A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research

Michael Scriven
School of Behavioral and Organizational Sciences, Claremont Graduate University

The causal wars are still raging,[1] and the amount of collateral damage is increasing.[2] Are the benefits from this heated debate also increasing? Is the net balance over time positive or negative? It may be useful to attempt a summary of the major points of the present situation in this campaign, although it must be considered as coming from an "embedded" correspondent rather than a neutral observer.

This is an evaluation of a methodology, and such evaluations, common in all methodological literature and in the philosophy of science, are part of a sub-division of the discipline of evaluation sometimes referred to as intradisciplinary evaluation. It is an essential component of any discipline, since it is what justifies the term "discipline"—a field of study that could not distinguish reliably and validly between good and bad research methods could not be identified as a science rather than a pseudo-science.[3]

It should be noted in this introduction that the importance of the issue here lies not just in the key role of causation in evaluation, but also in the key role of causation in what might be called the middle ground of the body of investigative methods, those that lie between the highly theoretical and the highly localized techniques. This includes the key methods for many, perhaps most, practical investigations. It is not accidental that causation shares with evaluation the peculiar distinction of having been attacked as a concept that has no proper place in science. In each case, this was based on a superficial and elitist conception of science, the same elitist conception that argues for technology as an inferior discipline to—or as a mere spin-off from—"real science." In all three cases, the truth of the matter is that the concept or practice under attack was a full-fledged and

[1] The causal wars are about what is to count as scientifically impeccable evidence of a causal connection, usually in the context of the evaluation of interventions into human affairs. The most recent battles are between those arguing that only the use of RCTs should be accepted as providing acceptable evidence (sometimes, the exotic regression discontinuity (RD) design is also allowed). The RCT or *randomly controlled trial*, is an experimental design involving at least two groups of subjects, the control group and the experimental group (a.k.a. study group, or treatment group), between which the subjects are distributed by a strictly random process (i.e., one with no exceptions), and which are not further identified or distinguished by any common factor besides the application of the experimental treatment to the experimental group.

[2] The collateral damage comes from the policy that the RCT camp has been supporting with considerable success, here referred to as "the exclusionary policy," which recommends that no (or almost no) programs be funded whose claims of good effects cannot be supported by RCT-based evidence. This means terminating many demonstrably excellent programs currently saving huge numbers of life-years. It is argued here that the exclusionary policy is not only based on false premises about the merits of (alleged) RCT designs, but even if its premises were true, it does not follow from them.

[3] This is a putative one-sentence refutation of the concept of "value-free science."

vital part of disciplined reflective and practical thought long before science began—in the case of technology, the seniority is provably by at least a million years—and continues in that role today. The attempt to promote RCT to a keystone role in the analysis of causation is, if the arguments below are sound, essentially one more of these examples of academic affectation, and those of a practical turn of mind should treat it with great caution. It is often suggested that the RCT campaign is a revival of the old quantitative vs. qualitative debate, but the suggestion here is that it goes deeper than that, although linked in one respect. Both are partly and perhaps unconsciously driven by the urge to create a new specialty, an esoteric cabal in which the originators have a privileged position as keeper of the keys to knowledge. That's a common accompaniment of the creative urge, but it is a temptation with a downside—the risk that one creates unnecessary barriers for those seeking help, and new idols rather than new appreciation of true values.

Along with the attempt to redefine the concepts of—or at least the acceptable ways to establish—evidence and causation, the RCT campaign also involves the less-remarked parallel effort, going back further, to redefine the concept of an experiment. In standard scientific usage, experiments are just carefully constrained explorations, and the RCT is simply a special case of these. To call the RCT the only "true experiment" is part of an attempt at redefinition that distorts the original and continuing usage, and excludes experiments designed to test many simple hypotheses about—or simple efforts to find out—what happens if we do *this*.

This effort at persuasive redefinition is allied with an implicit denigration of the so-called "quasi-experimental" designs, which are in fact perfectly respectable experiments, only 'quasi' with respect to the one respect in which they have less control over one possible way of excluding one type of alternative explanation. But in other respects, equally important in the practical business of selecting appropriate designs to get definite answers in the given circumstances, they are often massively superior, e.g., with respect to the number of subjects required in order to achieve useful results; the extent to which they avoid intrusion into a natural course of events that it may be very important not to disturb; their cost, not just in money terms but in terms of other important values, etc. Of particular importance, the commonly accepted implication of the "quasi" terminology—that the conclusions from them will be less secure—is, as argued below, categorically false. It is based on an abstract concept of proof or certainty that ignores the practical process and standards used by working scientists and engineers—and by historians and judges in courts of law, and by everyone when acting as real people facing crucial decisions—all of whose approaches are treated with more respect in the present paper.

This review focuses on what might be called a reconsideration of the working credentials of the RCT design, and includes some radical new perspectives on these, with relatively brief coverage of the usual suspects. It presupposes reasonable familiarity with the concepts of experimental design. The key claim or claims in each of the following list of labeled points is italicized.

