Abstract Title Page

Title:

Changes in the Precision of a Study from Planning Phase to Implementation Phase: Evidence from the First Wave of Group Randomized Trials Launched by the Institute of Education Sciences

Author(s):

Jessaca Spybrook, Monica Lininger, Anne Cullen

Abstract Body

Background / Context:

In the past decade, the actions of the federal government reflect a shift in policy towards experimental research as a means to produce reliable evidence of the effectiveness of educational programs and policies. This shift in policy is evident by the passing of a series of Congressional Acts beginning in 1998 that call for rigorous evaluations of educational programs. Rigorous has been defined as either experimental or high-quality quasi-experimental designs. A key force behind this movement is the Institute of Education Sciences (IES), the research division of the US Department of Education (DOE), which was established by The Education Sciences Reform Act of 2002. The mission of IES is to produce research that provides rigorous evidence to inform education policy and practice (http://ies.ed.gov/). More specifically, the goals of IES are to conduct scientific research, disseminate the results, promote the use of knowledge gained from scientifically based research, and translate education into an evidence-based field.

Although IES was established only 8 years ago, its impact on the field is clearly visible. In the year 2000, only 1 program evaluation funded by the US DOE included a randomized trial (Boruch, deMoya, and Snyder, 2002). Between 2002 and 2006, over 55 evaluations funded by the National Center for Education Research (NCER) and the National Center for Education Evaluation (NCEE) included a randomized trial (Spybrook, 2008) and between 2007 and 2008, NCER and NCEE combined to fund more than 80 evaluations that included a randomized trial (http://ies.ed.gov/). The dramatic increase in the number of experiments funded is unprecedented in education and represents a serious commitment on the part of IES to support research with the aim of translating education into an evidence-based field.

However, funding experiments is not enough to transform education into an evidence-based field. That is, the presence of an experiment does not guarantee high-quality evidence. In order to yield the high quality evidence desired by the shift in policy, the studies must be *well-designed* and *implemented* (Boruch, 1997; Boruch, DeMoya, & Synder, 2002; Cook, 2002; Valentine & Cooper, 2008). The 55 randomized trials that were funded by IES between 2002 and 2006 are now either in the field or complete, providing a critical opportunity to examine both the design and implementation of the trials. In a previous study (Spybrook & Raudenbush, 2009) examined issues related to the design of the studies. This study focused on the group randomized trials (GRT) funded by IES from 2002 to 2006, hereafter referred to as the first wave of experiments funded by IES. The sample was limited to the GRTs because of the natural clustering in the U.S. education system involving students nested within classrooms nested within schools nested within districts, and the fact that an intervention is typically administered at the classroom, school, or district level (Bloom, 2005; Boruch & Foley, 2000; Cook, 2005). The study examined the designs and precision of the planned studies.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this study is to extend the work of Spybrook and Raudenbush (2009) and examine how the research designs and sample sizes changed from the planning phase to the implementation phase in the first wave of studies funded by IES. We examine the impact of the changes in terms of the changes in the precision of the study from the planning phase to the implementation phase. We explore trends in the changes that occurred in order to inform the planning and implementation of future studies.

Sample/Data Collection and Analysis:

The sample for this study consists of GRTs funded by NCER between 2002 and 2006 and NCEE between 2001 and 2006. In the previous study, Spybrook & Raudenbush identified 68 possible GRTs from the IES website, 55 funded by NCER and 13 funded by NCEE. They obtained the funded proposals for these studies and screened them for inclusion. The criteria included that the study must include an experiment where groups were randomly assigned to the treatment condition and must test an intervention targeted toward children in pre-K through grade 12. They identified 46 NCER and 9 NCEE studies that met the inclusion criteria and coded the funded proposals for these studies to determine the planned features of the study design, or baseline data related to the experimental design and sample size.

