The Logic of Causal Investigations

21

Michael Scriven

Abstract

A familiar quest in the realms of educational research is to seek and produce knowledge that is 'beyond reasonable doubt'. Framed as a journey and marked by challenges of 'burden of proof', it has become closely associated with a logic that argues the validity and superiority of a research design by applying a randomized controlled trial. But what is actually required if research about education is to be generalized beyond the task and context to hand? Is establishing the theoretical basis and limits of inference non-negotiable? Indeed, do the standards of evidence in educational research always require claims that are testable and verifiable by others, or when might some other consideration, criterion, or quest be more apt? Such issues are widely debated in the literatures on educational research and evaluation, attracting particular scrutiny when a 'gold standard' is proposed for what to prefer – and fund – as studies of education. The chapter illustrates why critical observation and logical reasoning matter in interpreting the quality and usability of particular methodologies and findings as much as in designing a research strategy, especially when due consideration of theoretical, methodological and practical constraints is absent or muddled in accounts of research design and meta-analysis. The chapter then, raises the value of pausing: at recognition of configuration rather than defaulting to the pursuit of causation amongst factors, and at the challenges presented by elevating certainty to the status of sine qua non for legitimate, research-based knowledge in education. In short, monolithic research strategies are found wanting; while the promise and prospect of moving beyond impasse between 'warring parties' – such as via mixed methods – are also considered.

Keywords

Casual claims • Randomized controlled trial • Experimental design • Quasi-experimental design • Mixed methods

Introduction

It may be time for a new attempt at an overview of the alternative ways to establish causal claims. Much mainline scientific research centers around causal claims, as does much research in history, legal studies, and evaluation, and the current 'causal wars' threaten the validity of a great deal of this work. Moreover, much of this work, for example, in the evaluation of services and medicines for international aid, affects the approval and hence the delivery of vital products and services, so the welfare of large numbers of people depends on these decisions about validity. The proposal here is a composite of some of the existing, often conflicting, arguments with some new ones, but it is not a compromise that will appeal to all parties since it contradicts or omits several other existing arguments, and some of the new twists will surely be resisted. But it does incorporate

M. Scriven (✉)
Division of Behavioral and Organizational Sciences, Claremont Graduate University, Claremont, CA, USA
e-mail: mjscriv1@gmail.com

A.D. Reid et al. (eds.), *A Companion to Research in Education*,
DOI 10.1007/978-94-007-6809-3_21,

strong support for major planks from the competing platforms and strong support for the importance of some cooperative efforts. It may be the most inclusive defensible approach to a compromise.

The Position Focuses on Three Main Themes

A. It seems useful to distinguish about eight methods or groups of methods for establishing causal claims to the highest scientific standards of validity that are widely used in mainstream scientific domains. One of these methods is the randomized controlled trial (RCT), the entry currently sanctified by a large group of good scientists, but the validity attained by its use is not in the slightest degree superior to that achieved by any of the others, across the board. As with all the others, there are important research problems posed in a context of particular investigatory constraints, when an RCT design is demonstrably the best choice, just as there are many other cases where it is useless or substantially inferior. One reason that it has recently attracted a great deal of support, including frequent reference to it as 'the gold standard' for causal investigations, is here described as its 'conceptual superiority.' While it certainly has this property, this is widely and incorrectly supposed to confer upon it some *general* advantage in achieving greater validity. It has that advantage only in a head to head competition in cases to which it is best suited; in other cases other approaches will give equally certain results when RCTs cannot operate at all, or can operate but represent a bad choice, for reasons of, for example, cost, delays, ethics, or side-effects.

   The reasons why conceptual superiority is not the same as sole ownership of certainty ('validity superiority') are set out below and these reasons, along with some indisputable problems with the RCT design in the contexts where it is favored today, make the application of the term 'gold standard' to it basically a sign of scientific illiteracy. A more severe case of this affliction is symptomatized by the suggestion that the phrase 'evidence-based' should be restricted to causal claims supported by RCT research. This view is classified as more severe since it not only dismisses the possibility of providing evidence of causation in half a dozen other entirely satisfactory ways, but it also dismisses as lacking in evidential support a raft of highly significant and frequent causes such as warfare, heroin, plate tectonics, sexual attraction, political pressure, heavy smoking, and HIV/AIDS, none of whose mighty effects were ever established by RCT studies.