Summative Propositions

A. *The RCT design is a theoretical construct of considerable interest, but it has essentially zero practical application to the field of human affairs.* It is important to be clear that a true RCT study has to be (at least) double-blind, as are all sound pharmacological studies, whereas the applications in public health, education, social services, law enforcements, etc., that are currently advocated as RCTs are neither double-blind nor even single blind, but 'zero-blind.' Such studies are of course open to the unintended

explanation of their results by appeal to the Hawthorne effect or its converse, since it's usually easy for members of the experimental and control groups to work out which one they are in. *Hence the common argument that the RCT designs being advocated in areas like education, public health, international aid, law enforcement, etc., have the (unique) advantage of "eliminating all spurious explanations" is completely invalid.* It was careless to suppose that randomization of subject allocation would compensate for the failure to blind the subjects (as in single blind studies), let alone the failure to blind the treatment dispensers, a.k.a. service providers (the requirement that distinguishes the double-blind study). The RCT banner in the applied human sciences is in fact being flown over pseudo-RCTs.[4] This failing is not the result of carelessness, but of the almost complete impossibility, at least within the constraints of the usual protocols governing experimentation with human subjects, of arranging for even single blind conditions.[5]

B. *Even the best double-blind drug studies do not have the unique explanatory power claimed by the proponents of RCTs for their zero-blind studies.* These studies, usually thought to be paradigm examples of RCT design, are themselves open to challenge as not meeting the requirements of RCT design as that concept is normally understood. This has led to some demand for what are called "triple blind" studies. That term is variously defined,[6] and has been used to refer to blinding (or excluding via the use of computer-controlled experiments) the statistician analyzing the results, or the pharmacist sometimes used to administer the drugs, or the radiologist or pathologist doing the first stage of interpreting the data.* (check Google or Wikipedia for updated references). Here I will use the term "fully-blind" to cover such cases and more, by defining it as a study in which no-one involved in providing, securing, or receiving the treatment, or analyzing the results, can identify the group membership of any subject until the final decoding step. (It is also desirable, though not here made into a definitional requirement, for that final step to be done under the supervision and control of a named and certifiably independent observer with the skills required to fully understand the complete process.)

The need for fully-blind studies, over and above double-blind studies, arises from the fact that members of both groups receiving treatment in a double-blind experiment usually know that status to be the case, i.e., they know when they are receiving some kind of treatment, even if they don't know whether it is the experimental treatment or a placebo. They frequently know this because they have been told it in order to meet local interpretations of constraints on the use of human subjects; but they may also be able to infer it from other

[4] This would appear to be the correct term, but, to avoid arguments about terminology and in the interests of conciliation, we replace it with "quasi-RCT" in what follows. Note that if one did wish to claim that the zero-blind design is terminologically entitled to be called a true RCT, then one must immediately abandon the claim that RCT designs eliminate spurious explanations.

[5] The word "almost" in this sentence should be noted, and suggests that rather more effort be devoted to identifying and using blinding. After all, it was too quickly assumed in the last decades of the 20th century, that randomization of subject allocation was rarely feasible in social experiments, and we have benefited from reconsideration of that assumption.

[6] With these ongoing discussions and some fugitive documents, the best references are those that are constantly updated, in this case Google and Wikipedia or the specialized and restricted medical or legal research databases. In this article an asterisk is used to indicate that the best references are there.

clues, e.g., variations from normal procedures. This means that differential benefits to the experimental group, if any emerge, may be due *either* to the experimental treatment, *or* to the sum of that effect plus the effect of any interaction of that treatment with the psychological impact of knowing that one is part of an experiment on a new drug. One cannot assume non-interaction here, of course, and only if that assumption were true could one infer that the differences between the two outcomes are due to the experimental treatment on its own.

A simple example of a fully-blind study would be one where the subjects were long-term ward patients in a hospital and the drug or placebo was a tasteless addition to their regular meals; or, if they were all on drip treatment, the placebo and experimental drug, pre-tested for being intravenously asymptomatic, were introduced by injection into the drip bag while the patients were asleep. They would then not know when or if they were receiving treatment. But note that if the drug was very successful its presence would be inferable in the experimental subjects (just as if it had serious side-effects), the blinding would be lost and the experiment would have to be curtailed. Ethics in the fully-blind study might be addressed by asking for volunteers who might *or might not* be used in a study where these conditions would be employed. Since this involves a slight breach of full blinding, a more subtle approach would be desirable.

It's arguable that "RCT" is usually taken to mean a design that meets or is functionally equivalent to a fully-blind design. For this reason, it seems fair to conclude that the designs commonly described as RCTs in the human affairs domains are not just two but three design stages removed from true RCTs. In any case, however you choose to use the term "RCT," the zero-blind designs supported by the RCT protagonists for use in the human services area are obviously incapable of eliminating all spurious explanations, and the idea that we should treat them as the only source of good scientific evidence for causal claims or sound evaluations of interventions is roughly the equivalent of restricting our national entries in the Olympic marathon to runners with one leg.