To determine how much the planned research designs and sample sizes changed when the studies were implemented in the field, we collected data on these same studies after they entered the field. The follow-up data were primarily gathered via interviews with the Principal Investigators (PI) and other lead personnel of the projects. Supplemental data were gathered from annual reports and journal articles. We contacted all 46 NCER PIs and 9 NCEE PIs to schedule interviews. Thirty-three of the PIs agreed to interviews. One PI sent mid-year and end-of-year reports from the study. We obtained mid-year and end-of-year reports for 9 studies via the Freedom of Information Act (FOIA). Results from 41 studies are reported in this paper.[†]

The interview protocol for each PI drew from the baseline data for each individual study. This enabled direct comparisons between what was planned and what was implemented. For example, if a study planned to randomly assign 40 schools to either a new math curriculum or the current condition, then 40 schools would be the baseline and questions would focus on if, how, and why the number of schools was different from 40. All interviews were transcribed and then entered into NVivo8, a qualitative analysis software. Three researchers reviewed each transcript to identify the research design and sample sizes when each study was implemented in the field. Any discrepancies were discussed and resolved by revisiting the interview and if necessary, contacting the PI for clarification. We were also able to triangulate the information with end-of-year IES reports or journal articles.

Findings / Results:

We examined the research designs in the planning phase and implementation phase. Changes in the research design are defined as a difference in the design classification. For example, a study with students nested within classrooms within schools that planned to match schools prior to random assignment would be considered a 4-level multisite cluster randomized trial (MSCRT). However, if the researchers decided not to match the schools before randomizing, then the study would be considered a 3-level cluster randomized trial (3-level CRT). In the overwhelming majority of the studies, there was no major change in the research design. However, in 3 cases the design did change. In the first case, the planned design was a 4-level MSCRT with students nested within classrooms nested within schools and planned to block schools prior to random assignment. The design that was implemented was a 3-level CRT.

[†] We have sent another request via FOIA for the remaining 12 studies. They were not available at the time this proposal was written. We also determined that one study was not funded by NCER or NCEE and removed it from the sample.

Because of difficulties in recruiting, they were not able to secure all the schools prior to random assignment and instead recruited the schools over a 2 year period and did not do apriori blocking. The second design change was also a planned 4-level MSCRT. In this study, the researchers had students nested within classrooms nested within schools and planned to block schools prior to random assignment. However, prior to the start of the study they recognized that the study was underpowered and revised the design. The design that was implemented was a 3-level MSCRT in which schools became the blocks and classrooms were the unit of random assignment. The third design change was a planned 3-level CRT with students nested within teachers nested within supervisors. Recruiting and randomizing supervisors proved to be more challenging than teachers and thus the research team randomly assigned teachers instead of supervisors and thus had a 2-level CRT with students nested within teachers.

There are a few common features across the studies that experienced design changes. First, the three studies were funded by NCER. None of the NCEE studies exhibited changes in the experimental design from the planning phase to the implementation phase. Second, two of the three studies evaluated interventions targeting middle or high school students.

Next we examined changes in sample sizes, and particularly the number of clusters since this is the unit that drive statistical power. We categorize change in the number of clusters randomized in the following five categories; more that 10 percent loss, less than 10 percent loss, no loss, up to 10 percent gain, and more than 10 percent gain. Figure 1 shows the frequency and percent of studies in each of the 5 categories.

From Figure 1, it is clear that there were increases and decreases in the number of clusters randomized in the implementation phase. Across all studies, about 21 percent of the studies had a loss in the total number of clusters, with about 14 percent of them yielding a loss of more than 10 percent. However, about 50 percent of the studies increased the total number of clusters, with 15 percent of the studies increasing the number of clusters randomized by more than 10 percent. About 27 percent of the studies showed no change.

A comparison between the NCER and NCEE studies in terms of sample size changes is presented in Figure 2. A difference among the two centers is clear. Of those studies that decreased the number of clusters randomized, all of them were funded by NCER. This means that about 30 percent of the NCER studies had a loss in the total number of clusters randomized, with approximately 20 percent showing a loss greater than 10 percent. None of the studies funded by NCEE yielded a loss of clusters randomized from planning phase to implementation phase. The percentage of studies with no changes was also quite different among the two centers, with 50 percent of the NCEE studies showing no change and only 20 percent of the NCER studies revealing no change. Similar percentages showed increases in the total number of clusters across both funding centers. In general, there appears to be much more change in the number of clusters randomized in NCER studies than in NCEE studies.

Using the actual design, sample sizes, and best estimates of other design parameters, we estimated the minimum detectable effect size (MDES) for each study in the implementation phase. Details of the calculations are provided in the full paper. We compared the MDES at implementation phase to the MDES at planning phase based on Spybrook & Raudenbush (2009). Figure 3 displays the planning phase MDES and the implementation phase MDES. The MDES is a range in all cases because of the estimated parameters are ranges.