B. There *is* a gold standard for determining causation, and it is the same gold standard that applies throughout the whole of science, namely the use of critical observation, either on its own or connected to conclusions by logical reasoning. That is, causation can be directly observed, in lab or field: it is simply one of many contextually embedded configurations such as lead melting in a crucible, eggs frying in a pan, or a hawk taking a pigeon, that can be directly observed. And causation can of course also be inferred from non-causal direct observations, as by the coroner performing an autopsy. How the latter is done, with or without the use of a theory,[1] is set out here with some care, since the particular logical connections required to establish causal conclusions by inference deserve to be more widely understood in their own right as part of the logic of causation, and also need to be distinguished from three logical properties that are often incorrectly proposed as relevant to this task: (i) the 'conceptual superiority' property of RCTs; (ii) the scientific desideratum of replicability; (iii) the scientific desideratum of falsifiability.
C. Assessing the merits of any or all of the several types of causal investigatory design involves two standard types of evaluation (grading and ranking) and is addressed in the two preceding themes.[2] A more complex logical type of evaluation is involved in decisions about the funding of research (apportioning, the type of evaluation on which portfolio or budget allocations are based). In assessing proposed solutions to each or these problems, there is an important ethical requirement, but while this can be accommodated in RCT design (by using great care in the details of the design), it has a more important role in the evaluation of any research funding strategy that gives a dominant role to RCT designs. The key theorem governing the latter, established in the chapter, is as follows. Even if a particular method such as RCT were the best method for investigating causation whenever it can be used (a possibility that is disproved in the discussion of theme A) it is impossible to defend the strategy or policy of allocating research funds to RCTs in (anything like) all and only those cases. Such a strategy, now

[1] For an example of observations that, mediated by a theory, demonstrated causation, think of the observations of the solar eclipse of 1919 that showed sunlight was refracted by gravity as predicted by the general theory of relativity.

[2] The eight types of causation are based on: (i) direct observation, e.g., visual, affective, tactile; (ii) reported observation, e.g., case studies; (iii) eliminative inference, e.g., autopsy, engineering breakdown; (iv) theoretical inference, based on use of an analogy/theory, e.g., physics, geology, astronomy; (v) direct manipulation e.g., in the kitchen and lab; (vi) 'natural experiments' e.g., meteorology, epidemiology; (vii) quasi-experimentation, e.g., medicine, pedagogy; (viii) RCTs e.g., pharmacology.

widely enforced or recommended, is not only logically indefensible, but also can be shown to be morally unacceptable because of the effect it has on those dependent on the service that may be withheld because of that strategy.[3] While it has the advantage of being 'fail safe' in avoiding one type of error, it risks being overconservative when urgent needs are involved, and the balance between these two considerations in fact goes heavily against it.

The Ontogeny of Causation

The first of the two foundation stones to be laid in constructing the logic of causation is the proposition that causation is directly and reliably – indeed trivially and universally – observable. It is perhaps best to approach the proof of this by looking at how we acquire the concept of causation. It is developed in the child's brain before language skills are well developed and it springs from the palmar ("grip") reflex which soon develops considerably: into the child's realization that s/he can manipulate the environment by shaking a rattle to make a noise; the recognition and manipulation of crayons for producing marks on paper; and the discovery that squeezing the cat makes her scratch. These are all cases of understood causation and indeed, by the age of 3 years, the average child has discovered some things that are much more sophisticated, beginning with how to cause others to do things upon request – and indeed becomes notably 'bossy' about such demands. Also acquired are the basic notion of responsibility for his or her actions, resulting in blame when they are bad and praise when they are good, and the disclaimer of responsibility for 'bad' actions – e.g., knocking something over – when the wind or a sibling did it. Soon there is language to express all of this, and the youngster rightly claims to see others do things, as well as being able to get others to do things. In other words, their experience now includes the management as well as the observation of causation, and the evaluation of consequences. Maturation simply brings greater range and sophistication to these basics, so there's nothing essentially new about such claims as the adult makes, for example, that the brakes are working well in his or her car, every datum for this generalization being a (tactile) observation of causation.

