

CAN WE INFER CAUSATION FROM CROSS-SECTIONAL DATA?¹

Michael Scriven
The Evaluation Center
Western Michigan University

THE STANDARD SCIENTIFIC ANSWER

One answer to the question in the title is: Certainly not if the only scientific basis for causal claims is the results of experiments, in particular experiments with random allocation of treatment to the control and experimental groups (RCTs). Equally certainly, however, causal inferences from cross-sectional data are made all the time in science, using the cross sectional views provided by magnetic resonance scans to infer conclusions about brain tumors, or using observations from the cross sections cut by the scalpel and saws of the forensic pathologist performing an autopsy, to infer the cause of death. Indeed, there are simpler examples: the biologist looking at the cross section of an ancient water cypress can infer from the rings to the cause of certain irregularities, e.g., to the occurrence of a big drought in 1720 which almost sucked the swamps dry. My task here is to look with some care at such examples and see whether we can learn something from them about what can and can't be done with School-level State Assessment Score Database (SSASD), either in its present form or in some modestly revised form.

To begin with, these examples suggest there's no need to rush to the conclusion that the states should abandon their present data-gathering on the grounds that it cannot in principle answer their questions--or the national questions--about the effectiveness of major interventions on the schools. But the real question, the primary question, once we get past the oversimplified remarks about 'the only scientific way to establish causation,' as the director of the Institute for Educational Science often states, is not whether the cross sectional (XS) data can *in principle* establish causation, which it certainly can, but whether the *present* kind of data is good enough to let us answer the *presently* important questions about the *current* crop of interventions.

¹ Particular thanks to David Sweet and Tom Cook for valuable discussions that improved my thinking on this topic.

Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

Beyond that primary question there is a secondary question, which is whether, if the answer to the primary question is No, we can suggest some moderate changes that would alter that answer without breaking the state and/or federal banks. Those two questions are much more difficult to answer than the question of what's possible in principle.

To answer them, it's useful to go back to the examples just given from standard scientific practice, and ask just how the causal inference is possible from those XS datasets. The basic logic is not very complex, but has to be rather carefully formulated. Let's take the case of the coroner looking at the skull of someone with massive head injuries. The single entry wound is circular and obviously deep, which means that something spike-shaped or spear-shaped could have been driven in with great force, or a bullet might have had a similar effect. There are no powder burns on the scalp around the entry wound, so one way of pinning the cause down to a gun is not there to help us. But that doesn't rule the gun out, since the powder burns don't occur if the gun was fired from a distance of more than 15 feet. Still, the microfractures around the entry wound indicate that the velocity of the alien object was very high, too high for an arrow, spear or swung spike to achieve. So the bullet looks like a good candidate. Of course, if the bullet is still in the skull, that solution is fully supported. But even if the bullet has passed on its way, if the exit wound also shows the star pattern of micro-fractures, we can rule out the slower projectiles completely, since they could not possibly be still carrying that kind of velocity after traversing a skull. With some care, we can probably give a modest range of calibers for the fatal bullet; and we can certainly attest that the injury was enough to be fatal. Of course, we need to scan the rest of the body for other possible causes of death, but, finding none, we have clearly identified the cause of death from cross-sectional data.

The underlying logic here is simple enough. We see evidence—the entry wound—that immediately points to a small set of possible causal agents—the projectiles. Let's call them A, B, C,..., the possible causes. Now each of these has certain associated patterns of operation, their *modus operandi*. We now begin scrutinizing the cross-sectional data in order to find factors that eliminate some, and/or characterize others, in the set. Assuming something for the moment that we can easily check later--that there are no other possible causes of death present--and assuming that some cause had to be present (the assumption of determinism, well-established outside the quantum realm) we find that only A is consistent with the pattern of evidence here. This inference to a causal conclusion is based on what we can call elimination analysis, and it requires only two things: (i) a list of possible causes consistent with the first features we notice about the situation, in this case the entry wound; and (ii) a list of the features that identify the *modus operandi* of each possible cause. Can we match this kind of inference in the schools?

