On randomized experimentation in education: A commentary on Deaton and Cartwright, in honor of Frederick Mosteller

Stephen W. Raudenbush

1. Introduction

In the introduction to their article, Deaton and Cartwright (2018) bemoan the status that randomized trials seem to have achieved in various domains of science. They warn against the view that randomization is a "gold standard" for obtaining valid causal inference. They seek not only to correct such misperceptions but also to consider how best to improve use of results from randomized trials when such trials are warranted. Although their extended paper makes many assertions with which I agree, I found myself wanting to rise up in defense of randomized assignment.

In 1999 the American Academy of Arts and Sciences sponsored a conference on the state of research in education. Chairing the meeting were Frederick Mosteller and Howard Hiatt, two men who had helped lead the movement in the 1950s to establish the randomized trial as a foundation for causal inference in medicine. They asked why there were still so few randomized trials in education and whether it was time to launch a new epoch of educational research that would parallel the history of medicine. The conference led to an important volume advocating more randomized trials in education (Mosteller and Boruch, 2002).

One important stimulant for this initiative was the Tennessee Class Size Experiment (Finn and Achilles, 1990). Motivating the study was a foundation for causal inference in medicine. They asked why there were still so few randomized trials in education and whether it was time to launch a new epoch of educational research that would parallel the history of medicine. The conference led to an important volume advocating more randomized trials in education (Mosteller and Boruch, 2002).

In 1999 the American Academy of Arts and Sciences sponsored a conference on the state of research in education. Chairing the meeting were Frederick Mosteller and Howard Hiatt, two men who had helped lead the movement in the 1950s to establish the randomized trial as a foundation for causal inference in medicine. They asked why there were still so few randomized trials in education and whether it was time to launch a new epoch of educational research that would parallel the history of medicine. The conference led to an important volume advocating more randomized trials in education (Mosteller and Boruch, 2002).

In 1999 the American Academy of Arts and Sciences sponsored a conference on the state of research in education. Chairing the meeting were Frederick Mosteller and Howard Hiatt, two men who had helped lead the movement in the 1950s to establish the randomized trial as a foundation for causal inference in medicine. They asked why there were still so few randomized trials in education and whether it was time to launch a new epoch of educational research that would parallel the history of medicine. The conference led to an important volume advocating more randomized trials in education (Mosteller and Boruch, 2002).

In 1999 the American Academy of Arts and Sciences sponsored a conference on the state of research in education. Chairing the meeting were Frederick Mosteller and Howard Hiatt, two men who had helped lead the movement in the 1950s to establish the randomized trial as a foundation for causal inference in medicine. They asked why there were still so few randomized trials in education and whether it was time to launch a new epoch of educational research that would parallel the history of medicine. The conference led to an important volume advocating more randomized trials in education (Mosteller and Boruch, 2002).

In 1999 the American Academy of Arts and Sciences sponsored a conference on the state of research in education. Chairing the meeting were Frederick Mosteller and Howard Hiatt, two men who had helped lead the movement in the 1950s to establish the randomized trial as a foundation for causal inference in medicine. They asked why there were still so few randomized trials in education and whether it was time to launch a new epoch of educational research that would parallel the history of medicine. The conference led to an important volume advocating more randomized trials in education (Mosteller and Boruch, 2002).

First, a positive impact of class size reduction depends on unobserved changes in interactions between teachers and students. Unless those changes occur, the results obtained in Tennessee will not be replicated elsewhere. This illustrates a key point in Deaton and Cartwright regarding the importance of theorizing the mechanisms that allow a new intervention to succeed and studying those mechanisms. An intervention that depends for its success on an unknown technology may produce varied results in varied settings. Second, in Tennessee, the
average skills of teachers in small and large classes were statistically equated by random assignment. But taking the reform to scale in California necessarily brought many new and apparently unskilled teachers into the smaller classrooms. It’s remarkable that Tennessee turned 79 schools into a laboratory for studying class size, producing the best study ever of class size. But shifting the policy of an entire state sets in motion a dynamic process that must be understood to avoid policy failure.

