Lessons Learned From

Large-Scale Randomized Experiments

Robert E. Slavin

Johns Hopkins University

Alan C. K. Cheung

The Chinese University of Hong Kong

Corresponding author: Robert Slavin, JHU, 300 East Joppa Rd., 5th flr.,

Baltimore, MD 21286; rslavin@jhu.edu

Abstract

Large-scale randomized studies provide the best means of evaluating practical, replicable approaches to improving educational outcomes. This article discusses the advantages, problems, and pitfalls of these evaluations, focusing on alternative methods of randomization, recruitment, ensuring high-quality implementation, dealing with attrition, and data analysis. It also discusses means of increasing the chances that large randomized experiments will find positive effects, and interpreting effect sizes.

Lessons Learned from Large-Scale Randomized Experiments

In recent years, evidence of effectiveness has become increasingly important in educational policy and practice (Buck & McGee, 2015; Executive Office of the President, 2015; Haskins & Margolis, 2015; Nussle & Orszag, 2014; Slavin, in press). For example, the 2015 Every Student Succeeds Act (ESSA) identifies three standards of evidence. "Strong" programs have at least one randomized study showing positive effects, "moderate" programs have at least one quasi-experiment (e.g., a matched study) showing positive outcomes, and "promising" programs have at least one correlational study with controls for inputs showing positive outcomes. ESSA encourages use of programs meeting these standards, and requires it for school improvement grants (Slavin, in press; West, 2016).

The unprecedented ESSA standards were made possible by development and research over a long period of time, but especially over the past 15 years. During that time, the U.S. Department of Education and other agencies and private foundations have supported a broad range of research using designs in which students, teachers/classes, or schools are assigned at random to experimental or control groups. This funding has led to an outpouring of large randomized experiments, especially cluster-randomized experiments in which teachers/classes or, especially, whole schools are randomly assigned to treatments (U.S. Department of Education, n.d.). These investments have greatly increased

the number of proven programs in education, but they have also taught experimenters some hard lessons, as a substantial proportion of programs evaluated under funding from the Institute for Education Sciences (IES) or Investing in Innovation (i3) are not reporting significantly positive impacts, or they are reporting significant but very small effect sizes. For example, out of the 44 i3 projects reported as of January, 2017 (Lester, 2017), only 13 produced positive results (30%) and the rest (70%) reported small effect sizes or no impact (also see West, 2016).

This article discusses lessons learned from large randomized experiments, in hopes of helping experimenters increase the number of programs producing positive outcomes and helping readers of the findings of large randomized experiments interpret the findings and understand the contributions they make.

Why Large Randomized Experiments Matter

In education, and in many other fields, randomized experiments are referred to as the "gold standard" for research intended to determine the effectiveness of programs or practices. Other designs, such as correlational or qualitative studies, can be useful for theory-building or description, but experiments are ideal for testing the outcomes of interventions in comparison to what would have happened without the intervention. When individuals or groups are assigned at random to experimental or control groups, it can be assumed that

the groups were initially equivalent not only on measured variables, such as pretests, but also on unmeasured variables, such as interest in implementing the program or practice. The most important goal of experimental design is to eliminate bias, and well-done randomized experiments with objective measures provide the best means known to do this (Slavin, 2008).

The main alternative form of experiment, the quasi-experiment, typically uses matching or controls for pretests and other covariates to control for any pre-existing experimental-control differences. These controls can make a quasi-experiment just as good as a randomized experiment if pre-existing differences are small and do not strongly affect outcomes, and if all important differences are accounted for. The problem is that neither the experimenter nor the reader can be sure that all important factors have in fact been controlled for. For example, imagine that 20 schools are using Program X. A researcher might find 20 very similar schools in the same district, pretest students who have not yet had Program X, and note the growth students make in the 20 Program X schools in comparison to the growth in the 20 control schools. Designs like this are very common in education, and can be unbiased. But what if the teachers using Program X are better teachers than controls, and that's why they were willing and able to use Program X? What if the principals or department heads in schools that embraced Program X were also better at supporting teachers' development in general? These would be difficult factors to measure or control for. In any case,

for these and other reasons, a recent review of 645 high-quality studies across grades pre-k to 12 in reading, math, and science found that quasi-experiments obtained inflated effect sizes in comparison to randomized experiments, by a ratio of more than 1.4 to 1 (Cheung & Slavin, 2016). Randomized experiments, where randomization is done at the teacher or school levels, do not suffer from these problems, because unmeasured pre-existing variables, such as teachers' willingness or ability to use a given innovation, can be assumed to be equal in experimental and control groups. Random assignment of students within schools makes it likely that experimental and control students can be considered equal, but the teachers of the experimental and control classes cannot be considered equal if they volunteered or were non-randomly assigned to one or the other group. That is, random assignment within schools may still allow for bias unless teachers themselves are randomly assigned to teach experimental or control classes.