The answer is that *we can in some cases*—notably in cases of traumatic impacts like the Columbine massacre at the single school level, and the Katrina hurricane in the case of regions--but in most cases it would require some auxiliary information. For the relatively slow-acting interventions currently of interest in education and many health and environmental areas, that means the answer to the primary question is, No, but the answer to the secondary question

is Yes. The further data we need is not vastly expensive to get, but it does take trained personnel to get it: what we need is systematic investigation of the presence and operation of alternative causes for any changes from the normal patterns that we find.

Now that reference to ‘normal patterns’ reminds us of a very helpful feature of the state data: we are not dealing with a single cross-section here, as with the autopsy; we have section after section, stretching over many years, as with the tree rings (and, spatially, with the MRI). Many of the ‘normal patterns’ are at least partly temporal patterns. In that kind of time-spread dataset, making causal inferences from the dataset alone, with no auxiliary information, is much, much easier, in many cases, although still not always possible.

A simple example would be as follows. Many school districts are dismissing the principals, and sometimes the whole staff, of schools identified as ‘failing’ under NCLB analysis. We can locate such schools in most state datasets, and then track their scores on the local tests across the transition to a new administrator and staff. If we do find that they are consistently recovering to the same level of performance as comparable schools, we have thereby identified some benefits probably produced by NCLB. Of course, if the n was small, one would have to check on the ground for the presence of alternative explanations, but if the pattern is common and consistent, we can be confident that the staff changes are causally efficacious. Of course, this is not to say that there are no intermediate causes such as increased parental effort, but these just provide an amplification, not a refutation, of the causal claim.

Looking back at our examples from standard scientific practice, it’s worth noting that the tree section, apparently the simplest example, actually provides a long sequence of cross-sectional data such as we have in SSASD, whereas an MRI scan, despite consisting of multiple plates, only gives us the temporal equivalent of one tree ring or one year’s data from a school. By sectioning a number of trees in a community, and annotating their locations with care, so that we pick up variations in shade exposure, water run-off channels, soil composition, and weather lee protection, we can explain many variations in the ring patterns, thus narrowing down the explanation of a common feature to some external intervention such as a flood or drought. And of course, we look for ways to triangulate such conclusions from geological traces or human chronicles, if there are such. The end result is complete confidence, in plenty of cases, that we have demonstrated the effects of a long drought or a great flood at a particular time in history.

THE CASE OF SSASD SCHOOL DATA

A number of people here are inclined to the view that causation can only be established by controlled experiments with random allocation of subjects between the experimental and control groups. For example, Don McLaughlin says:

“Randomized assignment of schools to treatment and control conditions is an essential component of studies which would estimate program impact.”

The examples considered above suggest this is too restrictive a view. And there is a much older and more extensive body of knowledge which supports the existence of alternative ways to skin the causal cat. More than two thousand years ago, the discipline of history began to emerge as a serious subject in the West, one in which causal explanations are common, and carefully documented, without any possibility of using control groups. The recent publication of the four volume set, *Historical Methods in the Social Sciences*, (Sage, 2005) reminds us of the long and serious discussion of the methodology of history, in particular, the logic of causation in history. It is perhaps rather unfortunate that little training in historical methodology is included in social science research methods courses, since studies of social interventions such as NCLB are paradigmatically historical studies. The result has been that social scientists are often too easily impressed with physics and chemistry, the great non-historical sciences where experimentation is available, as the ideal embodiments of causal inquiry. Even without looking at historical methodology, if there was more emphasis on epidemiology, forensic pathology, geology, engineering, astronomy, and plant biology, we would be less likely to fall into the fallacious view that the Institute for Educational Science has adopted as doctrine. Not only is it completely wrong to suggest that causal claims requires experimentation for scientific support, it is wrong to suggest that we should view causation as an essentially scientific notion: it is as much a historical or legal or commonsense notion as it is a scientific one, and in none of these fields is there any great difficulty in establishing causal conclusions *beyond reasonable doubt*, the correct standard of evidential sufficiency for such conclusions, in science or outside it.² We will now look at whether those procedures can help us with the present problem.

