Describing what is Special about the Role of Experiments in Contemporary Educational Research: Putting the "Gold Standard" Rhetoric into Perspective

Thomas D. Cook

Northwestern University

These remarks deal with vexing issues of method choice that are currently bedeviling the educational research community as it seeks to ground educational policy decisions in better evidence. Of particular importance are debates in American educational circles about the need for experiments versus multi-method studies, and it is on this issue that we focus here. The remarks are organized around the exposition of a limited number of basic points. The emphasis is on clarity of presentation rather than mellifluous prose.

1. *If educational policy is to be "evidence-based", educational research will have to deal with many different kinds of issue and question, each often associated with a different method preference. So educational research must be multi-method.*

   Educational policy has to be concerned with many different kinds of issue, including: Who gets what? What does it cost? What is classroom life like? How well are students performing? How are teachers trained? What works to improve performance? Only this last question is explicitly causal; the others are not. They require methods like ethnographies, interviews, surveys that are not necessarily based on explicit causal reasoning. So it is trivial to argue about whether evidence-based research should be multi-method or not. Even causal research is deepened by learning about non-causal issues, such as what the substantive theory behind a program design is, who gets to participate in it or not, how well the program is implemented, who gets greater exposure, and what the program costs. So nearly all causal studies require multiple methods that complement each other. Multi-method, complementary research is desirable even when a causal claim is centrally at issue.

2. *Causal questions are always important in research on educational policy, but especially so now.*

   Policy-makers are selected or elected to make decisions. Their decisions often touch on how to improve the operation of schools by a variety of different means—some financial or administrative, and others concerning classroom practices directly linked to changing teacher and student behavior. The need to improve the educational system is especially

strong in nations where comparative research like PISA or TIMMS indicates that educational performance is inadequate, many of these the larger countries in OECD. But it is also the case with nations currently doing well that will need novel ideas in the future if they are to maintain their high performance levels. Where are these novel ideas to come from? Where are the larger and struggling countries to get ideas for improving their schools, now recognized as inadequate for producing the citizens and workers need for the future? Educational research is expected to help in this by producing causal knowledge that schools can actually use to improve themselves.

3. *For answering questions about effective school practices, the randomized experiment is well warranted theoretically and empirically.*

   The theoretical warrant for experiments comes from a minor variant on the same statistical theory that buttresses survey research, arguably social science's greatest contribution to date. Members of the survey research industry prefer to use random selection when choosing units for study, and non-random selection is almost universally judged as creating biased knowledge of the population being described. Experiments involve only a minor variant to the statistical theory behind the survey. With experiments the intent is not to draw a single sample at random, but to draw two or more samples at random. So created, they represent the same population within known limits of sampling error, and so any difference observed between them at a later date must be due to whatever intervention one sample has had that the others have not. But this is not the only warrant experiments enjoy. Over the many years implementing them, researchers have learned how to improve their implementation so as to meet the method's assumptions better and more often. Also, empirical research has now accumulated comparing the results of a single experiment and non-experiment on the same topic. To date, this research has shown that the two methods almost always fail to converge on the same causal answer to a policy-relevant question about the effectiveness of some social or educational program. Since the theoretical warrant for the experimental result is more compelling than the warrant for the non-experimental result, the presumption is that non-experiments are often biased and that, even if they are not, there would be no way to know this in particular instances unless a randomized experiment were also done. But if an experiment could be done there would be no need for a non-experiment in the first place!

4. *However, experiments are not infallible.*

   Experiments require several assumptions—that a correct random assignment procedure is chosen; that it is then correctly implemented; that there is not differential attrition from the study across the groups being compared; and that there is minimal contamination of the intervention details from one group to another. Each of these assumptions can be violated in a particular study. However, knowledge of these assumptions is widespread now, and practical means are routinely available to limit their role in actual research practice in education. The key here, of course, is the availability of researchers trained not just in the statistical theory of experimentation but also in the fine details of how to implement experiments in complex social settings like school systems. But the main

point is that experiments are only sufficient for unbiased causal knowledge when the above assumptions are demonstrably met.

5. *Experiments do not always answer the question of greatest policy relevance.*

   Many experiments are limited to those schools, teachers or students that agree to participating in whatever treatment condition to which they are assigned by chance, usually through some computer-generated analog to a lottery or coin toss. The causal findings so generated are likely to be bias-free, but to apply only to those who volunteer to be in a study with random assignment. These may not be the types of persons of greatest policy relevance. In the same vein, there can be no logical warrant for inferring that effects found in the past will continue to hold in the future; nor that effects will be the same when an intervention is implemented on a much broader scale than in an experiment. If increasing the scale changes the causal processes responsible for the causal impact observed on the smaller scale, then the experiment will poorly predict outcomes on the broader scale. Advocates of experiments prefer policy to depend on synthesizing the results of multiple studies with different populations of persons, settings and times as well as different ways of instantiating the intervention and measuring the outcome. Compared to syntheses, individual experiments are necessarily more limited in their ability to generalize a causal finding. Of course, scale-up problems are the result of a study's sample size and not its plan for inferring cause. So these same problems hold for any other type of causal study with a similar number of units.