Difficulties of Carrying Out High-Quality Randomized Experiments

Much as randomized experiments may be ideal in principle for testing effects of programs or practices, carrying them out in real-world settings can be difficult. The focus of this article is on describing these difficulties, as well as offering strategies for coping with or minimizing them:

1. Level of analysis

2. Strategies for random assignment
3. Maintaining the integrity and effectiveness of treatments
4. Dealing with attrition and intent-to-treat
5. Interpreting outcomes

Level of Analysis

Randomized experiments in education typically involve randomly assigning individual students, teachers/classes, or schools to experimental or control groups. The level at which randomization takes place makes a big difference.

Methodologists insist that analysis of data be done at the level of random assignment and treatment (Bloom, Bos, & Lee, 1999; Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). In a study that randomly assigns at the student level, student-level analyses are appropriate, especially if teachers are also randomly assigned to treatment or control groups. With a good covariate (such as pretest), a student-level study might need as few as 200-300 students to detect an effect size of +0.20. However, clustered analyses at the teacher/class or school level typically require 40 to 50 *clusters,* or 1200 to more than 10,000 students (Spybrook, Bloom, Congdon, Hill, Martinez, & Raudenbush, 2011).

So why does anyone use clustered designs? The answer is that individual random assignment requires that both experimental and control students be in

each school or class. If the treatment is one-to-one tutoring or family therapy, this may work fine. But students are taught in classes and schools. It would be very difficult to randomly assign students within classes or schools for a whole-class math or reading program, for example, and even if a researcher could do this, the treatment effect might be diminished because control teachers are likely to learn about and use the program. It may be better to focus on increasing the quality of implementation, to increase the effect size, than to try to increase power by randomly assigning individual students within cluster, such as schools or classes.

Strategies for Random Assignment

Random assignment of students. A researcher randomly assigning students may either select students at random to be in a treatment or control group, or may block on school or classroom, as when 10 tutoring-eligible students in each of 6 classes are randomly assigned to get tutoring or not to get tutoring. Qualifying students might be put in matched pairs on key characteristics (e.g., pretest levels, English language proficiency, ethnicity, gender) and then one student in each matched pair might be assigned at random to receive tutoring in the fall semester, while the other member of the pair would serve in the control group in the fall and then receive tutoring in the spring semester. If students are randomly assigned within blocks, the block should be accounted for in the analysis (see May et al., 2016, for an example). Alternatively, researchers could

forego the matching and just assign students at random, assuming that the randomization will ensure equality between the experimental and control groups.

Cluster random assignment of schools or teachers/classes. Random assignment of teachers/classes or schools, called cluster randomization, involves identifying eligible schools or classes and randomly assigning them to treatment or control groups.

Enhancing willingness to participate in randomized studies. Schools or teachers may argue that random assignment is unfair, as it deprives the control group of the treatment. One way to handle this, in individual or cluster studies, is to use a delayed treatment design. In such designs, control schools/teachers/students are offered the treatment at the end of the study such as in the follow semester or school year. In addition to making recruitment easier, this design may increase the willingness of the control schools or teachers to cooperate in data collection, since they know they will eventually receive the program.

Another way to make serving in the control group more appealing is to offer control schools or teachers modest funding to use as they wish. Often, this is cheaper and more attractive to schools than receiving the treatment in the future.

Even using delayed treatment or cash incentives for control groups, teachers and principals considering participating in a randomized experiment are likely to complain that it is unfair to their children to withhold the treatment. Yet

principals can be reminded that if they do not participate in the study and cannot afford to purchase the program, they have 0% chance of getting it, so by participating, they are increasing their chances to 50%, or even 100% in a delayed-treatment design.

It may be important to have teachers vote to participate in a randomized experiment (before randomization, of course). This increases the chances that the whole staff in each school will be committed to full implementation, and the randomization ensures that experimental and control schools are equally committed.

Of course, the foregoing discussion depends on the assumption that researchers are providing the experimental program for free. This is usually crucial: school staffs are unlikely to sign up for a randomized study unless they feel it will be beneficial to their students, and free.

