American Journal of Evaluation
31(1) 105-117

DOI: 10.1177/1098214009354918
http://aje.sagepub.com

Contemporary Thinking About Causation in Evaluation: A Dialogue With Tom Cook and Michael Scriven

Thomas D. Cook,[1] Michael Scriven,[2] Chris L. S. Coryn,[3] and Stephanie D. H. Evergreen[3]

Abstract

Legitimate knowledge claims about causation have been a central concern among evaluators and applied researchers for several decades and often have been the subject of heated debates. In recent years these debates have resurfaced with a renewed intensity, due in part to the priority currently being given to randomized experiments by many funders of evaluation studies, such as the Institute for Educational Sciences. In this dialogue, which took place at Western Michigan University in October 2008, two of the field's leading theorists and methodologists, Thomas D. Cook and Michael Scriven, described their current thinking and views about causation and causal inference in evaluation. They also discussed recent methodological developments for cause-probing investigations that sometimes produce results comparable to those produced by randomized experiments. Both Cook and Scriven prepared clarifying postscripts after reading the edited transcript.

Keywords

causal inference, causation, randomized controlled trials, methodology

The Evaluation Center's Evaluation Café, held weekly during the academic year, is a public forum intended to foster engaging dialogue and debate on a variety of evaluation-related topics. On October 24, 2008, The Evaluation Center and Western Michigan University's Interdisciplinary PhD in Evaluation program jointly hosted a special Café event on contemporary thinking about causation and causal inference in evaluation. Largely centered on randomized controlled trials (RCTs), causal inference in evaluation has been a point of high priority as well as contention, given the recent resurgence of debates regarding legitimate knowledge claims and conceptions of "evidence" in applied

[1]Northwestern University, Evanston, Illinois
[2]Claremont Graduate University, California
[3]Western Michigan University, Kalamazoo

Corresponding Author:
Chris L. S. Coryn, 1903 West Michigan Avenue, Kalamazoo, MI 49008.
Email: chris.coryn@wmich.edu

research and evaluation (Donaldson, Christie, & Mark, 2009). Two of evaluation's leading theorists and methodologists, Thomas D. Cook and Michael Scriven, were invited to discuss the topic not only because of their substantive expertise but also because they were expected to have differing viewpoints on the matter.

Cook is a professor of Sociology, Psychology, Education, and Social Policy at Northwestern University's Institute for Policy Research. He has authored numerous papers and books on theory of method and causal generalization, including *Experimental and quasi-experimental designs for generalized causal inference* (Shadish, Cook, & Campbell, 2002) with William Shadish and Donald Campbell, "Clarifying the warrant for generalized causal inferences in quasi-experimentation" (Cook, 1991), and *Quasi-experimentation: Design and analysis issues for field settings* (1979) with Campbell, among many others.

Scriven is a distinguished professor at the School of Organizational and Behavioral Sciences at Claremont Graduate University. He has written extensively on the philosophy and methodology of causation, including, for example, "Defects of the necessary condition analysis of causation" (Scriven, 1993), "Maximizing the power of causal investigations: The modus operandi method" (Scriven, 1976), and, more recently, "A summative evaluation of RCT methodology: & An alternative approach to causal research" (Scriven, 2008).

In the edited transcript that follows, Cook and Scriven describe their current thinking and views about causation and causal inference in evaluation. They also discuss and reflect on recent methodological developments and alternatives to RCTs for cause-probing investigations in educational and other settings. After reviewing the transcript, both scholars requested the addition of postscripts, which follow.

A Dialogue on Contemporary Thinking About Causation

Tom Cook: What I'd like to do today is engage in a dialogue with Michael about similarities and differences in our perspectives. One thing I am doing is pushing ahead a program of research on when studies done without random assignment regularly reproduce the results of studies with random assignment. It involves taking a randomized experiment and first calculating the effect sizes it produces for any given outcome. Then, one creates a nonexperiment designed to answer the very same causal question. The trick is to use the same treatment group in the experiment and nonexperiment. What varies is how the comparison group is formed. The logic is that the credibility of alternatives to the randomized experiment is greater if they often reproduce the same results as a randomized experiment. LaLonde (1986) began this line of research and Cook, Shadish, and Wong (2008) do so in other domains while also providing comprehensive criteria for determining how well a "within-study comparison" has been carried out. In any event, the main aim of this line of work is to provide evaluators with a set of arrows for their causal quiver instead of having to rely on a single one—the randomized experiment. But the set of acceptable alternatives is sharply defined by the ability to recreate experimental results on a regular basis. These are the alternatives most worth supporting.