C. The difficulty of divorcing the "intrinsic" effects of a treatment from the psychological effects of *giving* the treatment has two debilitating results for causal inferences from experimental designs. The first we have just covered—the problem that the distinctive effects seen in the experimental group are possibly due to the combined effects of two factors, not just to treatment effects. The second problem is that one cannot tell what the effect of the experimental treatment will be if administered outside the experimental context because you can't tell how much that context is adding to the effects. *Hence generalizing from even a true RCT to real world use of the experimental treatment tested in the RCT is hazardous, and any true RCT study has to be supplemented with extensive high quality field reports on real world use.* This is one reason one can't speed up drug testing without substantial risk, particularly because much of the normal process of gathering the real world test data is left in the already over-occupied hands of the general practitioner or specialist. What does this mean for research based on quasi-RCTs?

Essentially, it adds a fourth ravine of disconnect between the quasi-RCT design and any justified practical use, and, most notably, since we do not have a huge cadre of trained scientists like the groups acting as field reporters in the medical field, it means that this fourth gap can only be filled in by seriously funded studies using expert field researchers who will of course not be using RCTs. But that fourth

gap is not mentioned or staffed in any of the large RCTs I have seen funded under the new exclusionary funding policy, and would face difficulties in getting funding under the current practices since it is not an RCT design.[7]

So the irony of the present situation is that the value of quasi-RCTs, if it is substantial, cannot be established via the usual arguments for the superiority of the design; indeed, we now see that it cannot be established via a simple empirical demonstration without depending on non-RCT researchers. Since the RCT group, now in power in many domains, neither recognizes this nor is willing to support non-RCT research, it has cut itself off from its only remaining avenue to legitimacy.

D. But there is more bad news ahead. The other logical advantage of the RCT design that is claimed by its protagonists is that it supports what is said to be the key logical property of causes, the fact that they support counterfactuals (i.e., a cause is something without which the effect would not have occurred). *However, this counterfactual-supporting property is certainly not a logical property of causes, as stated, and even more certainly is not a logical property possessed by the quasi-RCTs being supported by RCT protagonists.* It's not a logical property of causes, as it stands, because of the common phenomenon of overdetermination, i.e., situations where an effect E is caused by event C, but E would have occurred even if C had not occurred, because of circumstance $\bar{C}$ which is "lurking in the background," ready to do the deed if C does not do so. For example, when an outfielder—let's say his name is Jaime Cortez—makes an easy catch of a fly ball in a professional baseball game (a causal claim), the fact that a second outfielder got into position behind him and would have made the catch if Cortez had missed it, does not lead us to say Cortez did not in fact make the catch, although it is clear the counterfactual does not hold.[8]

The counterfactual claim has caught on throughout a large slice of the social sciences and has a certain amount of appeal to the group of professionals and managers that is not literate in the specialized field of the logic of causation. The pitch goes like this: "You want to support interventions that *really matter*—in other words, ones about which you could say, if they hadn't been done, then we would not have gotten these results—well, the RCT is the way to go, and it's the only way to go." Not true, not even of real RCTs, and not even half-true of quasi-RCTs. Cause is an even more complex concept than counterfactual and it implies both more and less than the latter notion.[9]

[7] One frequently hears protests that there is no restriction of funding to RCTs. It's true that there are some explicit references to the possibility of alternative approaches being funded in the official releases, but all the many reports that seem to have surfaced tell a different story of the actual practice in the review panels, in fact a story of huge funding for poorly trained RCT groups over really strong alternative approaches requesting small amounts of support. If anyone was interested in refuting these stories it would be easy enough to do so via a small meta-study, but those controlling the funding have shown no interest in doing this. Power corrupts even scientists.

[8] A common defense against such counter-examples is to say that E means 'E at the time it did occur.' But it's easy enough to construct a case where $\bar{C}$ produces E at the same time that E does, and such cases have been well known in the literature on the logic of cause for more than forty years, although poorly understood in the social science discussions. For example, in *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (Morgan & Winship, 2007), there are no references to overdetermination (or redundant causation, a related term), by contrast with the 25 in the excellent philosophical anthology *Causation and Counterfactuals* (Collins, Hall, & Paul, eds., 2004), or see my "Causes, Connections, and Conditions in History" (1966) in various anthologies.* (check Google or Wikipedia for updated references).

[9] No attempt is made here to give anything like a full analysis of causation, only of methods for identifying it.