[3] My thanks to Ryoh Sasaki for raising some problems with an earlier draft of this overview that I hope to have resolved in this version.

Direct Observation of Causation

Despite the commonplace use of our language to the contrary, Tom Cook (at least sometimes, in conversation) and many well-trained social scientists and educational researchers, following Hume, find it hard to accept the notion of observed causation. They appear to favor the idea that we are 'really' inferring it. But that's like saying we 'really' infer that this person we see in the crowd meeting the passengers from our plane into San Francisco is our spouse. Of course, the neural net is, in some sense, putting bits and pieces together, but that's part of what happens in *perception*; the end result of these neural machinations is pattern *recognition*, not pattern *inference*. Hume's pitch was seductive because we don't see causation in quite the same way we see color and motion. Causation, like many other complex predicates, refers to a learnt holistic feature of a *configuration* not just to a learnt *element* in it. That configuration is what enables the billiards player to say, in suitable circumstances, that he did indeed see the cue ball strike the object ball, and thus *cause* the latter to head for the pocket. Once one learns how to see this kind of example of causation, it becomes part of the *perceptual* vocabulary, like the myriad instances of your friend's face, or even part of perceptual evaluation, for example, part of what is called a good seat in an equestrian, or good style in a dismount from the parallel bars or in dunking a pass in basketball.

This epistemic status of causal claims as observable is fully recognized in the one place outside science where doubts are best respected, the court of law. Eyewitness testimony, especially but not only if it meets all the well-defined standards (normal vision, good lighting, clear field, propinquity, recency, corroboration, absence of motive to lie, and so forth), is treated there as in science, as an appropriate datum in the court of last resort for establishing a case. And the examples of it regularly include testimony that causation was observed in the standard cases such as battery, vandalism, and shooting. Causation is part of the language of observed acts, and as part of the language of observation, in suitable circumstances, it is established as having occurred with all the credibility that observation deserves, in science as in law.

So the first key conclusion here is that the simplest and probably the most reliable of all ways to establish causation is by critical observation. (I use the term 'critical observation' here as shorthand for observation subject to the usual checks for the usual sources of error, including reflection on the likelihood of those.)

Interestingly enough, close study of the bible on quasi-experimentation, Cook and Campbell (1979), turns up a passage in which this view is conceded, although its implications for causal methodology were never developed there:

> . . .we do not find it useful to assert that causes are "unreal" and are only inferences drawn by humans from observations that do not themselves directly demonstrate causation
>
> (Cook and Campbell 1979, p. 33).

This position leads us to the second foundation stone for the logic of causal inquiry.

Scientific, Legal, and Practical Certainty

One of the main attractions of the RCT approach is that it appears to provide a greater degree of certainty than alternatives. There is a sense, or at least circumstances, in which this is true, but it is not true across the board for several reasons, of which the first is that causal claims based on direct critical observation attain the benchmark level of certainty, and it's very hard to find an RCT that matches that standard. The 'benchmark level' in scientific research, as in the criminal law and in common practice where important matters are at stake, is simply "beyond reasonable doubt." This is the standard required to establish a case in criminal law, and is traditionally and extensively distinguished from 'the balance of evidence' which is the criterion for establishing the occurrence of misdemeanors.[4] This concept of certainty is part of the common language of science, so that in the lab or field, the observer or reasoner knows when to make and how to understand a claim that someone is certain that they did, or saw, or calculated, that some description applies.

