The Logic of Causal Investigations and the Rhetoric and Pragmatics of Research Planning

22

Melvin M. Mark

Abstract

Must investigations of treatment effects adopt a logic of causal investigation that defaults to the design and application of randomized control trials in education? Does the language of scientific research impel this too for intervention studies? In this brief response to Scriven's chapter, we explore the pragmatics of research planning to identify compounding factors and challenges to such characterizations of 'gold standard' research in education. The prospect is tantalizing: if adjudications of explanatory power are the true benchmark for evaluating educational studies, then a requisite variety in research designs and logics must be entertained, and engaged.

Keywords

RCTs • Evaluation • Quality criteria • Validity • Generalizability

It is both an honour and somewhat foreboding to comment on Michael Scriven's Chap. 21, "The logic of causal investigation." Scriven has made thoughtful, indeed, seminal contributions in more than one field related to his chapter. Nevertheless, one can try to read his chapter with the kind of critical eye with which he scans the landscape of the logic of causal investigations. In my view, his focus is as much or more on key aspects of the contemporary landscape of rhetoric and funding priorities related to research design.

In part, Scriven's chapter is a critique of the language used by some advocates of randomized controlled (or clinical) trials (RCTs). By way of background, the terminology and claims to which Scriven takes exception arose primarily from supporters of a priority for RCTs in the funding of select research streams by the Department of Education's Institute of Education Sciences. In my view, one can appreciate the potential contribution of RCTs and still believe that excessive and inaccurate language was used by some advocates of the IES priority. This includes language that seemed to equate "scientific research" with randomized trials, as well as other language which seemed to suggest that confident causal inference requires a random assignment study. Viewed from this vantage, Scriven's presentation, including his references to children's learning, plate tectonics, and a hypothetical nonequivalent groups design, is compelling. Language that equates causal inference or scientific research with RCTs should be avoided. Case won.

But how important is winning that case, for the broader understanding of causal inference and for planning of cause-probing research in education and elsewhere? Of limited value, I would argue. Likewise, such a rhetorical victory is not likely to persuade advocates about the relative value of RCTs. Regardless of their rhetoric, those advocates appear concerned about the use of RCTs in circumstances rather different than when a baby's learns of her ability to cause a rattle to make noise. Instead, they appear to be interested in causal claims about treatment effects such as those relevant to, say, making choices about adopting one or another math curriculum, or retaining 'DARE' versus replacing it with a different substance abuse prevention programme, or mandating a certain level of credentialization for day care staff.

Fortunately, Scriven also addresses what he calls "the limited RCT position." Scriven recognizes the narrow victory that comes with defeating excessive rhetoric: "To avoid

M.M. Mark (✉)
Department of Psychology, Penn State University, University Park, PA, USA
e-mail: m5m@psu.edu

A.D. Reid et al. (eds.), *A Companion to Research in Education*,
DOI 10.1007/978-94-007-6809-3_22,

tilting at windmills, it seems we should modify the overgeneralized claims of the more enthusiastic supporters of RCT and allied terminology so as to retain a reasonable position to consider." (p. 168) The limited RCT position, according to Scriven, involves the use of RCTs "to investigate current (i.e., limited) issues about the effect of typical large-scale interventions in areas such as education, health, and social services." (p. 168) Scriven does not so specify, but in my view, champions of the "limited RCT position" also assume a multifaceted causal background, with an array of forces other than the treatment that may also affect outcomes of interest (Campbell and colleagues' lists of internal validity threats are an attempt to catalogue these). For instance, one expects that children (or other treatment recipients) will change over time even in the absence of any treatment effect, and that pre-existing individual differences will affect the outcome. Moreover, natural groupings of individuals into treatment conditions may be confounded with other factors in ways that can create biases in estimates of treatment effects.

Scriven contends that RCTs are not superior, even in the limited RCT position. One reason Scriven gives is that, in educational and most behavioral applications, RCTs typically do not include double blindness (whereby both investigator and research participant are unaware of the participant's condition). Absent double blindness, expectancy effects including Hawthorne are possible. As a result, Scriven argues, a good alternative method "is not essentially disadvantaged against the RCT; both leave open other explanations of any effects." (p. 167) Actually, Scriven would not have had to point to the absence of double blindness to make this point. The common use of hypothesis testing statistics in RCTs leaves open another, simpler explanation, that is, Type I error. Regardless, Scriven's point seems to be that RCTs are not "bulletproof," that other methods can provide equally compelling evidence about treatment effects, and that judgments about relative merit of methods should be made in the context of specific studies.