E. *The threats of confounding variables to RCT designs are extremely serious and numerous, and costly to handle, and require continual highly skilled attention that is often not budgeted or staffed in RCT studies.* Now we'll look at some better-known problems with RCTs, although we'll bring out some aspects of these considerations that are less well understood. The aim is to show that the only feasible RCT-type designs usable in the human affairs fields are, like most of the alternatives, just one risky and tricky design to consider for causal investigations. We'll begin with some of the more obvious difficulties.

In the definition given earlier of an (ideal) RCT, we included the condition that the experimental and control groups be distinguished only by the experimental treatment of the former, since any other distinguishing factor would be a potential contaminant, a.k.a. spurious cause. There are some quite deep difficulties with real world attempts to meet this condition that we have not so far addressed. We need to stress the problem that getting volunteers of the required type for our studies will often mean that they are—increasingly, as experience now indicates—not typical of the group to which we wish to generalize. To this we add the fact that the need for *two large* groups disadvantages the RCT design from the start for both availability and cost reasons, keeping in mind that some excellent alternative designs require only one group, e.g., the interrupted time series design, and others can use smaller groups with comparable power, e.g., variants of case study approaches.

Taking another example from the preliminary planning phase, we often want there to be a number of villages (or clinics or wards or classrooms) in each group, randomly allocated as to which group they are in. Now, any two such groups will always be distinguished by *some* factors, e.g., location, or they would not be different groups. And these unavoidable distinguishing factors may be linked in an unexpected way to causally relevant differentiating factors such as local variations in weather (if the average physical displacement is substantial), or (if it's minor) room temperature, or ambient noise level, or facilities management style, which then invalidate the inference to the experimental treatment as being the only possible cause of any outcome differences. The standard ways of handling these problems in the design phase are via stratification and/or large numbers, since by using large enough groups, and/or forced sorting on at least the leading known potential causal variables, we hope to have enough entities in each group to constitute an equivalent range of location, management, etc., and hence to "balance these factors out" of the relevant calculations. But to attain high confidence levels in this way requires paying the price of large size, already mentioned. And, probably more importantly, to reduce the risk of complete failure, we still need to keep a continuing careful and highly trained eye on all groups to make sure no new factors turn out to be active that we have not stratified for, or that are not covered by our numbers. Call this first group of potentially contaminating factors—the first set of reminders that we are in the real world of experiments—the Delta Group.

There is another group of such factors that are commonly recognized—e.g., by Tom Cook, who regards them as so serious as to constitute in themselves good reasons to abjure any use of the term "gold standard" for RCTs—which we'll call the Epsilon Group. These factors can only surface *during the course of running* an RCT,

Its complexity has led many analysts, e.g., Nancy Cartwright (2007) in *Hunting Causes and Using Them: Approaches to Philosophy and Economics*, to argue that "there is a great variety of different kinds of causes…" I think this may be taken to suggests that "cause" is ambiguous, which I think is not true; I have suggested as an alternative that the reason for the variety is that a cause is (roughly) any single-factor empirical explanation, and there are many different ways to explain something.

rather than from the beginning, as with the Delta Group. Two of them are sufficiently well-known to excuse further exposition here: cross-contamination, a.k.a. leakage (usually of the experimental treatment over to the control group, but the reverse is possible), and differential attrition (usually of the control group members), which can take us below the shared number of post-tests that we need for reasonable confidence in the results.

The Delta and Epsilon confounders are especially serious threats in exactly the cases where RCTs are often thought to be best suited for application, i.e., in dealing with slow-acting, small-size effects. These factors have four important properties: they are potentially fatal flaws in an RCT, they have often ruined very expensive RCTs designed and run by very well-trained researchers, their effects can virtually never be factored out ex post facto, and they require specially trained observers almost constantly watching for their emergence, continuing presence, and magnitude—observers who have to be empowered to act quickly to stem the swift hemorrhaging of validity.[10] It is ironic that the skills and responsibilities required of such absolutely essential observers are most often to be found in the repertoire of a few qualitative researchers (although by no means the norm amongst them) and are often ignored or absent or inadequately provided for in the payroll of many of the quantitative researchers that are normally funded to do RCTs. In any case, the Delta/Epsilon factors raise the cost, entry, and maintenance requirements for the real-world quasi-RCTs very high. This makes that approach hard to justify except in special circumstances where, like half a dozen other designs in *their* special circumstances, they are a better choice than their real-world alternatives.

F. The discussion of the previous sections appears to show that: (i) the frequently claimed logical credentials of RCTs do not support the real world quasi-RCTs championed by the RCT camp; and that: (ii) those real world designs have their own particular and serious limitations. But they do retain the random allocation procedure. So we still need to inquire whether this leaves them with some residual general superiority. Or does the existence of the four divides listed earlier—the gaps between them and the ideal RCT—leave them without any special ability to exclude any other cause besides the experimental treatment? Basically, the answer is both yes—in one sense (of "special")—and no, in another sense; and the two senses cancel each other out, leaving the quasi-RCT with no net advantage.