This is not careless use, or abuse, of the term, it is the proper use of the term, it illustrates what the term does in fact mean. Some strands of perfectionist argument in epistemology here, as with perception and causation, have sought to persuade us otherwise, pushing us in this case towards the idea that the proper use of 'certain' refers to the complete impossibility of error as in definitional claims and mathematical theorems. But 'certain' is a contextually defined term and the proper standards for its use in the context of empirical discussions is empirical support beyond reasonable doubt, not the same standards as apply in talking about the realm of deductive proof. One might as well argue that the term 'large' is improperly used of an emu's egg, or an elephant, or anything smaller than the universe. The perfectionist move is just an example of bad linguistic analysis. The law courts remind us that there is a well-established body of rules for the proper use of terms like 'observe' and 'certain' beyond what is sometimes scoffed at as the imprecision of ordinary usage. The courts define the hard core of ordinary use, since that is what the juries understand, and that is what good scientific use employs. Even Tom Cook concedes, in his magisterial review of the arguments for and against the RCT design,[5] when talking about case studies (where we often rely on reported observations of causation), "I do not doubt that these procedures sometimes reduce all reasonable uncertainty. . ." (ibid., p. 38). And that is just the conclusion we need to establish, for that is all that can be reasonably required of any scientific method for establishing causation.

The RCT Advantage

Now RCT designs do have an edge, although not the edge that is often claimed for them. As Tom Cook goes on to say in the quote begun just now (ibid.):

> I do not doubt that these procedures sometimes reduce all reasonable uncertainty, though it will be difficult to know when this has been achieved. However, I do doubt whether intensive, qualitative case studies can reduce as much uncertainty about cause as a true experiment. That is because such intensive case studies rarely involve a totally credible causal counterfactual. Since they typically do not involve use of comparison groups, it is difficult to know the group under study would have changed over time without the reform under analysis
>
> (Cook 2000, p. 38).

The first problem with this passage – and with this position, which is the basic argument for the superiority of the RCT (a.k.a. 'true experiment') design, is that the RCT design, as used in the cases under discussion here, does *not* support a counterfactual. The RCT design as used in traditional pharmacological research, does have this property. But in educational and community interventions, the design is crucially weakened and is no longer double-blind, as it is in the drug studies. It is not even single-blind. That is, both the subjects and the researchers know who is in the experimental group, and usually both know which subjects are in the control group. This leaves open a gap through which the Hawthorne effect (and its converse) can slip in. (I refer to this situation as a 'zero-blind' or 'unblinded' condition.) As Cook and the texts define the RCT design, the key point about it is that after randomization (and assuming adequate group size) the only relevant difference between the two groups is the treatment; but in the context we are discussing, there is another difference, namely the difference in the

[4] Robert Brinkerhoff makes this case very well in his book, *The Success Case Method* (2003) in the course of defending his high quality case study-based approach to causation.

[5] "A critical appraisal of the case against using experiments to assess school (or community) effects" *Education Next,* 2000, Hoover Institute, Stanford. This paper of his represents by far the most sophisticated support for the RCT position in print, and for that reason I focus on it frequently in this discussion.

beliefs of the subjects and experimenters, which we know can cause effects of the type and size we are finding from treatments, and so we cannot conclude that differences in outcomes must be due to the treatment.

So the intensive case study (and the same applies to good quasi-experimental designs and critical observations) is not essentially disadvantaged against the RCT; both leave open other explanations of any effects.

Might it not still be argued that the RCT has an edge in only having this one loophole, whereas in the other designs there are, at least typically, more possible counter-explanations? This is not as telling a point as it might appear, since the total probability of the alternative explanations is not additive by the number of their descriptions; it is entirely situation-dependent. There will be situations where the Hawthorne possibility is more threatening to the RCT than the totality of the alternatives is to a case study or quasi-experimental design; this will be quite common in the case of the regression-discontinuity design for example, but will occur in many other cases. So the RCT edge, significant though it is when the design is double-blind (although even then not a general edge on validity), is entirely situation-dependent in the normal context of social and educational inquiry. It will still be significant in special cases, but non-existent or negative in others.

The second problem with the quoted argument for the RCT's superiority is that causation may occur in the absence of support for a counterfactual, as it does in cases of overdetermination. I have discussed these cases at some length elsewhere[6] and will only remark that it is a significant although not crucial weakness of RCTs for the purposes of the present discussion that they will not have any advantage at all in such cases, whereas case studies (and some other approaches, e.g., those based on theory) will do so.