2. The transformation of educational research

In 2000 Congress passed the “No Child Left Behind” (NCLB) law. It is well known that NCLB unleashed a regime of school accountability based on high-stakes testing. Less well known is the fact that the law also mandated the formation of the Institute of Education Sciences (IES) with the purpose of creating a new scientific basis for educational research. In 2002 Russell Whitehurst became the founding Director of IES. Whitehurst recognized that a massive, multi-billion dollar “school improvement industry” was selling text books, tests, teacher professional development guides, after-school programs, pedagogical reforms, and more to school districts. The profits were substantial, but there was rarely any credible evidence to support claims that these products were effective. Schools of education trained millions of teachers but little was known about the effectiveness of the training, and university researchers never designed randomized trials to find out. Prior to the formation of IES, few educational researchers embraced randomized trials despite the fact that the non-experimental studies widely in use in education provided little credible causal evidence (Cook, 2002). Partly as a result of Whitehurst’s commitment to randomized trials, by 2013, the IES had carried out 175 large-scale randomized trials (Spybrook, 2014). Other government agencies and foundations have also lent support to movement toward randomized trials, and subsequent leaders of IES have continued to emphasize the importance of random assignment in program evaluation.

How should we evaluate this transformation of educational research in the US?

My view is that the larger impact of IES has been to raise the bar for studies that make (or suggest) causal claims. As a result of the transformation in funding opportunities, economists and experimental psychologists have become more engaged in education research and have brought their disciplinary tools for causal inference with them. Schools of education have increasingly hired economists and statisticians to teach methods for causal inference and designed randomized trials to seek IES or other government or foundation support. IES-funded pre-doctoral training programs have produced hundreds of new scholars who aim to use rigorous scientific methods to studies of educational improvement. In contrast to the widespread opposition to experimentation cited by Cook (2002), school district superintendents, school principals, and teachers are now often quite willing to participate in randomized trials. And serious thinking about causal inference among educational researchers is now widespread.

3. Defining causal effects and clarifying assumptions

Many applied educational researchers now understand some variant of counterfactual thinking often attributed to work in statistics (e.g., Neyman and Iwaszkiewicz, 1935; Rubin, 1978; Holland, 1986) and economics (Haavelmo, 1944; Roy, 1951; Heckman, 1977). The basic idea is that each participant in a study has a potential outcome under each possible course of action. The impact of one course of action relative to a second for a given participant is then the difference between two potential outcomes. We’ll never observe this effect for any specific participant, but we can estimate the average impact under three key assumptions.

The first is that one participant’s assignment to intervention does not affect another person’s potential outcome. Rubin (1986) called this the “Stable Unit Treatment Value Assumption” (SUTVA) because it requires that one person’s potential outcome remains stable when another person is assigned to intervention or control. This assumption will be violated, for example, if those assigned to a new program share knowledge or resources with those assigned to a non-intervention comparison group, or if the behavior of program participants otherwise influences the outcomes of those in a comparison group. In most randomized trials in education since 2002, whole classrooms or schools, rather than individuals, are assigned at random to interventions (Spybrook, 2014). A benevolent consequence is that violations of SUTVA in the form of spillovers of knowledge and resources can often be minimized.

The second key assumption required for valid causal inference is often called “ignorable treatment assignment.” This requires that a participant’s potential outcomes cannot predict one’s treatment assignment. In studies where program staff select participants or when the participants themselves select the program in which they will participate, this assumption is often hard to justify. Random assignment effectively implemented assures that this assumption is satisfied. Non-randomized studies of social programs are credible only to the extent that the selection of persons into comparison groups approximates random assignment. The burden is on the investigator to support this claim, either by asserting that assignment was accidental (if not formally random) or by arguing that statistical control of observable background characteristics fully accounts for biased selection into program groups.

Third, when we use a sample to estimate the average value of a variable in a broader population, we must make the case that the sample represents the target population. This is surely true when the variable in question is a causal effect. If the experimenters draw a probability sample of units from a well-defined population and then assign those units at random to programs, we typically have a strong case for extrapolating the sample result to the target population. The National Head Start Impact Study (Puma et al., 2010) approximated this ideal, but such random selection from a population is exceptional in studies of interventions. Educational researchers have devised clever new ways to support claims of generalization (Stuart et al., 2011; O’Muircheartaigh and Hedges, 2014). A more common approach is to describe the sample in enough detail so that informed persons can evaluate the plausibility of extrapolation to similar sub-populations. One also can assess theoretical claims about who should benefit most and in what settings. Hence, scientific judgement is essential in interpreting even the results of well-designed studies. In any case, one must keep in mind that a valid generalization also requires that implementation of an intervention in a broad population will be similar to the implementation as carried out in the sample. The extrapolation from Tennessee to California is instructive.

4. Natural versus planned experiments

My evaluation of the contribution of the randomized trial in education begins with a broader view of causal scenarios. Two stand out: the natural experiment and the planned intervention.