The three nonexperimental alternatives identified to date as recreating experimental results are

1. Regression discontinuity, which has worked in all three attempts to reproduce experimental results, as reviewed in Cook and Wong (in press).
2. A geographically local intact comparison group that is matched on pretest scores. This has also worked in all three attempts to use it, presumably because it leads to minimal differences from the treatment group on a highly stable measure that is due to the group (and not individual) level of the matching. The finding suggests that hidden bias from unobservables may not be that frequent.

3. When the process of assignment to treatment is perfectly known. With initially nonequivalent contrast groups, researchers sometimes have excellent independent knowledge of the process of selection into treatment. Measures of this can then be used to equate the nonequivalent comparison group to the treatment group. The need is both to choose the right covariates describing selection into treatment on variables correlated with outcome and to measure these covariates very reliably.

The second line of research I am doing is on No Child Left Behind (NCLB). I have an interrupted time-series study with Manyee Wong on the national impact of NCLB. Given it is a national program, the trick has always been to find a counterfactual group of schools not exposed to NCLB. Using time series, we have created a couple of imperfect counterfactual groups from 1990 to 2007 at 2- or 3-year intervals when the main National Assessment of Educational Progress (NAEP) was given. One counterfactual compares public and private schools, and the other compares public schools in states that set lower standards for making adequate yearly progress and states that set higher standards. We are also doing a study of NCLB's effect on students with disabilities using a version of regression discontinuity (Wong, Cook, & Steiner, 2009). But there are great complications because a school can fail to make adequate progress for reasons over and above the percentage of its students who are proficient on the state test—for example, because of confidence intervals around these proficiencies, for reasons of safe harbor or growth criteria, or because schools successfully appeal the state decision about the school. So, assignment is by four or five mechanisms rather than the classic single one! Fortunately, these are almost all deterministic rather than probabilistic mechanisms, and the issue then becomes how to do a regression discontinuity with multiple deterministic assignment criteria rather than a single one.

The third line of work is quite different. I am studying how children's development from zero to eight is influenced by multiple physical, social, and institutional contexts considered jointly—the physical home and its immediate surrounds, family composition and internal processes, neighborhood composition and its social climate and institutional resources and, finally, educational institutions like child care, preschool, and school. Our social science is constructed around individual social contexts like schools, neighborhoods, families, and the like that do not capture the complexity of children' lives as they experience, for example, pretty competent parents, lousy schools, good peers, and very few neighborhood institutions (Fleming, Cook, & Stone, 2002). Children's lives are much more complicated than our academic way of framing them.

To summarize, I am trying to create an empirical warrant for alternatives to randomized experiments so that we have multiple arrows in our causal method quiver.

Michael Scriven: I am going to hand out a one-pager (see Appendix A), which summarizes some of the problems with the RCT approach. I am hoping that Tom will comment on these points and if we agree on enough we can regard this as an evidentiary service for the randomly controlled trial and if we don't then we will clarify.

To begin with, Tom's allegiance is not regarded as being 100% reliable by the RCT gang because he has a tendency to be rather more open minded about causal alternatives than the most vocal of the enthusiasts. As you can tell from his talk, he has really engaged in a set of other activities that are not really true RCT applications, but rather extending the range of legitimate approaches to other, somewhat different approach methodologies, an approach which I heartily endorse.

One reason why Tom was one of the early leaders of the RCT lobby is because he was fed up with (I hope you will correct me if I am wrong or misrepresenting you, Tom), a disgraceful mess of 20 years of really second rate, invalid research on interventions in human affairs. That in turn was a backlash to the first of the RCT swings of the pendulum in the late eighties/early nineties when there were many fruitless studies. They were fruitless mainly because they were either badly designed or badly executed studies, as Tom found when he looked at them very carefully, so that the best of

current RCT designs avoid many of those problems. But, to do this, they impose tough requirements, one of which is a somewhat ironic one. It arises because you have to really watch out for two of the endemic problems with RCTs that involve the deterioration of the experimental conditions through two leakages—one being differential attrition, so that (typically) the control group membership declines to the point where you are in trouble with the statistical significance of the comparison, the other being cross-contamination. The ironic part of this is that you have got to pick up these leakages or weaknesses very fast indeed, and stop them, or you have lost the design. And, to do that you have to have your finger on the pulse of almost everyone in the control group all the time; that is, constant watching by very skilled observers. But the skills that are required to do that are the skills of a good qualitative researcher, and of course the irony is that the people running RCTs usually do not have and in fact are not hospitable to very good qualitative researchers. So, the general problem is that not many of them meet and hold up the conditions that the RCT imposes.