The bottom line here is that the advantage of RCTs is by no means general and must be established in the particular case, a non-trivial task. It remains true that there are cases, including important ones, where the RCT design will settle the issue of causation and no alternative approach will do so as well. However, the same is true for many other designs. The conclusion for researchers is simple; each case needs to be highly specified, including not just the exact question we need answered and the degree to which we want to be able to generalize the answer, but the exact constraints on time and resources and social context, before one can decide on the optimal design for an investigation. That analysis obviously is not best done by those who specialize in RCTs alone; it must involve serious discussion by a panel including those expert in alternative approaches of several if not all of the kinds listed earlier. As Cook stresses, relying on a single approach is a methodological error and a serious one; and relying on the wrong one compounds the felony. Using a panel which favors just one approach would be a further felony in itself.

[6] For example, in "Causes, Connections, and Conditions in History" in *Philosophical Analysis and History* ed. W. Dray (1966).

The Other Contenders

Every child acquires a repertoire of possible causes for a large number of effects before reaching school age; for example, they know that the vase on the table by the window can be knocked over by the wind, the shades, the cat, a sibling, a playmate, or a grown-up. When they encounter the effect, they begin to sift that list and check for indicators, either immediately observable or quickly accessible, that will eliminate one or more candidates and eventually may identify the responsible cause. This is the basic case of hypothesis creation and verification and it is the essential element, even if subliminally and non-inferentially, in all careful causal explanation.

There is a background assumption for this enterprise – the assumption that everything has a cause. Truth of that assumption in the macro-domains of everyday experience and scientific investigations is unaffected by the discovery of micro-uncertainty, not because the latter phenomenon cannot manifest itself at the macro-level – indeed it can – but because it has a sufficiently small incidence at that level to leave the deterministic principle unaffected as a methodological guide.

The two key components in the basic procedure outlined are the 'list of possible causes' (LOPC), based of course on memories of prior personal or reported observations, and the 'general elimination methodology' or algorithm (GEM). Both become increasingly complex as the individual's experience and learning expand, for example by the addition of theories about possible causes that are extrapolations, or extensions by analogy or speculation, from a human's direct experiences, Take, for example, the theory of tectonic plates that added to the list of possible causes of mountain ranges. No-one saw those plates collide and raise the Rockies or the Urals, but we all can visualize what happened on a smaller scale, and once conceived, we add it to the LOPC and can readily project the kind of clues in the geologic record that would confirm this etiology, thus kicking in the GEM process that in fact confirmed the hypothesis.

When the hypothesis is about the formation of star-clusters, we begin to move beyond models that are based on analogies with ordinary experience, and instead create formal models that extrapolate from those models or even from models that seem to have worked in other areas beyond direct experience. So the piggy-backing continues, stretched to its limits with string theory at the macro limit and boson/hadron models at

the micro limit. In all of these realms, however, the concept of causation continues, usually unchanged by the changes in the forms to which it applies, except for quantum uncertainty, where it, too, must be modified significantly. And in all these areas, for all these kinds of causal claims, the same procedures of investigation apply; that is, the process of LOPC identification, and GEM application to whittle the list down in particular cases.

This vast web of theory-driven causation is essentially independent of any direct experimental confirmation since it deals with entities that are largely beyond the range of manipulation. The large hadron collider at Geneva, is the culmination of the main exception to this segregation, the zone of experiments with fundamental particles. But even there, where the term 'experiments' is always used, it does not refer to experiments with randomized controls, but to those ruled by simple pre/post design, entirely adequate in those circumstances to establish the conclusions to the satisfaction of the Nobel Prize committees. It is simply absurd to suggest that the conclusions arrived at in these circumstances do not deserve to be called 'evidence-based' because there is no RCT in sight. To avoid tilting at windmills, it seems that 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.