² Note that ‘beyond reasonable doubt’ the standard of evidence that the courts require in felony trials, is far stronger than ‘the balance of evidence’ the standard they use, and sharply distinguish, for misdemeanor trials. People sometimes think that RCTs are the paradigm design because they meet some higher standard than beyond reasonable doubt, perhaps ‘beyond the practical possibility of error.’ But they are far from that standard, which is not even met by proofs in deductive logic and mathematics; we give some of the reasons for caution below. Another verbal obfuscation arises from the use of the term ‘quasi-experimental design’ which strongly suggests a design that cannot pro-

McLaughlin goes on from the passage just quoted to mention what he sees as the insuperable barrier to inferring causation from correlational data, and in his remarks we can see the roots of a solution to our present problem. He says:

“By itself, the SSASD can provide statistics which could be explained as program impact (or lack of impact), but such inference is unwarranted without credible rejection of all competing explanations for the same results (e.g., the program was implemented in schools which had a predisposition to succeed or fail).”

Well, the ‘credible rejection of all competing explanations,’ which I think he regards as unfeasible, is in fact the less obvious part of most demonstrations of causation, as we saw in the autopsy case, the tree rings case, and in most cases of demonstrating historical causation. Moreover, something like it is even essential with RCTs, since there are always alternative explanations knocking at their door, sometimes thumping hard on it. The most obvious examples, as Tom Cook warns us when rejecting the suggestion that the RCT be regarded as a ‘gold standard’ for scientific explanation, include differential attrition, intersite variations in treatment, treatment contamination of the control group, and deliberate treatment matching. He does not mention in that particular passage two other sources of error that seems important to me. One is the need to pick up unanticipated effects, one of the reasons for using an auxiliary goal-blind design, which I’ve discussed elsewhere. The other arises from the fact that RCTs in the social/behavioral/educational area are weak designs compared to the real gold standard, which we find in all serious pharmaceutical investigation—the model they are often carelessly thought to exemplify—because the latter, and not the former, are virtually always ‘zero-blind’ studies. That is, they are not double-blind studies, not even single-blind studies: they have no defense against the dread Hawthorne effect (and the reverse Hawthorne effect, the result of control group competitiveness) and it has often proved to be their undoing. It is sometimes said in their defense that one

vide such strong conclusions as an experimental design. While the point of this view is clear in terms of the abstract theory of experimental design, it is completely false in practice: quasi-experimental designs can often meet the standard of establishing a conclusion beyond reasonable doubt, and a perfectly executed RCT can fail to reach that standard for the supposed conclusion, one reason amongst many being the almost irresistible temptation to state the conclusion in overly general terms. Finally, in the verbal jousting, it’s worth pointing out that the attempt to use “evidence-based” (or, ‘what works’) to mean “supported by RCTs” is simply a sign of scientific illiteracy: it’s about as foolish as Cadillac attempting to copyright the term “standard of excellence”. Cadillac makes some good cars and some bad cars, but they don’t own the standard of excellence. These are all just examples of superficial analysis sidetracking serious research in the way that treating statistical significance as a substitute for educational significance side-tracked educational research for many years.

cannot reasonably arrange double-blind conditions in social/educational experiments, but this is not only false³ but of course no defense against the threat to internal validity, merely a plea for sympathy. Given these problems and the generally conceded difficulties with external validity, a matter of great importance to those of us looking for exportable solutions, it seems best to conclude that the RCT approach is simply one of half a dozen useful designs from which one can choose by striking a balance, in a given context, between cost, required expertise, time window, confidence interval, generalizability, and ethical propriety.