Scriven supports this notion, and further argues against the limited RCT position by presenting a hypothetical study of the effect of the HIP instructional technique. In a non-equivalent groups, pretest-posttest quasi-experimental design, morning and afternoon sections of introductory physics serve as treatment and comparison groups. I agree with Scriven that studies other than RCTs can provide reasonably compelling evidence of the effectiveness of interventions such as the new instructional procedure. But at least three challenges seem to apply to Scriven's general rejection of the limited RCT position.

First, consider the several things that went right in the hypothetical example: equivalent pretest scores, a big treatment effect (with treatment group students showing about one standard deviation more improvement than the comparison group), replication of that big effect across two semesters, the presence of observers who see no confounds, and no other reported validity threats. If all of these circumstances hold (and if no other plausible validity threats are identified), the study's results seem credible. But what are the odds, a priori, that all of these will fall into place in practice?

Second, what is the plausibility that a hidden or lurking source of bias exists in the hypothetical study? The classic concern, in the absence of random assignment, would involve some form of selection bias, that is, the possibility that pre-existing differences in the groups exist and are responsible for the observed posttest difference between groups. Selection bias is made less plausible in Scriven's example by the pretest equivalence between groups and by the replication across semesters despite switching the HIP intervention from the morning to the afternoon session. However, consider this possibility. The most motivated students (most motivated, not most knowledgeable about physics) tend to register first. They also tend to look for courses being taught by professors known on the street as the best. And perhaps one semester the competing favourite professors are mostly in the morning, and the next semester mostly in the afternoon. This scenario could explain what appears to be a treatment effect in the example.

Interestingly, the quasi-experimental design Scriven presents would also appear to be susceptible to an experimenter/instructor expectancy effect. Often in studies such as this, the instructor is a believer in the new instructional technique being tested. Could that translate into subtle forms of teacher expectancy that the observers did not see? This seems at least as plausible as the concern about Hawthorne effects that Scriven uses as a general criticism of RCTs. (After all, if Hawthorne effects were commonplace and potent, wouldn't more interventions be successful?)

Third, Scriven assumes that only large effects are of interest. He endorses the hypothetical quasi-experimental design "knowing that it's a net that will catch only big fish, but you don't want little fish." His emphasis on large treatment effects in general reduces concerns about the quasi-experimental design, because usually it is less plausible that a validity threat can explain away a larger effect. But is an a priori restriction of interest to "big fish" reasonable, especially in domains in which past efforts suggest the typical fish is not huge? As in the classic case of the early trial of aspirin's effects in preventing heart attacks, aren't small effects sometime important?

Taken together, these three challenges may raise the question of how compelling Scriven's argument is against the limited RCT position. This is especially so when we think of research planning and funding, as opposed to after-the-fact evaluation of individual studies. In the IES priority, for instance, the key consideration would seem to be the expected validity of alternative designs for estimating the effect of educational (or other) interventions, given

reasonable expectations about the likelihood of such factors as selection biases. A related subquestion would involve the ability of review committees (or other interested parties) to predict the presence and magnitude of various internal validity threats. For instance, could a grant review panel, prior to completion of Scriven's hypothetical study, know with confidence that the treatment and comparison groups would be equivalent at the pretest? And would they agree that only a very big treatment effect would be of interest? Alternatively, to what extent is random assignment (or other design features) preferred as a kind of preventative measure, as an admittedly imperfect insurance policy against validity threats that could plausibly appear and weaken a study's conclusions?

In this brief commentary, I raise these questions without attempting a definitive answer. Instead, I turn to what I see as two of the noteworthy strengths of Scriven's chapter. One is his thoughtful discussion of funding priorities, not for individual studies but for portfolios of studies. As he notes, even if RCTs were preferable in general, it would likely be better not to invest all research funding on RCTs. This is a compelling point, perhaps lost on those who apply so-called gold standard rules indiscriminately. Consider external validity, that is, the accuracy of inferences about the generalizability of a finding. Even if RCTs give the most internally valid, unbiased estimates of the effects of a program, external validity concerns might best be served by mixing study types in the overall research portfolio. If random assignment is feasible only in limited circumstances, then adding more RCTs could easily be less informative than funding an alternative study type. Another noteworthy strength of the chapter is Scriven's concluding section, which deserves your re-reading.

Were these commentaries longer, it would be easy to pick various nits. For instance, the question of generalizability merits additional attention. In a longer commentary, it would also be possible to applaud various additional points Scriven makes, such as his reference to the "general elimination methodology." That said, it is time for you to re-read Scriven's conclusion if not the entire chapter.

Note on Contributor

Mel Mark is Professor of Psychology at Penn State University. Mel Mark's interests include the theory, methodology, practice and profession of program and policy evaluation. A past president of the American Evaluation Association, he has also served as Editor of the American Journal of Evaluation where he is now Editor Emeritus. He researches applications of recent models of affect to prevention and to risk-taking behaviors, and the appropriate use of social science research in social policy, particularly in the context of program evaluation.