The Limited RCT Position

These controversial terms should be flagged in some way to indicate that they are not intended to be taken in their normal, all-contexts, sense, so we'll add the prefix 'limited' to their use in these restricted senses. This has the effect of converting positions that are absurd into ones that can be argued against without using such language as 'absurd'. Instead of saying, as the head of the Institute for Educational Science has pronounced, that there is no scientific basis for any causal claim that is not based on RCT studies, we'll take that as meant to apply to zero-blind (a.k.a. limited) RCTs used to investigate current (i.e., limited) issues about *the effect of typical large-scale interventions in areas such education*, *health*, *and social services*. Correspondingly, we'll take the term 'evidence-based' which is often said to be justified only for the results of RCT studies, to be intended to apply to limited RCT studies, *only* when talking about the effects of that kind of limited intervention in those limited kinds of areas.

The thrust of the preceding arguments is then that the limited claims are nevertheless wrong, even if not absurd. That is, even the view that only (limited) RCTs can establish limited causal claims is wrong, since they can be perfectly well established beyond reasonable doubt in other ways *and* the limited RCTs aren't bulletproof themselves. Similarly, the view that the only (limited) evidence-based claims are those supported by RCT studies is wrong, since even limited evidence-based claims (that is, claims about typical current types of intervention in health, education, and social services) can be established by quasi-experimental, observational, and theoretical studies. Finally, the claim that 'experiment' means a design with random allocation to the two groups should also be modified to the formulation "limited experiments are those in which subject are randomly allocated to groups."

This triple modification prevents what many have seen as an extremely tendentious, if unconscious, attempt to hijack an important slice of the scientific vocabulary.

Quasi-Experimental Designs (QXTs)

Let's walk through the consequences of the preceding arguments, using a common 'lower-class' QXT, the pre-post design with comparison groups. The argument given here would be much stronger with what is commonly thought of as a fancier a.k.a. more robust design, for example the interrupted time series design with bounded randomization of the intervals between applications. The example we'll use is an example of the use of the highly-interactive paradigm (HIP) for large introductory lectures at the college level, with enrolments in the low three-digit category. We divide the entering class in about half, in some convenient but not random way, e.g., by taking the morning class as the treatment group in the first semester of the experiment (sic) and using the afternoon class for the second semester of the experiment (this is a one-semester class). The experimental group receives the new treatment, the others get the same approach that has been used for several years; the same instructor teaches both classes and teaches the control group just as he has for some years. That claim of approximate constancy in treatment is verified by an experienced colleague who visits a few times unannounced, and by a Teaching Assistant who's done that job previously and now works for both classes. Let's add that we have an experienced pair of instructional researchers independently look for other differences and find none to remark on. Each class gets the same pre-test and post-test; they match closely on the pre-test but on the post-test, where the control group shows about one sigma of improvement, the HIP group scores about two sigmas better than that; and this effect recurs on two reiterations, the only two that are done as of report time for the 3-year experiment.

Now, did HIP have any effect? Given that you know the important 'local knowledge' fact that it's extremely hard to kick a sigma difference out of any variation in instructional method, and that two sigmas is considerably more than twice as hard to get as one, the answer has to be, yes, HIP made a big difference.[7] Clearly the size of the difference is crucial here, as is often the case. Conclusion: there are situations where non-RCT designs will provide support for causal claims, beyond reasonable doubt. If you now reflect on exactly what it would take to convert this study into even a limited RCT study, and on the fact that you are not very interested in small differences, because they have a track record of never showing up on the replications at distant sites, you should be willing to buy the conclusion that the pre/post/comparison study design is better than the RCT here. (That is, you use it knowing that it's a net that will catch only big fish, but you don't want little fish.)

There are a dozen variations on this kind of case, ringing the changes on such matters as dealing with cases where you are only interested in generalizing to the native population in Uganda, but the native population won't give permission for putting their children into the control group; or you can't afford the cost of measurement and monitoring for the control group of ex-addicted homeless, and the memory effects of vitamin shots are small, so interrupted time series will work well.

So the bottom line is that there are many cases where non-RCT designs will be better than RCT ones for the cases of interest, cases where they will indeed achieve results beyond reasonable doubt, and even more cases where they will be better than limited RCT designs, the only ones we're really talking about. The limited strategy, which protects the RCT position from absurd overstatement, still cannot save it from being beaten on its own ground, that is on ground where it *can* be used. Like a good two-wheel drive car, it can be driven in snow, but it's easily overmatched by the four wheel drives in those conditions.