The arguments of D show that the quasi-RCT design must include provision for protection against many factors that have to be actively prevented from becoming alternative causes, in both the design phase and in the monitoring (and intervention) phase. This is exactly the same situation, from one point of view, as the situation when using any quasi-experimental design—although of course the particular threats vary from design to design. The randomization is still a special virtue in that it does break some causal chains from spurious causes, but its four removes from the fully-blind design mean that randomization does not remove a large family of significant spurious causes from consideration.[11] Now many other

[10] These observers must be both present regularly and able to act quickly, because even if spotted immediately, these factors have to be controlled very quickly in order to save the experiment.

[11] For readers who are dubious about the claim that Hawthorne-type effects are significant threats, it may be worth recalling the interesting experiment done in the early days of placebo studies, that showed the placebo effect works just fine *even if* the control group is told they are getting the placebo, and are instructed and tested on their knowledge of exactly what this means. (The exact reference for this study is now hard to trace,* (check Google or Wikipedia for updated references) but in the

designs have analogous "special strengths" corresponding to the remaining virtues of the quasi-RCT, with no more than its share of weaknesses. For example, in the interrupted times series (ITS) design, the use of individual subjects as their own controls is a notable and powerful feature. It avoids some of the spurious causes that are still serious threats in the quasi-RCT design, such as those to which it is vulnerable because of differential attrition and imperfect matching, or those emerging environmental factors that differentially affect the control group because of, for example, its different location. And it greatly reduces the required sample size. The ITS strengths are thus special in their own way, i.e., superior, to almost all other designs in these respects, just as randomization is still a special strength of the RCT design. On the other hand, ITS is subject to the grave limitation that it will only work with ephemeral effects, due to threat from the spurious explanations of learning and habituation. *The bottom line is that when we total up the strengths and weaknesses of any of the half-dozen leading designs we get the same situation, namely each has substantial entries in both columns: there is no clear edge for quasi-RCTs that holds across all cases. So, although the real-world RCT still has its own unique strength, that strength is no stronger than the strengths of the good alternatives like the IST, on their own home ground.* Now that is not an obituary of the quasi-RCT, but it should be an obituary of its public persona, as recently represented.

However, a good evaluation, summative or formative, does not restrict itself to looking at the limitations of the evaluand (the entity being evaluated), and we will now go on to consider several alternative entries in the race meeting at which we award the Gold Standard Cups. These include a gold standard—actually a higher standard than a gold standard, perhaps a platinum standard—for causal *claims*; a gold standard for causal *experimental designs*; one for *funding evaluations* (and other research involving causal claims); one for *funding programs*; and, along the way, one for *good evidence*. If these awards are well-justified, we will then have outlined a fairly complete alternative system for thinking about evidence and causation; and perhaps then we can say, "The king is dead, long live the king." Let's start with the most shocking of the alternatives.

G. *The real "gold standard" for causal claims is the same ultimate standard as for all scientific claims; it is critical observation.* Causation can be directly observed, in lab or home or field, usually as one of many contextually embedded observations, such as lead being melted by heating a crucible, eggs being fried in a pan, or a hawk taking a pigeon. And causation can also be inferred from non-causal direct observations with no experimentation, as by the forensic pathologist performing an autopsy to determine the cause of death.

A lingering effect of an antique epistemology is the still common belief amongst scientists that the examples just given of the direct observation of causation are "really" cases of inference to a causal conclusion. This item of philosophical mythology goes back to Hume and earlier, though it is thought to be further reinforced by recent neurophysiology. But this is a confusion of brain activity with thought processes; the latter depend on the former but are not identical to them. The infant's brain develops visual concepts—and probably inherits some as well—and in that learning process often infers to their presence, but the adult sees the world *through* those concepts, not by recapitulating the inferential process.

Thus the adult *sees* his friend Pierre, the person he has come to meet, in the crowd coming off the plane at the airport; he does not see a string of shapes, compare each with the stored templates in his memory, and hence *infer* that his friend is in front of him. In the same

mid-20th century it was widely accepted as reputable, and I have not been able to find a refutation.)

way, while as an infant it may make sense to say that he had to infer that he has knocked a cup off the table onto the floor, it is correct to say that as an adult, he *sees* that his infant son has just done that. This is seeing C cause E, seeing causation, not inferring it. No doubt the extended brain is still doing quite a bit of processing, probably some of it farmed out to the optic nerve, but the mind is not; its survival skills have taught it when to go into shortcut mode, and it has long since reduced what used to be an inference to a perception.