The hard-line RCT position is the view that, with the possible exception of regression discontinuity, the RCT design is the only legitimate way to establish causal claims and that position is typically supported by three main pillars. The first is the claim that only the RCT design excludes all alternative causes, essentially because there is only one difference between the two groups. One problem with that arises because in many interventions, under general ethical constraints, all subjects must be informed that they are part of an experiment, and it takes a subject with an IQ of 95 or more about 2 days to work out whether she or he is in the experimental or the control group. Now the minute that this realization dawns on you the design immediately has two problems. One problem is that cross-contamination now becomes a possibility because you started boasting to your pals that you are getting the hot new treatment. But even if you have good watchers on that front, a more serious problem is that the Hawthorne effect (and the reverse Hawthorne effect) now comes in to do its worst.

And its worst is pretty serious. That is, it may be that there is now another difference between the two groups, other than that fact that one is getting treated in a different way, namely that one knows it is being experimented on, a state of mind that may itself be producing the whole effect, with no help from the intrinsic virtues of the treatment. Unless the design eliminates this possibility, which is only possible with some special types of treatment, what you are seeing in the usual arguments for RCTs is mislabeled. It's mislabeled marketing because you are not getting RCTs. RCT designs are normally defined as double-blind and these aren't double-blind designs.

The second problem, a technical one, is the often-quoted claim that the important and unique feature of the RCT design is that it's the only design that supports the true meaning of causation, which is taken to be support for the counterfactual claim. In other words, it supports the claim that if this treatment had not occurred, then the effect would not have occurred. Well that would be an advantage if it were true that the counterfactual approach were the correct analysis of causation. But for anybody that knows the literature on causation, it's not true. The counterfactual claim is invalid basically for the very simple reason that there are many cases of what is called overdetermination. That is, cases where a number of factors are present in the context, any one of which would have produced the same result even if the result was not produced by the intervention.

The classic example of this is the guy who has jumped off the top of a skyscraper and as he passes the 44th floor somebody shoots him through the head with a .357 magnum. Well, it's clear enough that the shooter killed him but it's clearly not true that he would not have died if the shooter hadn't shot him; so the counterfactual condition does not apply, so it can't be an essential part of the meaning of cause. In the medical and social sciences, there are many, many cases where that happens.

The third claim is more complicated, the claim that no other design does it as well. Now there's a little verbal trap here that one needs to try to get clear about. The verbal trap was the invention of the term "quasi-experimental," which gives you a strong sense that these designs are something less than the best in all respects. But quasi-experimental designs are not feeble attempts at the one true

experimental design, they are alternative ways to establish conclusions, often better ways in particular circumstances. You even hear a lot of nonsense from the enthusiasts for the RCT who say that the only true experiments involve control groups with random allocation of subjects to treatment and control. That is only true if you assume what you can't prove. True experiments involve pouring stuff into flasks and finding out whether the result bubbles or turns green; or whether or not if we double the volume of a gas container, the pressure is cut in half, and so on and so on. The history of science is full of these true experiments. They have nothing to do with control groups of any kind.

The issue about quasi-experimental designs that we are interested in here is a simple one: can any of them establish causal conclusions beyond reasonable doubt? That's what the scientist usually wants to know, just as the judge wants to know it in a felony case; has the case that the defendant intentionally caused the victim's death been made beyond reasonable doubt? Of course there aren't any control groups involved, and, just as in scientific research, there are plenty of bodies of evidence that establish causation in the law, as in science, without control groups, let alone random allocation of subjects to them.

Many quasi-experimental designs, for example a good interrupted time-series design, when you can match the conditions required, will establish the result beyond reasonable doubt. Some of the many others that can do this are listed under "C" in the appendix. Now, the one that I know Tom doesn't like is C1. He doesn't like the idea that you can observe causation. There is a long, very distinguished philosophical history of the debates on this, in which Hume says things like, when you see one billiard ball hit another and you say it caused the target ball to go somewhere else on the table, you did not actually see the cause as something apart from the movement of the one ball into the other, something you have seen often before and always accompanied with that result. So, the causal claim is just the claim of succession and constant conjunction.

The cause is not a separate entity that you see, it is a combination of features of the circumstances you observe. So, this is just a bad heritage from early empiricist cum positivist thinking which we no longer accept it, because clearly it's the case that when I [loud bang], make that noise, and I ask you what made that noise, you answer with a causal claim, "you dropped your keys onto the table top," and you're damn sure that is what happened, you saw it happen, and you're right beyond reasonable doubt. So, forget the idea that you can't observe causation, of course you can.

The second family of cases, C2, is highly scientific claims in the forensic sciences. We all know that when somebody does an autopsy of an individual who was shot with a .45 in the forehead at a range of one foot and has a hole going in the front and coming out the back, establishing the cause of death is not a rocket science problem, it's a straightforward problem. And we don't have any doubt, we don't wish we could run a control group, we don't even have data about how many times that doesn't work. But we're entirely clear about this one, we have dug around in the brain and found out there is nothing else in there, or elsewhere in the body, that could have possibly caused death.

