Social Science & Medicine 210 (2018) 63–66

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

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 trials. This surely reflects my own experience in educational research, a domain that was nearly bereft of such trials prior to 2002. I regard the turn since then toward an emphasis on such trials as remarkably refreshing. So I thought I would tell the story of the sea change in educational research before offering some conclusions about natural experiments, planned interventions, and the contribution of random 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 there 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 stalemate in the Tennessee legislature. The lawmakers couldn't agree on whether to outlaw large classes, but they did agree to study that question. Helen Pate Bain, an associate professor at Tennessee State University and well-known advocate for education reform, argued for a randomized trial (Boyd-Zacharias, 1999). Past studies of class size had mixed results, and some studies seemed to suggest that larger classes were actually more effective than small classes, almost surely because more effective teachers tend to attract more students. So Tennessee funded a study in which kindergarten students and teachers were randomly assigned to classes large and small. Finn and Achilles (1990) reported that "the results were definitive:" Reducing class size could significantly increase student learning in reading and mathematics. Mosteller (1995) celebrated this finding for a broad audience and asserted at the 1999 conference that this was among the best studies in the history of education. The study had interesting by-products. Low-income and minority students benefitted most from class size reduction (Krueger and Whitmore, 2001; Shin and Raudenbush, 2011). Krueger and Whitmore (2001) found that students randomly assigned to smaller kindergarten classes were, on average, more likely to attend college. The random assignment of students to teachers enabled Nye et al. (2004) to estimate the variation in the causal effects of individual teachers. Their study showed that the magnitude of variation in teacher effects was consistent with the magnitude of teacher effects estimated in non-randomized "value added" studies, lending some support to the methods used in those studies. Chetty et al. (2011) exploited the random assignment of students to teachers to trace the positive effects on adult outcomes of having effective teachers.

Motivated in part by this study, the State of California passed a law setting a minimum class size and appropriating approximately one billion dollars to assist districts in class size reduction. This massive effort came to be regarded as unsuccessful (Jepsen and Rivkin, 2009). The law compelled districts to scramble to find enough teachers to populate the small classes. Competition for good teachers intensified, and more affluent districts appeared to successfully raid less affluent districts, likely increasing social inequality in access to good teachers (Jepsen and Rivkin, 2002). Paralleling the thinking of Cartwrght and Deaton, my colleagues and I concluded that the Tennessee trial was extremely illuminating but so was its misapplication in policy (Cohen et al., 2003).

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

E-mail address: sraudenb@uchicago.edu.

https://doi.org/10.1016/j.socscimed.2018.04.030

Available online 21 April 2018

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 predoctoral 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 seriously 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, schoolwide 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 randomized lottery is a fair way to decide who will get the program first. One can then follow lottery winners and losers to evaluate the impact of the program.

One advantage of many planned experiments in education is that randomization occurs in multiple sites (Spybrook, 2014). For example, classrooms may be assigned at random within schools; or schools may be assigned at random within districts. The sites and the students who serve them are typically very heterogeneous. This heterogeneity affords the opportunity to use theory to test explanations about generalizability (see Raudenbush and Bloom, 2015).

5. Reflecting on Deaton and Cartwright

Theirs 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

Borman, G.D., Slavin, R.E., Cheung, A.C., Chamberlain, A.M., Madden, N.A., Chambers, B., 2007. Final reading outcomes of the national randomized field trial of Success for All. Am. Educ. Res. J. 44 (3), 701–731.

Boyd-Zacharias, J., 1999. Project STARL the story of the Tennesssee class sie study. American Educator. Summer 1999 1–6.

Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. Q. J. Econ. 126 (4), 1593–1660.

Cohen, D.K., Raudenbush, S.W., Ball, D.L., 2003. Resources, instruction, and research. Educ. Eval. Pol. Anal. 25 (2), 1–24.

Cook, T.D., 2002. Randomized experiments in educational policy research: a critical examination of the reasons the educational evaluation community has offered for not doing them. Educ. Eval. Pol. Anal. 24 (3), 175–199.

Deaton and Cartwright, 2018. Understanding and misunderstanding randomized controlled trials. Soc. Sci. Med. 210C, 2–21.

Finn, J.D., Achilles, C.M., 1990. Answers and questions about class size: a statewide experiment. Am. Educ. Res. J. 27 (3), 557–577.

Haavelmo, T., 1944. The probability approach in econometrics. Econometrica: Journal of the Econometric Society iii-115.

Heckman, J.J., 1977. Sample Selection Bias as a Specification Error (With an Application to the Estimation of Labor Supply Functions).

Holland, P., 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81 (396), 945–960.

Jepsen, C., Rivkin, S.G., 2002. Class Size Reduction, Teacher Quality, and Academic Achievement in California Public Elementary Schools. Public Policy Institute. of CA.

Jepsen, C., Rivkin, S., 2009. Class size reduction and student achievement the potential tradeoff between teacher quality and class size. J. Hum. Resour. 44 (1), 223–250.

Krueger, A.B., Whitmore, D.M., 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from Project STAR. Econ. J. 111 (468), 1–28.

Mosteller, F., 1995. The Tennessee study of class size in the early school grades. Future Child. 113–127.

Mosteller, F., Boruch, R., 2002. In: Evidence Matters: Randomized Trials in Education Research. Brookings Institution Press, Washington, DC, pp. 80–119.

Neyman, J., Iwaszkiewicz, K., 1935. Statistical problems in agricultural experimentation. J. Roy. Stat. Soc. Suppl. 2 (2), 107–180.

Nye, B., Konstantopoulos, S., Hedges, L.V., 2004. How large are teacher effects? Educ. Eval. Pol. Anal. 26 (3), 237–257.

O'Muircheartaigh, C., Hedges, L.V., 2014. Generalizing from experiments with non representative samples. J. Roy. Stat. Soc. Series C 63, 195–210.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., et al., 2010. Head Start Impact Study. Final Report. Administration for Children & Families.

Raudenbush, S.W., Bloom, H.S., 2015. Learning about and from a distribution of program impacts using multisite trials. Am. J. Eval. 36 (4), 475–499.

Raudenbush, S.W., Eschmann, R.D., 2015. Does schooling increase or reduce social inequality? Annu. Rev. Sociol. 41, 443–470. http://dx.doi.org/10.3102/0013189X15575345.

Roy, A., 1951. Some thoughts on the distribution of earnings. Oxf. Econ. Pap. 3 (2), 135–146.

Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Ann. Stat. 34–58.

Rubin, D.B., 1986. Comment: which ifs have causal answers? J. Am. Stat. Assoc. 81, 961–962.

Spybrook, J., 2014. Detecting intervention effects across context: an examination of the precision of cluster randomized trials. J. Exp. Educ. 82 (3), 334–357.

Shin, Y., Raudenbush, S.W., 2011. The causal effect of class size on academic achievement: multivariate instrumental variable estimators with data missing at random. J. Educ. Behav. Stat. 34 (No. 2), 154–185.

Stuart, E.A., Cole, S.R., Bradshaw, C.P., Leaf, P.J., 2011. The use of propensity scores to assess the generalizability of results from randomized trials. J. Roy. Stat. Soc. 174 (2), 369–386.