Similarly, the adult is called on in court to bear witness to what he *saw*, including bearing witness to having seen the defendant strike the victim several times, a causal claim. The courts speak carefully about the difference between what can be seen and what is inferred, pushing hard for the assumptions involved in the latter case, and rightly allow that perception can include the direct, non-inferential, perception of causation. Matters of life and death hinge on the court's critical appraisal of what can and cannot be determined by observation, and that is surely as high a standard as any that we need for scientific evidence. What is true for visual perception is no less true for the best examples of other modes of sensory perception: for example, experienced drivers are all rightly certain that on many occasions they have slowed their car or cycle by applying the brakes to it. The whole of science rests on the ultimate testability of critically appraised observational claims, so I conclude that since these can include causal claims, we can be sure that we have attained the gold standard for them. Since critically appraised observation provides the best evidence for, and indeed the best kind of causal claims, it is time to examine the implications of this position for experimental design.

H. The first implication is that *the professionally performed case study, since it is often suffused with causal claims based on observation, is immediately reinstated as a live candidate for respectable demonstration of causation.* Let us be clear here, as in subsequent discussion, that we are referring only to the best examples of this genre. In the past a great deal of hopelessly unscientific work has been put forward as "qualitative methodology," including many anecdotal reports described as case studies, and it is not being reinstated by the present assertion. We are simply going to avoid guilt by association, and allow that there are case studies in the critical tradition of good scientific work, where what is reported, and checked, includes causal claims. They will be found in most cultural anthropology, in biological and sociological field work, in epidemiology, planetology, cosmogony, and geology, in many engineering studies of structural or machine failure, in the best clinical psychology and medicine, and very often outside science in the best disciplinary traditions of history and the law.

Someone may protest that case studies are aimed to establish singular facts, not the general propositions that science is concerned to establish. But that view of science is simply another myth: much of science—for example, a large part of the divisions of science listed above—is mainly concerned to establish singular facts. Even if that were not so, what is definitely so is that most of program evaluation, and most other branches of evaluation such as personnel evaluation, *is* primarily concerned to establish singular facts, namely the merit, worth, and/or significance (m/w/s) of a particular program. Even if, contrary to fact, that were not the business of science, then all that would follow would be that evaluation is not a science, but is rather a discipline like history or the law, also primarily concerned with the particular—with what are called idiographic phenomena rather than nomothetic ones—and it would be

none the less respectable for that. So the restriction of case studies to the particular is not a drawback for much of evaluation.

Case study research is not restricted to single cases, as Yin (2002) demonstrates. In fact, it has been usefully extended in what Robert Brinkerhoff calls "The Success Case Method" in his book by the same title (2003).

Now, case study approaches do become expensive when dealing with large studies, a drawback that has to be weighed against their great advantages in turning up fine details that lead to the next stage of research, something not so common when dealing with large numbers of cases on each of which very limited data is gathered. But this does not mean that we need RCTs for big general studies or for studies that depend on statistical analysis for their conclusions. There are many alternatives, ranging from the simpler 'quasi-experimental' designs to the vast net of studies that conclusively identified smoking as a major cause of lung cancer in humans without including a single RCT with human subjects. So RCTs, to use that term loosely—and certainly quasi-RCTs—have no categorical advantage for either large or small investigations.

I. *The key point for the present stage in this discussion is that the suggestion that we need RCTs to establish causation in evaluation is as far-fetched as the suggestion that we need them to establish all causal claims in history, such as the claim that the Iraq war caused the death of many US citizens; or in order to establish the guilt of every defendant accused of speeding (i.e., causing a car to break the speed limit) or of causing grievous bodily harm, i.e., assault.* Almost all of the causal claims made in the real world that are beyond reasonable doubt are based on observation or direct inference from observation, whether in the context of scientific lab, clinic, or field work, or in the practice of the law or history, or everyday affairs or journalism; and these can be assembled and analyzed statistically with either the microscope of case study approaches or the lighter touch of large studies.

Once again, we see how important it is to keep a firm grip on our common sense in the heat of the current dispute about the role of RCTs in establishing causation. The suggestion that they are needed to support the evaluation of international aid programs, for example, is about as sensible as the suggestion that we should not believe any claims about deaths due to the wars in Darfur or in Iraq unless we have RCT-based evidence. It is no harder to establish that a program that gives food or shelter to those who need it immediately has beneficial consequences than it is to establish that the enemy's mortar shells killed fifty soldiers this week.

J. Of course, there are programs, in international aid and elsewhere, for which it's not so easy to establish that they are producing benefits, or just how large those benefits are. As previously mentioned, these are likely to be programs where the effects are smaller, less immediate, and less obvious, especially programs where the net effects are simply not observable in each single case. For some but not all of these, an RCT design may be appropriate. For others—probably the majority but no-one has done the counting—we do better with other designs. It is now time to ask if there is some underlying methodology that can be used, or that logically underpins, all legitimate causal claims, since it clearly is not the RCT design. It will not only guide us in choosing more specific designs, but will serve us in many cases where no more specific design is needed and those where none of them work.