It's clear from the papers we received from earlier efforts as well as our own, that there are several other candidates for membership in the elite group from which choices should be made whenever possible – which I'll call the group of 'strong causal designs.' This group must include interrupted time series (ITS), regression discontinuity (RD), and the so-called value-added design (VAM),⁴ to which we have just suggested adding RCTs with or without zero-treatment controls, and with or without double-blind (DB) protection, which I'll count as only one family of designs (RCT/0, RCT/1, and RCT/2 (a.k.a. RCT/DB)). I've already suggested that we should also include what I'll call the general elimination model or GEM. This is simply the use of a 'causal list,' based on prior investigations, i.e., a list of possible causes of the phenomenon X (which might be significant decline or increase of test scores across a two year interval), combined with systematic elimination of all but one of the causes on the list, using *modus operandi* analysis, as the coroner or epidemiologist or historian does (i.e., checking for the presence of the intervening links that would have to be present in order to connect the actual cause to the effect X).

That gives us a list of five strong causal designs (SCDs, sorry for the alphabet soup but it saves repetitive syllabification). To say these are strong causal designs means they are designs that, properly applied, can yield causal conclusions in educational research that are beyond reasonable doubt. There are some other quasi-experimental designs

³ It is indeed not easy to devise double blind studies in educational research, but the effort is reminiscent of the difficulty in devising RCT designs, which were often said to be impossible when they only required ingenuity. However, the big question that arises before straining to solve this problem is whether it's worth the effort: in many cases, a more intelligent approach would be to drop the zero-treatment group in favor of an interesting alternative treatment group or a major dosage variation. (Dosage variations have the advantage that they can sometimes be extrapolated to get a zero-treatment estimate.) Instead of running the experimental group against a placebo group, by using a dummy treatment one would then learn something of considerable importance, and avoid most ethical and—unfortunately not the same thing—HSIRB problems.

⁴ Note that I have, for simplicity, not added as a separate case the ingenious variation on RD proposed by the RAND group, nor several variations on the ITS design that strengthen it somewhat, such as the use of doubly randomized intervals (where the time of onset and the duration of the intervals is randomly varied across a substantial range).

that should be added to this list, under certain conditions, but it would take too long to get into them in the necessary detail.

However, probably to the dismay of some, I'm going to argue for including two other non-experimental, non-quasi-experimental designs to the SCD list. I'll use the same basis as before, namely, rock-solid scientific credentials. Even though you might not find that easy to accept when you hear what I have in mind, I ask you to keep an open mind until you hear the authorities and examples I will quote. I'm adding these because I think it is extremely important to have them in mind during our present quest to find ways to squeeze bullet-proof causal conclusions out of SSASD.

The easier one to digest of these two is the theory-certified causal claim. The best source of examples is probably astronomy, which is full of causal claims, whether talking about the origin of craters on planetary surfaces or the origin of pulsars. None of these claims are based on experimental evidence, but they are well-based on well-supported theories about how things in the universe operate. These theories are often extrapolated from experimental or observational evidence, but the more remote and colossal the events, the more extreme the extrapolation. What comes to the rescue of the causal inference, sometimes enough to kick it up into the 'beyond reasonable doubt' category, is the conjoint application of the General Elimination Model, which uses causal lists that are in this case based on theoretical analyses (by contrast with causal lists based simply on prior experiments or observations). This kind of inference is good enough to bet the lives of astronauts on, after we have made and tested some of its conclusions without lives at risk, for example in our inferences to the appearance of the far side of the moon, confirmed by flyby photos from unmanned satellites.