4.1. Natural experiments

Natural experiments typically arise from broad changes in policy that encourage educators, parents or students to behave in new ways. For example, Robert Eschmann and I recently reviewed 13 studies conducted in 8 societies of the impact of providing universal pre-kindergarten schooling (Raudenbush and Eschmann, 2015). These studies were mostly remarkably well designed. In many cases, researchers compared children who were just old enough to qualify for pre-kindergarten to children who were just a bit too young to qualify. This comparison could be made during the year a policy was enacted and weighed against the same comparison in years when the policy was not
enacted. Assignment to an offer of pre-k was not random, but this procedure created comparison groups that were remarkably similar on observables and, based on theoretical considerations, probably similar on unobservables as well. To say that assignment was nearly equivalent to random is quite reasonable. The beauty of these studies, particularly when pooled together, is that extrapolation to the real world of policy is entirely realistic because the samples were broadly representative or, in fact, included entire populations.

In the same article, we review evidence on policies that increase the length of the school day, interrupt the academic year with summer recess, and extend years of compulsory schooling during adolescence. We do not have random assignment in these studies, but the conditions for causal inference are plausibly well met, and generalization to the scaled-up policy is immediate.

4.2. Planned interventions

We cannot always wait for world to dish out natural experiments. We need to supplement these with deliberate attempts to change the world for the sole purpose of learning how the world works. That means putting in place new conditions that would not arise without our efforts. When we do this we minimally need some kind of comparison group and we need to think about how participants will be assigned to treatment and control groups. My view is that in such instances we should virtually always consider random assignment. Random assignment may not be feasible or ethical, but when it is, the advantages are typically substantial. The assumption of ignorable treatment assignment is met. In contrast, in the absence of random assignment, that assumption is almost always questionable in the case of the planned intervention. In particular, in planned interventions without random assignment, we must assume that all selection bias is accounted for by the background characteristics we can observe.

Moreover, researchers are getting better at making randomization feasible. When Robert Slavin, Geoffrey Borman, and their colleagues sought to evaluate “Success for All” (SFA), a comprehensive, school-wide instructional program, they began by trying to recruit a set of schools that would agree to participate in data collection regardless of the outcome of random assignment. It proved hard to recruit schools. All schools were under pressure to improve because of accountability imposed by the “No Child Left Behind” law. The penalty for losing the random draw was that a school would be deprived of a strategy for improvement. The consequences of failing to improve could be severe, even including school closure. So Slavin and Borman decided to randomly assign schools to receive SFA starting either in kindergarten or in grade 3. Those schools randomly assigned to start SFA in kindergarten would expand SFA to grade 1 the next year, grade 2 the year after, etc. Those starting in grade 3 would expand to grade 4 the next year, and so on. So after 6 years all schools would have SFA in all years. In the meanwhile, each school receiving SFA early was a comparison school for those receiving SFA late and vice versa. Using this strategy, the experimenters quickly recruited a large sample of schools and produced a credible randomized trial with highly encouraging findings (Borman et al., 2007). Another increasingly used strategy for randomized trials is to exploit the fact that many new programs are over-subscribed. A

5. Reflecting on Deaton and Cartwright

Thiers is a lengthy article, quite comprehensive, and very thoughtful. Their emphasis on the role of a priori theory in contemplating who will benefit and why would increase the yield of intervention studies. Their recommendation to employ theory when extrapolating to a target population and thinking about how conditions change when an intervention scales up are compelling. Thinking hard about how to use randomize trials within a broader theoretical framework for policy is entirely convincing.

However, I think the distinction between natural experiments and planned interventions is critical. Studies of natural experiments often make the assumption of ignorable treatment assignment credible even without random assignment, and typically support reasonable generalization. In contrast, ignorable assignment in planned interventions is very often difficult to defend without random assignment. So random assignment is often decisive in obtaining credible causal information from planned intervention studies. However, generalizing from planned interventions, with or without random assignment, will often be somewhat speculative. Successful extrapolation will likely depend on compelling theory as well as sound methods. Synthesizing findings across studies is essential.

Deaton and Cartwright appear to be writing for an audience that is sold, perhaps oversold, on the value of random assignment. So the tone of their piece is cautionary, even skeptical. I live in a world where the embrace of new rigorous thinking about causation is tentative, and where many educational researchers resent large allocation of funding to randomized trials. That worries me. So my aim is to adopt more favorable stance regarding the importance of randomized experimentation within a portfolio of research on educational improvement. Without overselling what we can learn from randomized trials, I am decidedly opposed to returning to the pre-1999 era, and wish to thank the late Frederick Mosteller for taking the lead in opening up a new era.

References