And then, in the causal list, we have highly scientific claims in other areas, for example regression discontinuity, the one Tom has been working in lately, and then this colossal group C5 of the theory-based claims. Then, in the C6 group I put many other examples, because I think there's a group here which use this same General Elimination Method (GEM) as I call it, something like what is called "inference to the best explanation." We claim that the intervention has causes that are visible, and we do that by eliminating other possible causes in relatively systematic ways, a complicated but perfectly feasible process.

The ethical requirements about the randomized control group means that you have got to find subjects who are willing to face the fact that they have at least one chance in three, perhaps one in two, in getting no treatment and staying with no treatment for the duration of the study which may be quite a long time. That reduces your volunteer intake, so you may not be able to get anything like a random sample of the population that you'd like to generalize about. So you're really restricting your generalizability because of the nature of the RCT design, and that's a serious problem that doesn't apply to many other designs such as the interrupted time series.

I'll just finish by saying that for me a worrying feature of the RCT crusade is that nobody is looking at whether it is producing good or bad results, and it's peculiarly interesting that nobody has suggested doing an RCT study to see the intervention of forcing RCTs on behavioral studies. I think it's clear you could not use this design for such a study, which suggests there is something wrong with the view that it's the only valid design for causal studies.

Tom Cook: As I hope you'll see, my position is somewhat more nuanced than the position ascribed to me, at least I hope so. I am a fan of RCTs but a critical one and not for all of the same reasons Mike adduced. I think he is wrong on some points, and I also think he is not sufficiently critical of the RCT at other points.

The first point he made is that advocates of RCT designs are wrong when they claim that random assignment excludes all alternative causes. That is not a claim I would expect to find anywhere in the literature, though I know it pops up in casual dialog. Every advocate or practitioner of RCT knows that logic requires that he or she struggle with three conditions that can undermine a randomized experiment. Mike mentioned two of them.

The first is when you say you've done a randomized experiment but the procedure used for assignment to treatment is in fact not genuinely a randomly assigned one. The classic example of that was the first draft lottery at the time of the war in Vietnam. They put 365 numbers corresponding to each day of the year into a big, big bowl. If your birthday came out number one then you were eligible for the draft, and so on. A few years afterwards it was discovered that this lottery was not in fact truly random because they hadn't shaken up the numbers well enough. There are all types of mechanical ways that random assignment can get screwed up too, just as there are researchers who misunderstand random assignment and who confuse what is haphazard with what is formally random.

The second problem with an RCT that Mike talked about was differential attrition. It is also much less of a problem today because we are much more aware of it and know how to minimize it. In the famous 1960s New Jersey Negative Income Tax Experiment, where families were randomly assigned to different amounts of guaranteed income people dropped out of the study much more if they were guaranteed an income of zero or $5,000 than $25,000. This differential dropout was discovered in the first months of the study, and merely paying control and low benefit households a small amount of money to recompense for the time spent on measurement was enough to make the differential dropout rates disappear.

In this connection, consider the sample survey, probably the most successful thing social science has given the world. Why do we have such an effective, though sometimes bothersome, survey research industry? In part, it depends on an elegant statistical theory for selecting instances so that the achieved sample represents the population of interest within known limits of sampling error. But the industry also depends on the synthesis of thousands of little studies of individually mundane but cumulatively important topics like: What's the effect of having an interviewer who's black, brown, green, blue? What's the effect of doing the survey by telephone, face to face, by computer? What's the effect of having 5-point, 6-point, 7-point Likert scale? What's the effect of calling people at 2 o'clock, 4 o'clock or 6 o'clock in the afternoon? Researchers have done this huge amount of work over 50 years to improve the implementation of surveys in the real world. The same is happening today with randomized experiments. We're learning more and more about empirically validated implementation practices that reduce our problems, including those related to differential attrition.

There is a lot of writing in the random assignment literature about treatment contamination. While I am a great advocate for randomized experiments as the best single arrow in the evaluator's causal quiver, I'm not an advocate for it as being perfect or for it being the only arrow in the quiver. Mike thinks that others think it should be the only arrow in the quiver; or at least he seems to think that I believe this when I do not. In any event, you can and do sometimes get contamination in experiments. Teachers talk to each other about problems and they may be in different treatment

groups. The case you have to make is that there's something unique about contamination that arises because of random assignment.

Now you have to always ask yourself the questions: How effective is the cumulative learning of researchers about how to prevent, identify, and minimize contamination in experiments? Let me give you a case study of the latter, I did a study in Detroit, in Chicago, and Prince Georges County, Maryland, of Jim Comer's School Development program (Cook, Habib, Phillips, Settersten, Shagle, & Degirmencioglu, 1999; Cook, Hunt, & Murphy, 2000; Cook & Hirschfield, 2008). It's a whole school development program, and a randomized experiment was attempted in Prince Georges County. We were worried about contamination, and so every year we interviewed all people paid to implement the program in the schools, and all of the principals about a lot of issues, including contamination. There was a possibility of control cases borrowing parts of the treatment, and of the treatment schools not implementing parts of what they were supposed to and thus coming to resemble the control schools. We found 3 instances of this out of 10.