But let's take one other example besides astronomy of this type of causal inference. It's a hard one to reject. Of all the causal conclusions on which we have bet thousands of lives and millions of dollars, it would be difficult to find one about which we are more confident than our conclusion that heavy cigarette smoking often causes lung cancer. Yet we have never supported it by RCTs or any other design from the elite list. The nearest we have come is an approximation to an ITS design with naturally-occurring imposition of treatment, a weak sister to the hard-core ITS design we have on the list. The twin reasons for our great confidence in this causal connection are the underlying theory about cell destruction, backed up by studies of skin cancer in mice with nicotine painted onto shaved areas, and the elimination of alternative causes—the general elimination model. I suggest that there are many cases in which we have enough evidence, theory, and causal lists, to pull off the corresponding inferences in looking for the effects of NCLB or other major interventions in the SSASD material.

Finally, we come to the most shocking of all the candidates for the elite list. It is absolutely certain that, contrary to Hume and the positivists, that we can, under the right circumstances, simply observe causal connections. You all know this, with one half of your brain, but we sometimes sequester the halves of our brain. You know because you

have observed a thousand instances, for example that pressing the brake pedal causes your car to slow (nearly always), and that pressing the accelerator causes it to speed up. You know that you knocked over the milk jug on the breakfast table because you saw yourself do it; you know that your young son did it on quite a few occasions because you saw him do it. It's boring to recite the list of cases where people mistakenly infer causation from what they see; we all know that we can make errors about this. But we also know that we usually don't make errors about it in typical cases, and we know this from betting our lives on these causal conclusions many times a day in traffic, thousands of times in our lives, and almost never—or, if we are lucky, never—getting it wrong in these everyday cases. That experience rightly leads us to draw these conclusions and claim that we can draw them, in the proper circumstances, beyond reasonable doubt.

So I appeal to you as the authority here. The issue is not whether you *might* be wrong, conceivably; we all know that's possible. The issue is whether we are ever right to bet our lives on our observations of causal connections. And I suggest to you that you are at least as likely to be right about causation, in the standard observational circumstances of driving your car or watching the kids eating dinner, as you are to be right about causation on the basis of even a carefully done RCT experiment. In both cases, you are perfectly entitled to be certain beyond reasonable doubt.

Now, before you argue against this, let me mention my authority here. In the Cook and Campbell book, if you look very carefully, you will find a passage where they say something like this: “at the end of the day, it cannot be denied that there are circumstances in which we observe causation.” I rest my case.

CONCLUSIONS

So RCTs can only provide reliable causal inferences if they are very carefully run, with highly skilled observers watching constantly for leakage, manipulation, motivational aberrations, side-effects, and differential attrition. It's worth pointing out that those are well-known skills of qualitative research methodology, turning up here at the heart of the quantoid's paradigm for causation. They materialize in the interviews, focus groups, and systematic observation that are required for this intensive monitoring. And there are other ways in which qualitative methodology can support causal conclusions: the most interesting is in establishing the conditions for GEM and in the kind of direct observation of causation that is part of the best types of case study. So our elite group, now including seven approaches, corresponding perhaps to the seven-fold path to enlightenment, includes both qualitative and quantitative approaches, scientific as well as historical ones, experiential as well as experimental ones. This suggests that our strongest approach to causal inferences from the SSASD may well be through a multiple-method approach, suiting the exact choice to the exact problem, and perhaps most commonly combining case study work on the ground to

establish GEM in support of one of the more commonly admired quasi-experimental or value-added approaches. Certainly, attacking one problem by using more than one different SCD is almost mandatory, and likely to prove both successful and revealing about ways to improve our analytic techniques, if we do the appropriate study of our own successes and failures. And nothing said here should be taken to count against the importance of doing RCTs when time, cost, ethics, permissions, need, and the opportunity presents itself or can be created; they often fill a gap that needs to be filled, but sometimes later rather than sooner, and not always.

In short, there are many limitations of the SSASD, of which inconsistency as between states is perhaps the most severe one, and one that could be and should be improved, but there are many causal conclusions that can be well established by using this database as is, including the provision of answers to many of the main questions we face at the moment. But 'using the database as is' does not mean sitting in a chair punching keys; that approach must be combined with specific field and analytic work tailored to the particular problem at hand.