Dealing with attrition and intent-to-treat

In any large randomized experiment, individual students are sure to be lost due to normal mobility (e.g., students moving out of the district) and other changes. When students move from their schools for any reason, the final analysis will drop not only their posttest, put also their pretest. However, it is rare that student mobility alone upsets the equality of the experimental and control samples at pretest, unless somehow the treatment is so burdensome that it causes students

to drop out. For student mobility, all that is necessary is to compute pretest means and other initial information based on the students remaining in the final sample, the students who submitted both pre- and posttests.

A much bigger problem arises when whole schools withdraw. This may take place when schools close due for reasons unrelated to the treatment, but there is a bigger problem if schools withdraw because of turmoil, changes of principals or other staff, or other problems. Worst of all, schools may drop out of the experiment because they decide they do not like or cannot implement the program.

When schools withdraw due to turmoil, this is a serious problem because it may be that the withdrawing schools were weaker than other schools in the first place, upsetting the equality due to initial random assignment. However, if the remaining sample is still well matched on pretests and demographic variables, and experimenters can argue that the withdrawal could not have been due to the treatment itself, then the study may still be acceptable However, the study might be demoted to the status of a quasi-experiment, still perhaps a valuable contribution but not "gold standard."

If there is any reason to believe that any schools withdrew *because* of the treatment, then the study could have a significant element of bias, because the experimental group has lost some schools that were presumably the weakest, least

amenable to reform, or otherwise less likely to have done well with or without the experiment.

Intent to treat. One procedure that should always be used in randomized experiments and may help mitigate problems of attrition is called intent-to-treat (Gupta, 2011). This means that experimenters obtain and use posttest data for every student who took a pretest, even if the student's school dropped out, or the student transferred to another school. The reason for intent-to-treat is to ensure that the experimental and control groups stay equal at posttest (as they were at pretest, due to the use of random assignment). A randomized experiment using intent-to-treat eliminates any chance that any attrition might have unfairly favored the experimental or control group.

Effects of treatment on the treated. The problem with intent-to-treat, however, is that a lot of students in the posttest sample may have received little or nothing of the treatment, watering down any treatment effects.

To deal with this problem, randomized experiments reporting intent-to-treat estimates of treatment effects also often report effects of treatment on the treated (TOT). TOT analyses only include students, classes, and schools that actually received the treatment. TOT estimates are no longer randomized, but if attrition was modest, they are likely to be similar to the intent-to-treat estimates, and may be useful in reporting and explaining study findings.

Maintaining the Integrity and Effectiveness of Treatments

Another crucial focus in large randomized experiments is on making sure that teachers and principals have sufficient support to ensure high-quality implementation of the program being studied.

As noted earlier, most randomized experiments fail to produce statistically significant positive effects (Boulay, Goodson, Frye, Blocklin, & Price, 2015; Lester, 2017; Preschool Curriculum Evaluation Research Consortium, 2008). Some of these were underpowered, which just means that the sample size was not large enough to detect what might have been a meaningful effect size (say, +0.15 or more). But large randomized experiments may find effect sizes on independent measures that would not be educationally important even if they had been statistically significant.

The reason for this is often that the experimenters greatly underestimated the assistance school staffs need to undergo meaningful change. Change is hard, and teachers who are used to teaching in a particular way are likely to have difficulties teaching in a new way.

Increasing the chances of high-quality implementation can be done in several ways.

1. **Provide well-structured teacher's manuals, student materials, software, and other supports.** It is crucial to be explicit about what the program is.

2. **Provide sufficient professional development.** Teachers should leave their PD sessions with a clear idea about what the program looks like if it is well implemented. Videos showing proficient implementation, simulations putting teachers in the role of students to learn in the new way, or computer simulations to ask teaches to make judgments or simulate teaching moves can all help teachers visualize themselves using new methods.
3. **Provide on-site coaching.** After professional development, it is important to have coaches who are deeply familiar with the program cycle through experimental schools to observe what teachers are doing and give them feedback. Coaches should celebrate progress teachers have made and plan for next steps, recognizing that all new programs take time to learn. It may be useful to introduce a new program in steps, to allow teachers to master each step. Job-alike groups, such as all fourth grade teachers or all algebra teachers, might meet on a regular basis with coaches, either in person or electronically, to jointly discuss triumphs, questions, and problems.
4. **Provide internal facilitators.** For complex programs, it may be important for schools to appoint an internal facilitator, usually an experienced and respected teacher. The facilitator's role is to rotate through teachers' classes, give them support and feedback, ensure that all teachers have the

materials they need, and serve as the key point of contact for the experimenter's coaches.