Well that's 3 out of 10 control schools, not 10 of 10. But what did they borrow? The Comer school program involved changing the school's governance structure so that teachers and parents play an important role in decisions about school goals, about monitoring progress towards these goals, and about making midcourse corrections. It also requires teachers going to Yale for a professional development workshop and then several times undergoing a shorter professional developmental program in their district. In none of the control schools did anybody go to Yale. In none of them did they get any professional development on the program. In none of them did they set up all the three teams as specified in the program theory. So, only some schools borrowed, and those that did borrowed only parts of the program and not those most central to its theory. One has to be very careful in thinking about contamination and describing it lest one use the concept too promiscuously, inferring that the program has been borrowed merely because some details from the program has been borrowed in some of the schools or classrooms.

We have also learned a lot by now about how to monitor contamination to see how frequent it is, and how to reduce it. Why are we in education all caught up today in hierarchical linear modeling? This is partly due the nature of the social structure of our society. But another reason for the current level of interest is fear of treatment contamination. So, one way we deal with suspected contamination is to make the unit of assignment more aggregated.

So, if we look to the first of Mike's three great pillars I think he's logically correct in the narrow sense that random assignment cannot by itself guarantee a secure causal inference. But I am not sure anyone believes this on even a moment's reflection, and I am sure that the behavior of careful experimenters belies this as they struggle to justify that a correct random assignment procedure was chosen and implemented well, that attrition was not differential by treatment, and that treatment contamination was minimal or otherwise adequately dealt with. Knowledge is accreting about how to implement experiments so that these problems occur less often and with lesser severity—just like how we learned how to implement surveys better.

The second pillar concerns the denial Mike makes of the claim that only the RCT is supported by the counterfactual analysis of causation. That is an incorrect assumption. If you look at Rubin's causal model, the currently preferred statistical version of causation, it clearly says that the correct counterfactual is logically impossible. The counterfactual we'd like to have is the same person undergoing the treatment and comparison states at exactly the same time and in exactly the same situation. But this individual level analysis of causation contravenes the laws of physics. So, all the modern discussions of causation among the statistics crowd start from the assumption that what we do in the RCT is second best in logic. It creates groups that are identical on expectation and not in the concrete study under analysis. That is, the theory claims that the groups would be identical if an infinite number of random assignments were made; not that they are identical in this particular random assignment in this particular study.

Let's go to the third pillar that Mike criticizes—that no other design establishes causal claims with certainty. It's a little more complicated than that. I think no one denies that you can make causal inferences without random assignment. Not even the most fervent advocates of random assignment believe it is a necessary condition for causation. Consider the published interrupted time series of the effects of Cincinnati Bell charging for directory assistance calls. They used to tell you for free but then they started to charge you for it. What happened to the number of calls to the telephone service for information? It drops about 15 standard deviations, an effect so large it hits you between the eyes. And it occurs right on the day that Cincinnati Bell changed its pricing policy for directory assistance calls. No one has yet come up with a plausible alternative to the claim that charging for directory assistance calls reduced the number of such calls in this one instance. Random assignment is definitely not necessary for cause.

But we in education do not ever get effects that large. Remember that the Institute for Educational Sciences (IES) sets up experiments to detect effect sizes of 1/5 of a standard deviation. But the point is that there are clearly alternatives to the random assignment. Let's use some common sense here. If Mike drops his keys and hears a thud, well it's a reasonable causal claim that the keys caused the thud so long as we also know that no one else has dropped their keys (or something like keys) at the same time.

It's important also to note that many of Mike's examples, like autopsy reports, go from observing an effect to identifying its cause. Random assignment is limited to the situation where, given a specific cause, we want to know its effects. How relevant is it to think about counter instances of causation when you start with a dead body, you cut it open, and you see that the heart is abnormally enlarged? In a case like this, causal inference depends on having an accurate and general theory of the healthy human body. The pathologist is undertaking what we call "pattern matching," matching the pattern she knows the human body should have against the pattern observed and then saying "Well if the body is abnormal in this and this way, what are the possible diseases that could have done this? Maybe there are two or three of them, but of those two or three, what are the differential clues I should look for to rule out one versus another?" For social ameliorative fields that want to change the world in ways that will bring about positive consequences—like education—the interest is more often in going from causes to effects.