The fundamental logic of causal investigation, that is, the rules of inference required for establishing any causal conclusion, is not the use of experiments in the limited sense, it is the use of a critically developed list of possible causes together with critically applied general elimination methodology, required even for the justification (although not the occurrence) of critical observation.

[7] There is a huge online literature of debates about whether pre/post testing provides a valid basis for such conclusions, one of the reasons I use this example. Google Richard Hake (and gain scores), to see the whole debate well-referenced by the leading proponent of the common-sense position here, which is of course that pre/post works fine for any worthwhile effects if done carefully.

Funding Strategies

It is now time to turn from the logic of grading and ranking experimental designs to the distribution of resources between them. The first lesson to be learnt about the logic of portfolio construction is that the best single investment – better than every alternative although it may be – is not the best bet for the whole portfolio budget. Investment managers know very well that the rule about not putting all one's eggs in one basket is not just valid for the medieval housewives who inspired the adage and their successors. Provided only that one's second and third choices still meet the minimum acceptable standards for good eggs, they are better choices than further investment in the top pick for at least half the portfolio.

In research funding, a much-better-than-minimum-acceptable standard is the ability to produce conclusions that are beyond reasonable doubt, so even if RCTs were superior in their ability to yield such results across the board (which is itself a true counterfactual) it would be highly unscientific to back them across the board since they, like all other designs, can go wrong, badly and completely wrong, in a way that is usually not reversible even if detected, and is not always detectable at mid-stream.

But in scientific research there are two distinct further reasons for the heterogeneous strategy besides protection against failure. It may be helpful to think of the analogy with an investor who decides to put some money into the stock of companies who are working on a new oil field. She could put all her money into one company that has an excellent production hole that is currently the best on the field and is planning to drill more wells on that site. But she knows that a single site can peter out, hit an artesian aquifer that drowns their wells, or run into labor trouble; so for simple *safety* reasons – our first consideration – she will buy into at least one other outfit. She's covering herself against the possibility of failure. But there's another reason to do so: wells that begin with a less than stellar rate of production sometimes hit another pocket below the first and do much better later: so there's a chance of doing better by approaching the formation from two directions, even if the second one is less productive at first. That's the possibility of *superiority* via backing an independent approach. And there's a third reason, too, probably enough reason to justify investment in a third wildcatter. This is the chance to get a better overall picture of the layout of the field, which will be invaluable in guiding further action or withdrawal. This third consideration of course corresponds to getting some information about generalizability – external validity. Safety, possible superiority, and generalizability; three reasons for avoiding the monolithic strategy. The analogy carries over completely to the issue of funding research; a point that Cook stresses in the paper previously cited.

This argument does not dismiss the possibility of doubling the investment in the best option, just not restricting all investment to it. Doubling in the research case would make especially good sense if combined with slight variations in the research personnel and population used. But it still comes further down the list than variations in the primary strategy. Somewhere in between these two major paths to enrichment of a single design approach there is the use of the superbly ingenious list of ways to match the comparison group without using random allocation, a list developed and provided by Cook.

The argument given makes an invincible case for the indefensibility of the present situation in which, according to the extensive testimony from members of the review panels that have talked about it, there is no serious consideration of using non-RCT designs instead of RCTs. That strategy is largely based on bad reasoning about the superiority of what are in fact flawed RCT designs, which even if peerless would be no better than many others and clearly worse in many cases. This bad reasoning is combined with the fallacy of assuming that such superiority, if it did exist, would justify a monolithic strategy.

The present ill-based practice is also too often combined with denial of the existence of the monolithic strategy, sometimes accompanied by a gesture in the direction of regression discontinuity funding. If there is any doubt that an essentially monolithic strategy is de facto in place, it would be easy enough to establish the facts by doing a survey of funding over the past 2 years, using a contractor with a team from both sides. It is certainly long past time for a meta-evaluation of the success of the new emphasis on RCT funding and the absence of any movement towards doing that surely shows a serious lack of interest in finding out the truth about the claims for improvement before proceeding still further with the takeover, especially in the latest area where it has established a beachhead, the evaluation of international aid.