All approaches to substantiating causal claims, RCTs and case studies included, do indeed share the same basic logic, an underlying logic of all causal claims. To make it terminologically secure that we have something here that trumps the alleged gold standard, we use a name for it whose acronym supports that conclusion. It is the General Elimination Methodology or GEM approach, and its origin lies in the skills of every expert practice. General Elimination Methodology is the basis for all causal claims, and the best approach to use when direct observation, critically appraised, is not enough. In fact, it is the meta-program for the neural mechanism underlying the observation of causal connections, and it is the underlying logic of RCTs and all quasi-experimental approaches as well.

It is based on one general premise and, unsurprisingly, two premises summarising practical knowledge, that being the home domain of causation and well below the usual level of scientific theories. In outline, it looks like this:

i. The general premise is the deterministic principle: all macro events (or conditions, etc.) have a cause. This is only false at the micro-level, where the uncertainty principle applies, but the latter principle has essentially no detectable effect on the truth of macro-determinism (though it is easy enough to deliberately create bizarre experiments where it does).
ii. The first 'premise from practice' is the list of possible causes (LOPC) of events of the type in which we are interested, e.g., learning gains, reduction of poverty, extension of life for AIDS patients. We have used LOPCs for more than a million years, in tracking and cooking and healing and repairing, and today every detective knows the list for murder just as every competent mechanic knows the list for a big-end rattle or a brake failure, though the knowledge is as often tacit as explicit, outside the classroom and the maintenance videos. An LOPC usually refers to causes at a certain temporal or spatial remove from the effect, and at a certain level of conceptualization, and will vary depending on these parameters; of course, the context of the investigation determines the appropriate distance parameters. The distant LOPC for murder is the list of possible motives; a more proximate one, developed in a particular case by applying the general one, is the list of suspects. When dealing with new effects, we may not be certain the list is complete, but we work with the list we have and extend it when necessary.
iii. The second practical premise is the list of the modus operandi for each of the possible causes (the MOL). Each cause has a set of footprints, a short one if it's a proximate cause, a long one if it's a remote cause, but in general the MO is a sequence of intermediate or concurrent events or a set of conditions, or a chain of events, that has to be present when the cause is effective. There's often a rubric for this; for example, in criminal (and most other) investigations into human agency, we use the rubric of means/motives/opportunity to get from the motives to the list of "suspects." The MOL is the magnifying lens that fleshes out the candidate causes from the LOPC so that we can start fitting them to the case or rejecting them, for which we use the next premise.
iv. The fourth premise comprises the "facts of the case," and these are now assembled selectively, by looking for the presence or absence of factors listed in the MOs of each of the LOPCs. Only those causes are (eventually) left

> standing whose MOs are completely present. Ideally, there will be just one of these, but sometimes more than one, which are then co-causes. (Note that there is no reference to counterfactuals.)

Note also that the GEM works equally well with the determination of general or particular causal claims, i.e., with both causes of classes of effects (or with the effects in a large class of subjects), or with the causes of a particular effect on one occasion, or with one subject. So the GEM competes directly with the general methodology that is facilitated by the use of an experimental design such as RCT or IST, which is often unnecessary, and frequently impossible. For example, in determining the cause of a geological formation such as the Rocky Mountains, we are not in a position, for practical reasons, to do an RCT; in other cases, such as in evaluating most international aid, we could do it, but it would be unethical to do so. In both these types of case, the GEM works straightforwardly.

To take an example from work in which I have been involved, when looking at the effect of aid given by Heifer or Gates to extremely poor farmers in East Africa, after determining that a substantial improvement in welfare has followed the arrival of aid, and has been sustained for a few years, we check for the presence of more than a dozen other possible causes of this observed subsequent increase in welfare, including: efforts by the country's government that have actually trickled down to the village level, analogous efforts by other philanthropies, self-help gains resulting from inspired leadership in the local communities, increased income from family members traveling to well-paid job openings elsewhere and remitting money back home, increased prices for milk or calves in the local markets, the beneficial results of a few years of good weather or of improved water supply, or of technology-driven improvements in the quality of available commercial feed, veterinary meds or services, or grass seed for improving pastures. This requires considerable systematic effort, but no sophisticated experimental design, no sophisticated statistics or risk analysis. The GEM approach here is essentially an extension of common sense. Michael Quinn Patton (2008) provides another excellent example of the use of non-RCT methods in establishing causation in his article in this issue of JMDE.