Another thing that Mike talked about was the ethics of random assignment, referring to the ethics of withholding the intervention from control groups for some period of time. But what about the ethics of not doing randomized experiments? What about the ethics of having causal information that is in fact based on weaker evidence and is wrong? When this happens, you carry on for years and years with practices that don't work whose warrant lies in studies that are logically weaker than experiments provide. So the argument for randomized experiments being unethical has to be counterweighed against the idea that not doing randomized experiments is unethical, because it perpetuates harmful practices that an experiment would likely have detected.

The last point I want to make is this. Mike refers to the RTC "crusade" in education. When I talk to people at IES, it doesn't seem like they're on a crusade. True, they've done about 40 national evaluations in the last 7 years, and almost all of them have been randomized experiments, many at the school level. Further, the What Works Clearinghouse prioritizes random experiments too. Certain kinds of quasi-experiments do get included but are down-weighted, while other kinds of studies making causal claims don't even get included. You can see here a multi-pronged attempt to institutionalize randomized experiments as the method of choice in education research. But this is not a crusade. A crusade is not as systematic as the IES agenda to influence causal practice in education. A crusade gives me the sense of people who, at the end of the 12th and the beginning of the 13th century, sent out robber barons to bring Christianity to the Muslims while dangling before these Christians the opportunity to get rich quick. There is nothing as systematic here as attempting to institutionalize a process at the apex of educational research from which a significant source of all educational research dollars flows.

There are, of course, costs to putting randomized experiments so high on the agenda. They constitute for me the real downside of the IES agenda. For example, how is the development of substantive theory in education to be promoted? Where is hypothesis generation to come from? Neither is promoted if lots and lots of interventions are conducted that do not explicitly start out to be crucial theory tests, as those in education are not today.

Now, the people I know at IES are absolutely aware of this, and their argument is quite simple. Over the past 30 years, a systematic attempt was made in education research to downgrade quantitative studies, like randomized experiments. In their view, 30 years ago education research got off in a wrong imperialistic direction that systematically privileged a constructivist research style and downplayed "positivist" methods like the experiment. By overemphasizing them now they are trying to set the historical record straight. In 2000 when they came to power, we knew so little about what worked.

Michael Scriven: NSF still has a fight in them, so RCT hasn't won all the fights yet, but they've won the research fight and they're doing their best to take over New Zealand and Australia and other places. So, I'm not too impressed by the idea that this is just "we're going to have a little dictatorship to fix up the bad times in the past."

Tom Cook: Can I just interject here? I think they have not won. In a historical tense this is a skirmish of just 7 years. There is absolutely no indication that the IES agenda will take over all or most of the journals and that assistant professors will have to do this kind of stuff to get tenure. There is no indication that they have yet taken over all of graduate education in methods so that young people are being brought up in a monolithic pro-experimental way. So IES hasn't won and it's not historically institutionalized. Seven years is probably how long it takes for God to blink.

Michael Scriven: Yes, well this is a little bit like saying, "well its true we've only got armies taking over Iraq and most of Afghanistan but otherwise there are plenty of parts of the world we don't own, so we really haven't gotten very far yet." If you happen to be in educational research, the fact that you can't get funding for anything else for 7 years is a pretty serious take home. So, it's true there is some of the world that's left ahead to conquer with RCT, but it's done enough damage as it is. So, when you're making a rational decision about how to investigate something, you have to look very carefully at the fact that the RCT approach is an expensive way to go, sometimes an impossible way to go, and what are you going to be doing about alternative ways to go if you never fund any of them in educational research? And, after all, all that I am arguing for is something quite simple: match the design to the problem.

Now, he also said that he thought that this talk about the other cases that I gave, like autopsy, are cases where you reason from the effect back to the cause. Yes, and let's be clear about this, that's what the RCT people attack. They say "you're not justified in the claims that you've been making about what produced the results that we're seeing." For example in the international poverty alleviation effort, we can see that there is a great reduction in poverty in certain areas where we've poured millions of dollars into Central Africa, it's obvious that that is an improvement, the question is what caused it? And there are a number of possible causes. We can look for the patterns that they exhibit if they were offered it and we can rule them out because they weren't the cause of the cases that we're talking about.

You don't have to cut people's heads off to kill them and it's not best just because it's not quasi-experimental, it's not best because it is not the only way to get where you want to go, which is to reach conclusions beyond reasonable doubt. You can get there six ways from breakfast and RCT is only one of the ways, and that means it shouldn't be getting all of the funding.

Tom Cook: For me the most fundamental criticism of randomized experiments is one we haven't talked about enough yet. Mike said, and I trust we all believe, that questions come first and method choice second. We cannot and should not have methods driving the kinds of questions we ask. If you have a quiver with one arrow only, called the randomized experiment, then that decides the kinds of causal questions one asks. You only ask about manipulable causes, and thus not directly about

gender or race as causes because they cannot be manipulated directly. If you only did randomized experiments you would have a body of causal knowledge that never talked about race except to ask whether the effects were the same for kids who were green, brown, blue, or yellow. You would have no causal theories in which race starts generating the causal process, because race is not manipulable. So, for me the biggest criticism of randomized experiments is that you don't get to ask a lot of questions (a) that are not causal; and when you do ask causal questions, (b) the kinds you ask are restricted to manipulable agents, and (c) do not involve much explanation of a general kind about why the observed results came about.