It is important that the reasons against the monolithic strategy apply even, per impossible, the RCTs were superior across the board in the certainty with which they can determine causation. It is an argument designed to be acceptable to both camps.

Cooperation Combined with Competition

It would be unrealistic to suppose that the causal wars will cease in the light of the above treatment of the underlying differences between the competing positions. But it would be good to see some recognition of the very considerable range of cases where both parties can benefit from using the skills of the other. Tom Cook, in the article cited, lists many such cases, although not with quite the spin I'm putting on them, and I'll sum it up in my way by saying that it seems clear that the effective execution of RCT studies depends very heavily indeed on skills that are highly developed in qualitative researchers and extremely rare in the training of quantitative researchers. The converse position is also clear; there is still a considerable area in qualitative research where the skeptical reflexes of the trained quantitative researcher need to be heeded very carefully, not because their usual dismissive judgment is justified, but because by heeding their concerns, the design can be greatly improved, including its credibility to a wider audience, a worthwhile consideration in itself, and also, often enough, in validity. But let's take a final moment to look at the need for qualitative research skills in managing RCTs. We can begin with the two great threats to the validity of the (already limited) RCT design: differential attrition and cross-contamination. No one denies that some very expensive RCT efforts in the past have been completely brought down by these weaknesses. If these flaws are detected very early, it is sometimes possible to stop them in their tracks, before validity is hopelessly compromised. How can an early-warning system for them be set up?

The answer is almost always through a continual process of interviews, both with groups and face to face, with both students and staff, with parents and with administrators, a program which not only seeks by intelligent and systematic questioning to pick up the warning signs, but also builds trust and cooperation in what is after all a project with potential benefits for all.

Interestingly, there is a double reward from this activity. It not only provides good insurance for the validity of the study, but also vital evidence about the process whereby the causal agent – and any inhibitors of it – operate, which provides key clues to the possibility of generalizations in some dimensions and the improbability of generalizations in other directions, and often strong supportive evidence for the causal connection under investigation. Cook gives a long list of the valuable information that can be picked up by these observers in the source cited.

And who has the training for this kind of observation and interaction? Of course, it is a job description for a qualitative researcher. These are high skills, not often taught as of top importance in quantitative training programs. Cook actually gives as his reason for abjuring the term 'gold standard' for RCT designs the fact that these skills are of great importance, are hard to acquire, and rarely available. The bottom line is that RCTs are really hybrid designs, mixed-method in the sense of having essential quantitative and qualitative components.

So I end on this note. A marriage of the warring parties is not only possible but would provide a win-win solution, with major winning side-effects for those in need around

the world. The pre-nuptial agreement should include: (i) recognition of the place of duties for both parties, along with (ii) funding for non-RCT studies where they are better fitted to task and context than RCTs, with skilled quantitative researchers collaborating to cover both sharpening the design and analyzing the numerical data; plus (iii) at least one collaborative meta-evaluation panel funding proposals from both parties, and (iv) another one evaluating the success of contracts of both kinds. Serious concern with research standards (and human welfare) suggests that we should shortly see some proposals like this and/or signs that such proposals would be funded.

Note on Contributor

Michael Scriven is a professor of psychology at Claremont Graduate University, a Senior Research Associate at the Evaluation Center, Western Michigan University, and Director of Institutional Research at Palo Alto University. He is best known for his contributions to the theory and practice of evaluation in public policy and the social sciences. He is an editor and co-founder of the *Journal of MultiDisciplinary Evaluation*. He has also contributed to the fields of mathematics, philosophy, psychology, the history and philosophy of science, law, education, computer science, informal logic, cosmology, international philanthropy, and technology studies.

References

Brinkerhoff RO (2003) The success case method. Berrett-Koehler, San Francisco

Cook TD (2000) A critical appraisal of the case against using experiments to assess school (or community) effects. Education Next, Hoover Institute, Stanford

Cook TD, Campbell DT (1979) Quasi-experimentation. Houghton Mifflin, Boston

Scriven M (1966) Causes, connections and conditions in history. In: Dray W (ed) Philosophical analysis and history. Harper and Row, New York, pp 238–264