6. *In human history, valid causal knowledge has often come from non-experimental and non-quantitative sources*.

   It would be preposterous to maintain that experiments are necessary for causal knowledge. Our ancestors learned about the causal effects of making fires millennia before there was formal experimentation. And scholars knew that out-group threats usually cause in-group cohesion long before R. A. Fisher created the first formalization of experimental design. Experiments are most needed when it comes to identifying those causal forces whose effects are smaller than the many causal factors identified in the pre-experimental past. Many educational practices today are enmeshed in complex systems that make it difficult to identify their unique causal role. So research tactics require first isolating a component judged to be important and then systematically varying it in schools so as to establish its effects on student performance. Such components are likely to have modest effects that are hard to detect relative to background noise and other changes in social and school life. The fact that causal knowledge does not require experimentation should not blind us to the possibility that experiments are more needed the smaller the effects to be detected.

7. *In social science, experiments are not the only method known to be capable of generating unbiased causal knowledge*.

   From statistical theory and comparative empirical research, we know that regression-discontinuity designs can produce the same causal estimates as experiments in those circumstances where resources are distributed using a single score on a quantitative criterion to determine who is educationally needy or meritorious. We also know from theory that instrumental variable approaches can result in unbiased causal inference if an

instrument can be discovered that is correlated with the treatment assignment but not with errors in the outcome. We also know that causal inferences are unbiased if the process of assignment to treatment is perfectly known or if the outcome is perfectly predicted.

8. *The main rationale for experiments has to lie in their marginal advantage over these other bias-free methods.*

   When identical numbers of schools or students are studied, the unbiased causal estimates from regression-discontinuity designs are less precise than those from experiments. Moreover, regression-discontinuity makes strong assumptions about functional form that cannot always be tested well and that, if mis-specified only a little, can have a dramatic influence on the causal results obtained. As for instrumental variables, in most circumstances other than where random assignment is the instrument, it is difficult to justify the distributional assumptions the method requires; and empirical research on the ability of instrumental variables to generate the same causal answers as experiments has proved to be disappointing when the two were evaluated head to head on the same topic. As for generating a complete model of the selection process, this is what both random assignment and regression-discontinuity transparently do. But in almost every other situation it is impossible to know whether the selection into different treatments due to administrator decisions or self-selection is completely known and measured. And it is very rare to have close to complete prediction of an educational outcome. The experiment is to be preferred over other potentially bias-free methods because it enjoys greater statistical power and its assumptions are more transparent and better understood when compared to other forms of causal research.

9. *In many sectors where policy is made, experiments enjoy more credibility than other kinds of causal study*.

   This is the case, for instance, in health, public health, agriculture, the prevention sciences, criminal justice, audit surveys of compliance about non-discriminatory hiring, and even in research on ways to improve survey research. Experiments also enjoy special credibility in two contexts of educational relevance. In early childhood education in the USA, Congress asked for a randomized experiment to be conducted on its largest national program, Head Start; and the social science knowledge used to promote a universal preschool policy has been publicly proclaimed as credible because it came from experiments. Governments do not allow new drugs or medical procedures without randomized clinical trials on them, and the same seems to be increasingly true for pre-school practices. It also seems to be the case with prevention practices taught in schools, about preventing obesity, violence, smoking or drugs. The majority of studies are experimental; and clearinghouses that report on “what works” place a premium on experimental over non-experimental knowledge. To advocate against randomized experiments playing a major role in determining what works in education requires two compelling arguments. First, that schools are systematically different from other institutions in ways that either make experimentation infeasible or that bias the results it provides there; and second, that experiments are not possible for studying student achievement in schools even though they are currently used to study the efficacy of

health prevention practices there and for studying achievement gains among children attending pre-schools.

10. *Any single experiment in education assumes prior knowledge that is not itself the product of experiments*.

    Experiments are not born in a vacuum. They require prior substantive theory and the experience of persons knowledgeable about what is feasible in school life. They require the availability of good measures of the preferred outcomes, or the ability to construct such measures. They require at least local political and administrative support for the study. And they depend on prior causal studies. While the latter might be experiments, this is not a requirement if prior studies are to confer a marginal advantage in how well future experiments are constructed. Prior studies help determine many technical features of a study, like statistical power; but they also shape how an intervention and its implementation are conceptualized too. Any one experiment builds on the shoulders of prior scholars in theoretical and applied fields. Experimenters are not a new and superior priesthood that can afford to declare itself independent of educational research's past.