I have had occasion to be peripherally involved in a Mexican experiment then called Progresa and now renamed Oportunidades. It was a randomized experiment and it's probably saved many lives perhaps because it was a random experiment. So, I am turning Mike's point on its head. The Mexicans in their wisdom decided that they would roll income support, education, and health into one program. Oportunidades gives families a very healthy welfare income grant that sometimes triples income to the poor, contingent on their children staying on in school. There is even a supplement if your daughters stay in school because there's a national problem with girls dropping out too early. The added income is also contingent on the family going for health services on a regular basis and the children taking nutritional supplements. I suspect that the study results were trusted more throughout aide-giving policy circles because they came from a randomized experiment.

Postscripts

Tom Cook: We probably both agree that RCTs are neither necessary nor sufficient for a well-warranted causal inference. But I would contend that most advocates of RCTs believe the same thing. Mike and I may also agree that RCTs are the best single tool for warranting causal inference when such inference is understood as describing the results of a deliberately manipulated possible cause. Our dispute would probably be about how much better RCTs are when compared to other specific tools for claiming causal results. I was heartened by what Mike had to say about much of the past causal research in education being bad. But to get to the bottom of our similarities and differences we would have to talk very specifically about specific causal tools from within a very broad range of both quantitative and qualitative research. We did not do that here. We probably also agree that that the description of what happens when I deliberately manipulate something rarely gets at what is most esteemed in theories of causation in scientific practice or nearly all normative philosophies of science. There the concern is much more with explanations of how or why things happen or with identifying widely applicable and robust effect-generating mechanisms. By themselves, experiments rarely produce such knowledge unless deliberately designed as so-called crucial experiments. That is, they are specifically designed to identify the key operating assumption of some important substantive theory.

We would probably also agree that an agenda focused only on experiments has deep opportunity costs. Many other kinds of important questions and issues do not get probed, ironically including many on which the vigor of experimentation itself ultimately depends. But when we look across the whole panoply of education funders (rather than just IES) we might disagree about how much these other issues are being neglected at this time in history. We would almost certainly disagree, I think, about whether it is today warranted to redress the imbalance of the 30 years when education research was in a rather militant constructivist mode that he and I agree did not serve us well in the long haul.

We would probably agree that the only defensible warrant for causal assertions is that no alternative interpretations are forthcoming from the relevant community of scholars and practitioners concerned with a given subject matter. Specific research tools seem peripheral from this perspective,

even though some of them get us closer than others to the goal of ruling out all alternatives. They do this because of the logic underlying them, provided that this logic is implemented well in research practice. This might explain the preoccupation in this debate with implementing a correct random assignment procedure properly and with minimizing, directly observing, and correcting for treatment-related attrition and treatment crossovers. But the ultimate warrant for causal inference is not entirely tool-based or even truth-or logic-based. It is social. It has to do with acceptance of causal claims after intense critical scrutiny by a wide range of knowledgeable others, including those least likely to like a specific causal claim being made.

I suspect that a major difference between Mike and me is how much each of us would warrant causal claims through deliberately provoking multiple outside voices to be ultra-skeptical or through relying on one's own and one's immediate colleagues' judgment about how well alternative interpretations have been ruled out. I have no illusions about ever knowing I have reached the truth; but I have deep aspirations and even expectations of reaching very close approximations to the truth through a pluralistic, hyper-critical process of social commentary in which the choice of tools like random assignment plays an important but never determining role. I have very little faith in my own efficacy to rule out alternative interpretations. I am always too close to my own evidence and strongly held values. I always need help to be more critical, and in this process my intellectual antagonists are always my best friends. It would be wonderful to be a true believer; but I can never manage it; almost everything is grey to me, albeit different shades of that "color."

Michael Scriven: First, I do want to stress that I don't think Tom is one of the RCT super-enthusiasts, but indeed, as he puts it, a critical user of it. He's correct that my main target is the heavy-handed RCT approach that has taken over much of the research funding and is trying to take over more of it.

And I'm glad he said a little about the randomizing problem; I think of that as a problem that was always controllable, once he had spotted its importance in practice, and is now well controlled. But I don't think that what I call the three pillars of the position are equally well controlled.

On contamination of the control group, he says that it's "present throughout all of our summative evaluations." I think not, since many of them have no control group to be contaminated, for example the success case method studies and the interrupted time series.