5. **Use benchmark assessments.** It may be important to assess students' progress on tests aligned with the final outcome, perhaps four times a year. The benchmark assessments could be useful in informing the experimenters and staff alike about the progress being made, so they can adjust school practices where outcomes appear to be weak. Regular formal assessments of the overall quality of implementation may also be collected on a quarterly basis, similarly informing the school staff and coaches about where improvements are needed.

Interpreting Outcomes

At the end of a large randomized experiment, there are usually two major findings: Effect sizes and statistical significance.

An effect size is computed from the difference between the experimental and control group means, adjusted for any pretests or their covariates, divided by the unadjusted standard deviation. But what is a large or small effect size?

Cheung & Slavin (2016) recently addressed this question by looking at 645 studies evaluating programs in grades K to 12 reading, math, and science, as well as pre-K programs. Two key design characteristics had a particularly strong effect on effect sizes: Sample size (smaller studies produce inflated effects (Slavin

& Smith, 2009) and randomized and quasi-experimental studies. Previous reviews of research comparing effect sizes in randomized vs. quasi-experimental evaluations have found mixed effects (de Boer et al., 2014; Glazerman, Levy, & Myers, 2002; Heinsman & Shadish, 1996; Li & Ma, 2010; Lipsey & Wilson, 1993; Rake, Valentine, McGatha, & Ronan, 2010; Shaddish, Clark, Steiner, 2008; Slavin, Lake, & Groff, 2009; Torgerson, 2007; Wilde & Hollister, 2008). However, Cheung & Slavin (2016) had a sufficient sample of high-quality studies to permit a test of the effects of randomized vs quasi-experimental designs on effect sizes to be made with adequate power. In addition, the studies included had already met strict inclusion standards applied by the Best Evidence Encyclopedia (www.bestevidence.org). The analysis showed that average effect sizes were greatly affected by a combination of study design (randomized vs. quasi-experimental) and sample size. The smallest average effect size was for large randomized experiments. These averaged only ES=+0.12. In contrast, small quasi-experiments averaged ES=+0.32, a substantially inflated estimate of program impacts. Other combinations fell between these means.

If large randomized experiments (usually clustered) produce an effect size of +0.12, then it is reasonable to compare effect sizes from other large randomized experiments to this mean. For example, an effect size of +0.20 would be impressive for a large randomized study, modest for a small quasi-experiment, and about average for a large QED or a small randomized experiment. This is

very important, as readers, reviewers, potential funders, and often experimenters themselves are frequently unimpressed by effect sizes in the teens.

The situation gets more complex when statistical significance is considered. Randomized experiments usually recruit enough schools to obtain enough power to detect an effect size of +0.20. They are unlikely to find effect sizes as small as +0.12 without much larger numbers of schools, which in turn require much larger funding.

One solution sometimes proposed for this problem is to use higher p-values than usual in large cluster-randomized studies (see Bloom, 2005). For example, using p-values of p=.10 or even p=.20 in cluster randomized trials may provide better information on program effectiveness than lower-quality and/or smaller quasi-experiments that meet the conventional standard of p=.05. Another solution might be to combine the findings of similar large cluster-randomized experiments, where each small experiment might not be significant but two or more together might be.

The low expected value of effect sizes in large cluster-randomized experiments creates the quite common reality of very high quality experiments producing higher-than-average effect sizes (for studies of their type) but not being statistically significant. This is a problem that needs to be addressed.

Conclusion

Despite their difficulty and expense, large randomized experiments, especially cluster-randomized experiments, should remain the focus of experimentation and funding by government and others who are eager to provide educators with valid information on the effectiveness of educational treatments. Smaller and quasi-experimental studies are fine as pilot studies, but when a definitive evaluation is needed for a program that could be used on a wide scale, a large cluster-randomized experiment is the best way to obtain unbiased estimates of program impacts in real-life schools and classrooms. Instead of complaining about the costs and difficulties of such experiments and noting the many studies that fail to find positive impacts, we need to learn how to build robust treatments and evaluate them in powerful experiments, until our field is routinely producing trustworthy information supporting many effective, replicable, and practical approaches to improving educational outcomes.

References

Bloom, H. S. (2005). *Learning more from social experiments: Evolving analytic approaches.* New York: Russell Sage Foundation.