11. *Having information from experiments does not guarantee that this information will be used in policy debates, and certainly not to form a decision*.

    Although experiments might be expected to give a marginally superior causal answer compared to other methods, this does not guarantee that these results will be used in debates about educational change. And it certainly does not mean that, when they are used, they will by them selves shape policy decisions. The history of educational research is replete with examples of results not used in the short term, even where they might have been. Indeed, in any democratic society political decision-making does and should depend on many factors other than scientific knowledge alone.

12. *Having scientific information from experiments probably increases the odds of the information being used in policy debates in education*.

    It is difficult to argue the point above for education today, given the short and recent history of school-based experimental research with random assignment. However, in other fields of study, causal results from experiments are routinely preferred over the results from other kinds of study. This is especially true in medical and social science contexts, and when results from multiple studies are synthesized in search of a policy option. In many countries, government agencies that commission policy reviews from experts expect recommendations. In many cases, the synthesis separates out experimental studies for special consideration on the grounds that their results are better warranted than others. And often too, non-experiments are only taken seriously if, in the domain surveyed, their results are on average similar to those from experiments on the same topic. This is, in fact, a standard recommendation in the literature on best meta-analytic practice in both medical and social science domains.

13. *Most opponents of the current emphasis on random assignment experiments in education in the USA are not against experiments per se*.

    Some critics of the experimental agenda object to all quantification and what they call positivist methods, preferring instead post-positivist ones. However, careful reading of

most critics reveals they agree that there is a legitimate role for experiments in education and that experiments may well be the best single way to justify causal conclusions about what works. Their objections to the current experimental emphasis in the USA are basically two-fold. They think the current emphasis results in funding very few studies that have little or no explicit causal purpose; and they think that experiments are so monolithically preferred that other methods for causal learning are crowded-out. These points seem to hold whether the discussion is of Congressionally mandated evaluation studies conducted by contract research firms or of field-initiated research studies, usually university-based, that have to pass peer review to be funded.

14. *It is legitimate to debate how central causal issues should be in educational policy research; and also the extent to which experiments should be preferred over other methods for generating such knowledge*.

15. *However, the key debate concerns whether experiments have been so rare over the last 30 years of educational research that the quality of current knowledge about what works has been compromised thereby*.

    The rarity of experimental knowledge in primary, secondary and tertiary education is striking when the main outcome studied is student achievement. However, experiments are common in pre-school education and in school research where the outcome is the prevention of various social abuses. Why are experimental studies of school performance so rare, then? Why have they not better contributed to building up a body of empirical knowledge about what works in schools, given the need for this and the reality that alternative methods have not created a secure knowledge base about how to improve schools. These alternative methods are of many forms. Some are purely theoretical without much corroborating empirical evidence; others involve post-modernist studies of school process, or traditional qualitative research on schools, or quantitative research on effectiveness that requires accepting opaque and implausible assumptions in order to rule out selection bias. The main justification for putting an experimental agenda at the heart of evidence-based educational research stems from the importance of generating the demonstrably best causal knowledge to meet the pressing task of identifying what works in education. Other methods have not had conspicuous success in this task. If they had, policy actors in interested nations would be able to borrow (and marginally modify) them and thereby improve the performance of their nation. But this has not happened. The putative overselling of experiments in U.S. education today can be interpreted as a short-term correction for the negligence of experimentation in education over the last 30 years.

All the above can be reduced to three summary conclusions:

Summary 1. *It is appropriate to assert that multiple methods are required in research on educational policy*.

Education research is bound to require studies with different methods so long as different types of questions are asked in educational research. One truism of theory of method is that different questions usually require different methods if these questions are to be answered well. Even a good experiment requires attention to theory, sampling, qualitative data collection, value analysis and the like. So even a preference for experiments requires a commitment to multi-method research.

Summary 2. *It is appropriate to claim that generating better causal knowledge has an especially large role to play in educational research today.*

The causal knowledge generated in the past has not been adequate for developing a secure body of knowledge of what works to improve performance in schools. School officials do not have many powerful ideas for improving their schools. Many suggestions can be found in the literature. Some emanate from causal studies, many of which are qualitative or depend on large-scale descriptive data sets, many of them longitudinal but none specifically designed to optimize causal knowledge.

Summary 3. *There is a special need for experiments today.*

Experiments depend on other kinds of knowledge being available, and so there can be no eternal justification for making them the center-piece of method choice in educational research. The justification for according them this temporary status stems from the current importance of improving educational outcomes, from the inadequacies of the methods used in the past to warrant causal claims, and from the experiment's almost universally acknowledged special causal warrant that is derived from both theory and practice, though most of this practice is in fields other than improving students' performance.