisit this, perhaps it is a good time to explore some of the possible downstream *implications*. Suppose this "RCTs as the Gold Standard" doctrine managed to take hold and permeate all professional evaluation. In other words, what if clients only funded evaluations that used randomized controlled designs? Although it is unlikely to happen across the board, it does seem likely to happen in certain domains. What would be the implications for what gets evaluated and how we, as a society, acquire knowledge about what works?

The most important implication of all is that nothing would be evaluated that didn't lend itself to a randomized controlled trial:

1. Nationwide policies couldn't be evaluated because good comparison groups won't exist (i.e., the change takes place simultaneously throughout the country; so there is no chance for random assignment and a comparison population will be hard to find).

   If nationwide policies and initiatives don't get evaluated, we may end up only evaluating the "small fry" and not finding out whether major investments were worthwhile or not. This would be particularly important in smaller countries where most major initiatives are implemented nationwide rather than by state, province, or region.

2. Messy, real-world policies and programs that change and evolve and are implemented differently in different locations couldn't be evaluated because they don't "hold still" long enough for the treatment to be consistent enough to allow a "clean" comparison. If messy, real-world programs and policies don't get evaluated, especially those that (quite sensibly) adapt to the local conditions and improve as they go, we will never learn what works in the real world. Which is, after all, where most of our evaluands find themselves.

3. Interventions targeted at small populations (e.g., rural minority children with special learning needs) couldn't be evaluated because the numbers are not large enough to provide the required statistical power for a randomized experimental or even a quasi-experimental design.

   If interventions for smaller populations are passed over for evaluation, once again the minority groups in our society will miss out. Improvement of services to reach the already underserved is impossible without getting creative with our ways of finding out how small-scale initiatives work.

4. Truly innovative policies, programs, and projects that break new ground couldn't be evaluated because—given the very nature of initiatives that push into new territory—many important outcomes can't be anticipated in advance to allow the development of

valid, reliable instruments to measure those outcomes at baseline and later on. (It is rather ironic that really experimental initiatives are the ones that don't lend themselves well to experiments.).

The alternative might be to only evaluate such initiatives based on the outcomes that can be anticipated in advance, but this will require missing large chunks of important data, thereby making the overall evaluative conclusions invalid.

If new and innovative initiatives don't get evaluated (or get evaluated on only the predictable subset of their important outcomes), we will only find out what works among minor variations on same-old, same-old initiatives. This will severely limit our ability to find better ways of tackling important opportunities and problems.

5. Very early formative evaluations, which often focus more heavily on process evaluation (design, content and implementation) and possibly a few very preliminary outcomes, couldn't be evaluated because the bulk of the evaluation design/approach wouldn't look anything like an RCT, so would fail to score high enough on the methodological rigor scale.

   If formative evaluations that look seriously at process are not done, then policies and programs stand little chance of being able to be tweaked and improved soon enough to maximize the likelihood of seeing worthwhile longer-term outcomes. The other implication is that all evaluations would be delayed until programs had had enough time to start producing their medium and long-term outcomes.

   Therefore, programs would be denied the benefit of early feedback that they could actually use in a timely manner to maximize the likelihood of producing those longer-term outcomes. And evaluation would (justifiably) be seen as less useful because its findings will arrive well after they would have been maximally useful.

One of the weaknesses of many evaluations is that they fail to give adequate consideration to the causation issue. The recent debate about RCTs has certainly done much to draw our attention to this, which is a good thing. For quantitative and mixed method evaluations, one option that can strengthen an evaluation is to make judicious use of experimental and quasi-experimental designs where appropriate and feasible. However, if the so-called "Gold Standard" becomes almost the sole source of evaluative knowledge in any particular field or subfield, the implications for knowledge creation in our society are extremely serious. Not only may the best methods not be chosen for a particular evaluation, but very large numbers of important policies, programs and other initiatives are simply not going to be evaluated at all. In fact, maybe they won't ever be funded in the first place because they won't pass the Gold Standard "evaluability test." In other words, if the nature of the initiative is not conducive to the use of RCTs, the initiative itself won't be funded.

It is important to note that there are *many more options* that can be used to establish and/or strengthen causal claims in qualitative as well as quantitative and mixed method evaluations (for some suggestions that can be used with qualitative and/or quantitative data, see Davidson, 2004, 2006; Scriven, 1974). This doesn't mean that just anything should be accepted as causal evidence. Nor does it mean that it's impossible to rate an evaluation's technical quality or to distinguish poorly designed from well designed studies. But the quality barometer should be about how

conclusively the evaluation has ruled out the most likely alternative explanations for any suspected outcome, not simply whether "Tool A" has been employed to do so. The idea of a system for concisely rating an evaluation's technical quality is a good one, and (I think) an important tool for the mammoth task of sifting through and summarizing the evidence about what works. What we need now is a rating system that is inclusive of the diverse methods that can be used for this important task.

References

Davidson, E. J. (2004). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Davidson, E. J. (2006, November). *Causal inference nuts and bolts.* Demonstration session at the American Evaluation Association conference, Portland, OR. Available at (click on Presentations) http://davidsonconsulting.co.nz/

Scriven, M. (1974). Maximizing the power of causal investigations: The modus operandi method. In W. J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 68-84). Berkeley, CA: McCutcheon Publishing.

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S. D. (1998). *Preventing Crime: What works, what doesn't, what's promising*. National Institute of Justice, Washington, D. C. Available at http://www.ncjrs.gov/works/download.htm