In studies 1 through 5, the MDES in the implementation phase is less precise than in the planning phase. These first five studies represent the studies with a loss of more than 10 percent of the total number of clusters. All studies were funded by NCER. Studies 6 through 29 represent

those with either less than 10 percent loss in the number of clusters, no change in the number of clusters, or less than 10 percent gain in the number of clusters. Only one range for the MDES is present for each of these studies because the planning phase MDES and implementation phase MDES are either exactly the same or the differences are negligible. Studies 22 through 29 represent those that added more than 10 percent of the total number of clusters. In these cases, the MDES at the implementation phase is more precise than at the planning phase, representing an improvement in the precision of the study.

As seen in Figure 3, a loss of 10 percent of the total number of clusters or a gain of 10 percent of the total number of clusters may yield a MDES at implementation phase that is very different than a MDES at planning phase. For example, in study 3, the MDES at the planning phase ranged from 0.19 to 0.37. However, at the implementation phase the MDES ranged from 0.37 to 0.71. If the expected effect was 0.20, then at implementation phase the study is no longer powered to detect the effect. The opposite phenomenon occurs on the gain side. For example, study 14 had a MDES range from 0.34 to 0.49 at the planning phase. At the implementation phase, the MDES ranged from 0.26 to 0.37. If the true difference between the treatment and control group is 0.30 standard deviations, the study would be adequately powered at the implementation phase after the additional clusters but would not have been adequately powered at the planning phase.

Conclusions/Implications:

The findings from this study suggest that changes in the research design were rare. In fact only 4 studies had changes in the research design. Changes in the number of clusters randomized were much more common, which resulted in changes in the precision of the studies. About 30 percent of the NCER studies showed a loss in the total number of clusters. On the positive side, almost 50 percent of the NCER studies showed a gain in the total number of clusters. Among NCEE studies, approximately half showed no changes in sample sizes and none of the studies revealed a loss. Clearly, changes in sample size were more likely in NCER studies.

There are several possible explanations for this pattern. An obvious explanation is that NCEE studies are run by contract firms with a lot of experience in designing and implementing large scale randomized trials in the field resulting in fewer changes. NCER grants are typically run by University researchers who may be less experienced in conducting randomized trials and may have fewer systems in place to support the effort resulting in more implementation obstacles. However, there are a few other procedural differences that may contribute this pattern.

First, 7 of the 8 NCEE studies, or about 88 percent, had a built-in planning year prior to implementation of the trial in the field. Only about 9 of the 30 NCER studies, or about 30 percent, had a built-in planning year. The planning year allows time to secure schools participation, build relationships with school personnel, and finalize the planning process for a smoother implementation. Without the planning year, there may be little time to finalize the plans for participation of schools and districts or recruit new schools if necessary.

The differences between those studies with and without a planning year are evident when we consider the changes in overall sample size from planning phase to implementation phase. Table 1 shows the percentage of studies in each of the 5 change categories that had a planning year or did not have a planning year. In both of the loss categories, the overwhelming majority of the studies did not have a planning year. In the no change group, there was more even distribution among those with and without a planning year. Those with small gains, between 0

and 10 percent, look similar to the loss categories in that about two-thirds of the studies did not have a planning year. However, the category with the greatest gains reveals the opposite pattern. Of those studies that gained more than 10 percent of the total sample size, 67 percent had a built in planning year. This pattern suggests that perhaps with the planning year, researchers are able to better recruit and add additional clusters.

A second difference related to the planning year is in the recruitment process. In order to secure funding for a NCER grant, letters of support should be included from schools and districts that plan to participate in the study. However, letters of support are sent prior to the PI knowing whether or not the grant will be secured, which means that school and district personnel may be very willing to sign the letter, since they are not sure it will ever come to fruition. Typically at least 8 months lapses from the time a grant is submitted to the time a grant is awarded. During that time schools may undergo new leadership or put new programs in place and when researchers recontact them to begin the study, the school or district may no longer be interested in participating or may not have been serious in the initial phase. If there is no planning year, then there is usually little time for researchers to try to recruit new schools or districts which may results in a loss in the number of clusters. The recruitment process is different with NCEE contracts. Schools and districts are not recruited until after the contract is secured, thus when a school or district commits to the study, it is known that the study has been funded and will move forward. This makes it much less likely that schools or districts will back out of the study and reduce the sample size. The planning year also makes this type of recruitment strategy possible.