In the more complicated case of establishing that heavy smoking is a common cause of lung cancer, there were no RCTs involved—because of the ethical constraints—but the scientific process of GEM proceeded without any resulting problem with our confidence in the finding. If we can reach the high level of confidence we have about smoking as a cause of cancer without placing any weight on an RCT, it is clear we have no need to believe that the absence of RCT evidence weakens the case for conclusions about causation—provided that the GEM approach is rigorously applied. And remember all the examples from geology, epidemiology, etc., where we also do not use RCT. Roughly speaking, there are about nine or ten ways to go about establishing causation *beyond reasonable doubt*, all of them relying on GEM for underpinning, and only one of them involving even a pale shadow of RCT. They are: (i) direct critical observation, e.g., visual, affective, tactile; (ii) reported (and validated) observation, e.g., case studies; (iii) direct or simple inductive inference from (i) or (ii), e.g., to the effects of meteorites on the far side of the moon's surface, prior to satellite launching, or the famous inference to the effect of gravity on light rays from the observations of the 1919 eclipse; (iv) simple GEM inference, e.g., autopsy, engineering breakdown, the international aid examples; (v) theoretical inference, based on use of an analogy or theory, e.g., geology, cosmogony; (vi) direct manipulation e.g., in the kitchen and lab; (vii) "natural experiments," e.g., meteorology, epidemiology; (viii) 'quasi-experimentation,' e.g.,

pre/post with comparison group in pedagogy, addiction studies, international aid; and (ix) quasi-RCTs, e.g., pharmacology. The tenth candidate is inference from cross-sectional data, and it needs a little more analysis than we have space for here.

In sum, there is absolutely nothing imperative, and nothing in general superior, about the need for RCT designs, let alone the weak cousin of them that is all we are being offered in the areas currently being invaded by the demand that nothing less be accepted.

K. A last nail in the coffin of the RCT cause, in the causal wars, comes from a check on authenticity. The ultimate test of authenticity is self-application when relevant. For example, it is often relevant for evaluators to have their own work evaluated, and a test of their authenticity is to see how often they do this, and ensure that it is done with the care that they call for in their own calls on others to have their programs evaluated. Now, if the RCT cause is legitimate, would it not be good practice for those now enforcing the exclusionary implementation of it to check whether their policy is working? In other words, to evaluate their own policy. If they do not do this, may one not conclude that they are failing a crucial test of their own doctrine?

There is so far no sign of any such study, and that suggests two further conclusions. First, it is about time that one of the supervisory agencies—the General Accountability Office, or one of the Inspector-Generals' offices—checked out this situation, in which hundreds of millions of tax dollars are being invested. Second, and perhaps more significantly, one cannot refrain from speculating about the research design that would be appropriate for such a study. It seems clear that one could hardly do it by using an RCT design, not only because of circularity, but since that would involve withholding funding from half of a group of equally deserving applicants for a period of up to several years, a prima facie unethical action.[12] This does prove that the RCT model cannot be universally ideal for investigating causal claims, and a moment's thought will suggest that there will be many other programs of a similar kind—that is, tests of approaches to the investigation of causal claims—that also could not be investigated using RCT, for the same reason. In other words, the claim that the RCT is a fundamental necessity for causal investigations is self-refuting: it not only cannot be applied to itself, but the exercise of contemplating doing that uncovers at least one family of important studies it cannot cope with. So the RCT empire-builders do not meet their own standards.

With all of the preceding discussion in mind, it is time to spell out what can reasonably be required in a funding policy aimed to upgrade the standards of evidence for causal claims, especially those on which evaluations often depend. We have covered the arguments on this thoroughly enough to suggest an appropriate type of policy; it only remains to formulate it. While there are good reasons for upgrading the standards of quality in causal research in the human services areas, there is absolutely no basis in logic or experience for doing this by restricting the funding of proposals or programs or research investigations to those using any one model for doing such investigations. *The optimal procedure is simply to require very high standards in matching designs from the wide range that can yield high confidence levels, to the problem and the resources available.*

L. In conclusion, let me stress my full understanding that the point of view

[12] Unethical not just because it would be unfair to the applicants, but because it would deprive the equally deserving subjects of those applicants from the benefits of evaluation of and hence possibly support for the programs which are serving them.

presented here refers to a complex and difficult issue, and its arguments are almost certainly just as guilty of oversimplification as are, it claims, the distinguished scholars that are its target. It may indeed contain more serious errors than their position. It has been circulated prior to publication in the hope that these will be pointed out so that the paper can be corrected or, if appropriate, withdrawn; and, now that it is eventually published, it is in the hope that its remaining errors will be widely used as lessons learned for discussions that follow. This issue is not a mere academic dispute, and should be treated as one involving the welfare of very many people, not just the egos of a few.

References

Brinkerhoff, R. (2003). *The success case method; Find out quickly what's working and what's not.* San Francisco, CA: Berrett-Koehler.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics.* Cambridge, UK: University Press.

Collins, N. H., & Paul, L. (Eds.). (2004). *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* Cambridge, England: Cambridge University Press.

Patton, M. Q. (2008). Advocacy impact evaluation. *Journal of MultiDisciplinary Evaluation*, *5*(9), 1-10.

Scriven, M. (1966). Defects of the necessary condition analysis of causation, *Philosophical Analysis and History.* In W. Dray (Ed.) Reprint in: Sosa, E. & Tooley, M. (Eds.). *Causation.* Oxford: Oxford University Press.

Yin, R. K. (2002). *Case study research: Design and methods* (3rd ed.). Newberry Park, CA: Sage.