Bloom, H. S., Bos, J. M., & Lee, S. W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review, 23,* 445-469. DOI: 10.1177/0193841X9902300405

Boulay, Beth, Barbara Goodson, Michael Frye, Michelle Blocklin, and Cristofer Price. (2015). *Summary of research generated by Striving Readers on the effectiveness of interventions for struggling adolescent readers* (NCEE 2016-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Buck, S., & McGee, J. (2015). *Why government needs more randomized control trials: Refuting the myths.* Houston: Laura and John Arnold Foundation.

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45* (5), 283-292. DOI: https://doi.org/10.3102/0013189X16656615

de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students' academic performance:

A meta-analysis. *Review of Educational Research, 84*(4), 509-545. DOI: https://doi.org/10.3102/0034654314540006

Donner, A., & Klar, N. (2000). *Design and analysis of group randomization trials in health research*. London: Arnold.

Executive Office of the President (2015). *Every Student Succeeds Act: A program report on elementary and secondary education*. Washington DC: The White House.

Glazerman, S., Levy, D.M., & Myers, D. (2002). Nonexperimental replications of social experiments: A systematic review. Princeton, NJ: Mathematica Policy Research, Inc. Available at: http://www.mathematica-mpr.com/~/media/publications/PDFs/nonexperimentalreps.pdf

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2 (3), 109-112. doi: 10.4103/2229-3485.83221

Haskins, R., & Margolis, G. (2015). *Show me the evidence: Obama's fight for rigor and results in social policy*. Washington, DC: The Brookings Institution.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments? *Psychological Methods, 1*(2), 154-169. DOI: 10.1037/1082-989X.1.2.154

Lester, P. (January 19, 2017). *Investing in Innovation (i3): Strong start on evaluating and scaling effective programs, but greater focus needed on innovation*. Retrieved May 13, 2017, from http://socialinnovationcenter.org/wp-content/uploads/2017/01/SIRC-i3-report.pdf

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review. 22*, 215-243.

Lipsey, M.W. & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 1993, 48*(12), 1181-1209.

May, H., Sirinades, P., Gray, A., & Goldsworthy, H. (2016). *Reading Recovery: An evaluation of the four-year i3 scale-up.* Newark, DE: University of Delaware, Center for Research in Education and Social Policy.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Nussle, J., & Orszag, P. (Eds.). (2014). *Moneyball for government.* Washington, DC: Disruption Books.

Preschool Curriculum Evaluation Research Consortium (2008). *Effects of preschool curriculum programs on school readiness* (NCER 2008-2009). Washington, DC: National Center for Education Research, Institute of Education Sciences,

U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Rake, C. R., Valentine,, J. C., McGatha, M. B., & Ronau, R. N. (2010). Methods of instructional improvement in Algebra: A systematic review and meta-analysis. *Review of Educational Research*, *80*(3), 372-400. DOI: https://doi.org/10.3102/0034654310374880

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185. doi:10.1037/1082-989X.2.2.173

Shadish, W.R., Clark, M.H., & Steiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, *103*, 1334-1343. Doi: 10.1198/016214508000000733

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 47-50.

Slavin, R. E. (in press). Evidence-based reform in education. *Journal of Education for Students Placed at Risk.*

Slavin, R. E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500-506. DOI: https://doi.org/10.3102/0162373709352369

Slavin, R.E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research, 79* (2), 839-911. DOI: https://doi.org/10.3102/0034654308330968

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011) *Optimal Design Plus empirical evidence: Documentation for the "Optimal Design" software*. Retrieved May 27, 2016, from http://hlmsoft.net/od/od-manual-20111016-v300.pdf

Torgerson, C. J. (2007). The quality of systematic reviews of effectiveness in literacy learning in English: a "tertiary review. *Journal of Research in Reading, 32*(3), 287-315.

U.S. Department of Education (n.d.) *Investing in Innovation Fund (i3).* Retrieved May 27, 2016, from http://www2.ed.gov/programs/innovation/index.html?exp=0

West, M. R. (Feb 5, 2016). *From evidence-based programs to an evidence-based system: Opportunities under the Every Student Succeeds Act***.** Retrieved May 27, 2016, from http://www.brookings.edu/research/papers/2016/02/05-evidence-based-system-opportunities-under-essa-west

Wilde, E.T, & Hollister, R. (2002). *How close is close enough? Testing nonexperimental estimates of impact against experimental estimates of*

impact with education test scores as outcomes. Madison: University of Wisconsin—Madison, Institute for Research on Poverty. Available at: http://files.givewell.org/files/methods/Wilde%20and%20Hollister%202002.pdf