Overall, the findings from this study are positive in that they suggest that GRTs can be implemented in the field, and in many cases without major changes to the planned design, sample sizes, and precision. Thus if future studies are designed with adequate levels of precision, it is possible that they will be implemented in such a way to uphold that level of precision and have the potential to contribute to a base on evidence on which to base policy and practice in education. However, particularly within NCER, there were also cases in which the sample size decreased and hence the precision decreased. One way to minimize these negative changes might be to include a built-in planning year to help ensure fewer implementation problems. For example, if a grant is funded and the schools that were secured prior to funding decide not to participate, the built-in planning year would enable the researchers to recruit new schools without postponing the project. The planning year would also allow the researchers the opportunity to recruit additional schools if desired.

Appendix A. References

Bloom, H.S. (2005). Randomizing groups to evaluate place-based programs. In H.S. Bloom (Ed.), *Learning More From Social Experiments*: *Evolving Analytic Approaches* (pp. 115-172). New York: Russell Sage Foundation.

Boruch, R.F (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage Publications.

Boruch, R.F., DeMoya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In *Evidence Matters: Randomized Field Trials in Education Research*. Edited by F. Mosteller & R. Boruch. Washington, D.C: Brookings Institution Press, 50-79.

Boruch, R. F., & Foley, E.(2000). The honestly experimental society. In *Validity and Social Experiments: Donald Campbell's Legacy*. Edited by L. Bickman. Thousand Oaks, CA: Sage Publications, 193-239.

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. *Educational Evaluation and Policy Analysis*, 24(3), 175-199.

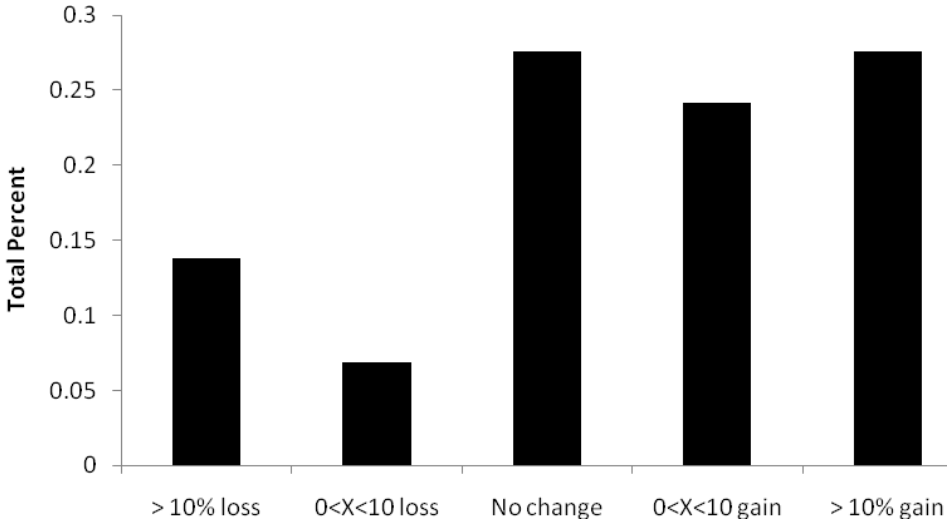
Cook, T.D. (2005). Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. *The Annals of American Academy of Political and Social Science,* 599.

Spybrook, J. (2008). Are power analyses reported with adequate detail: Findings from the first wave of group randomized trials funded by the institute of education sciences. *Journal of Research on Educational Effectiveness*, 1(3).

Spybrook, J. & Raudenbush, S. W. (2009). An examination of the prevision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis*, 31(3), 298-318.