On the counterfactual mistake, still endemic in the papers by most of the RCT folk, he thinks this problem is solved by the use of statistical matching of groups rather than individual matching. But the counter example of overdetermination, which brings down the counterfactual analysis, applies equally to groups.

On the third pillar, the certainty issue, Tom says "Not even the most fervent advocates of random assignment believe it is a necessary condition for causation." He needs to look more carefully at—well, I won't say "who he's in bed with" but "who's in the same room": that claim has been explicitly made by the director of IES.

Tom is on the winning team here, as he points out in some detail, and like the ruling class in other contexts, he's not as clear as he might be about life amongst the proletariat. It is the oppressed who notice the excesses, both in the language that he doesn't think anyone utters, and in the essential monopoly of funding that he agrees exists. It's entirely true that there is an ethical issue about using bad non-RCTs to base policy, just as there is in the other direction. But given that he agrees there are other ways to demonstrate causation, it seems clear we ought to be pursuing a middle ground: don't fund bad studies, of whatever kind, and do fund good and efficient use of research resources, whatever the design they use. That's the policy I'm advocating.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding
This dialogue was funded by the Interdisciplinary Ph.D. in Evaluation program at Western Michigan University.

Appendix A
The Present Status of the Randomized Controlled Trials (RCT) as a Standard for Causal Investigations in Human Behavior

The Three Great Pillars of the Position

A. The claim that only the RCT design excludes all alternative causes: This claim fails because it is agreed that it is only true if RCTs are double-blind, and in human investigations, that condition is rarely met. Note A1: In fact, it is arguable only true for triple-blind studies that cast doubt on most drug studies, and most human behavior studies.
B. The claim that only the RCT design supports the counterfactual analysis of causation: This claim is irrelevant because the counterfactual analysis is invalid; it is refuted by all cases of overdetermination, a common situation in medicine and human affairs.
C. The claim that no other design establishes causal claims with comparable certainty: This claim is false, since it is refuted by several other approaches, for example:
    1. Critically filtered observational claims, for example, damage caused by an explosion seen by witnesses; color change in flask from adding litmus solution; making a noise by clapping.
    2. Highly scientific claims in forensic sciences, for example, autopsy reports.
    3. Highly scientific claims in case study research, for example, modified success case method (MSCM; Coryn, Schröter, & Hanssen, 2009).
    4. Many good quasi-experimental designs, for example, interrupted time series, especially with random time of treatment application and duration.
    5. Many theory-based causal claims in science, history, law, for example, that the collision of tectonic plates caused the Sierra Nevada, that a meteorite caused Meteor Crater, AZ and Tycho on the Moon; smoking causes cancer, etc.
    6. Many (other) applications of the general elimination method (GEM).

References

Cook, T. D. (1991). Clarifying the warrant for generalized causal inferences in quasi-experimentation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century* (pp. 115-144). Chicago: National Society for the Study of Education.

Cook, T. D., Habib, F., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's school development program in Prince George's County: A theory-based evaluation. *American Educational Research Journal*, *36*, 543-597.

Cook, T. D., & Hirschfield, P. J. (2008). Comer's School Development Program in Chicago: Effects on involvement with the juvenile justice system from the late elementary through the high school years. *American Educational Research Journal*, *45*, 38-67.

Cook, T. D., Hunt, H. D., & Murphy, R. F. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal*, *37*, 535-597.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724-750.

Cook, T. D., & Wong, V. C. (in press). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique*.

Coryn, C. L. S., Schröter, D. C., & Hanssen, C. E. (2009). Adding a time-series design element to the Success Case Method to improve methodological rigor: An application for non-profit program evaluation. *American Journal of Evaluation*, *30*(1), 80-92.

Donaldson, S. I., Christie, C. A., & Mark, M. M. (2009). (Eds.). *What counts are credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: SAGE.

Fleming, J. E., Cook, T. D., & Stone, C. A. (2002). Interactive influences of perceived social contexts on the reading achievement of urban middle schoolers with learning disabilities. *Learning Disabilities Research & Practice*, *17*, 47-64.

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. *American Economic Review*, *76*, 604-620.

Scriven, M. (1993). Defects of the necessary condition analysis of causation. In E. Sosa & M. Tooley (Eds.), *Causation* (pp. 56-59). Oxford, NY: Oxford University Press.

Scriven, M. (1976). Maximizing the power of causal investigations: The modus operandi method. In G. V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp. 108-118). Beverly Hills, CA: SAGE.

Scriven, M. (2008). A summative evaluation of RCT methodology: & an alternative approach to causal research. *Journal of Multidisciplinary Evaluation*, *5*, 11-24.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Wong, M., Cook, T. D., & Steiner, P. M. (2009). *An evaluation of No Child Left Behind using short interrupted time-series analyses with non-equivalent comparison groups*. Unpublished manuscript, Northwestern University at Evanston, Illinois.