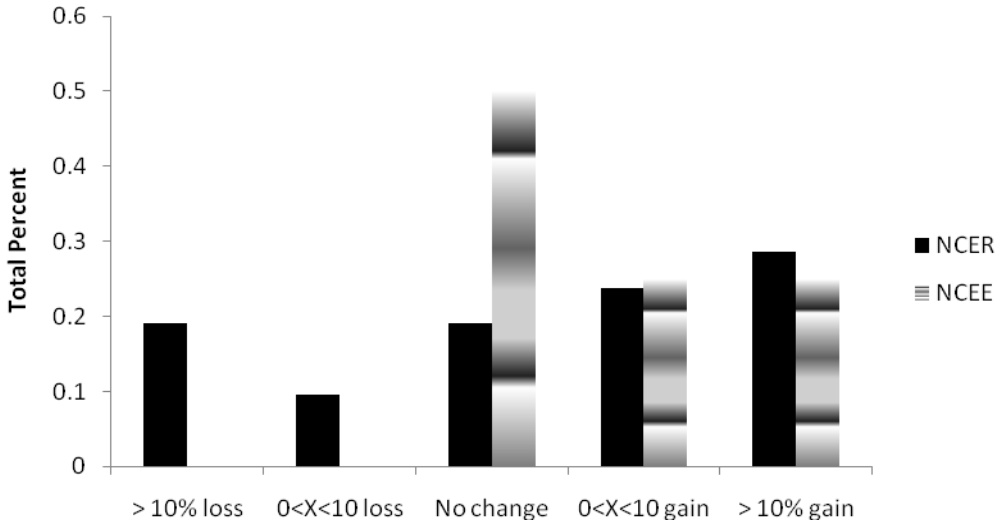
Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods,* 13(2), 130-149.

Appendix B. Tables and Figures



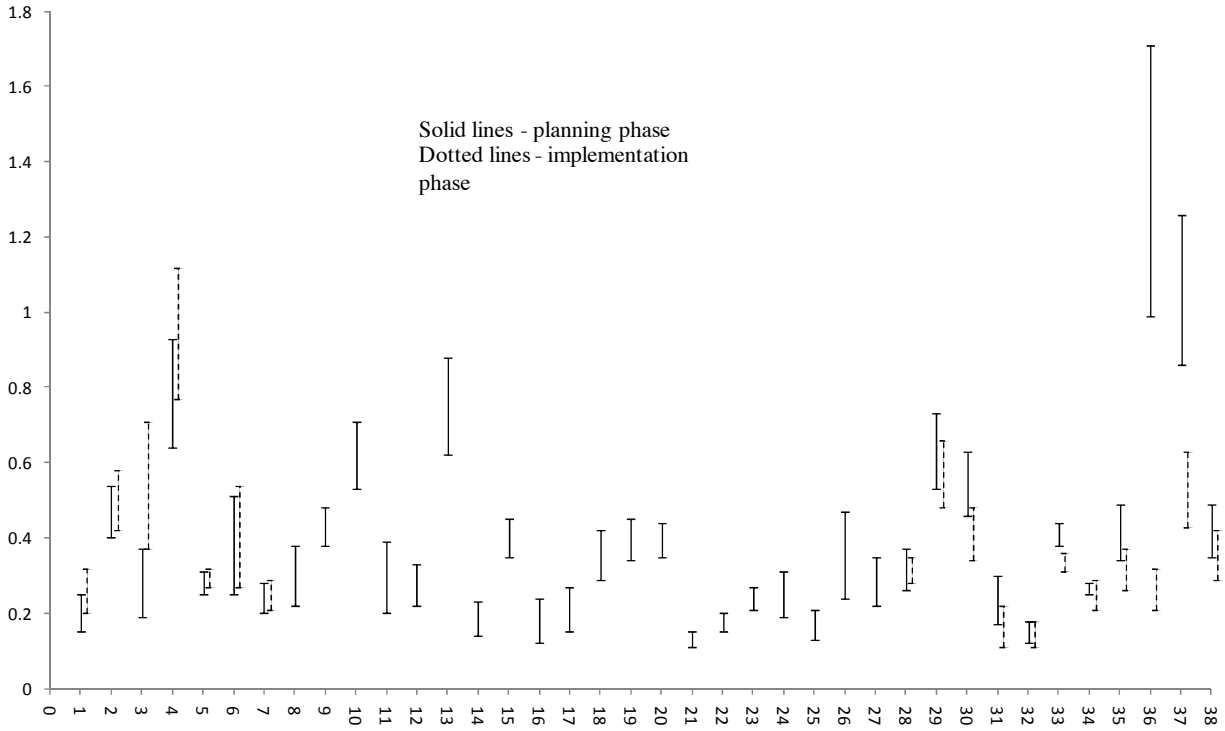
Note. Two studies are not included because they are still in the recruitment phase. An additional study is not included because the sample sizes at implementation are still unknown.

Figure 1. Frequency of changes in the number of clusters randomized for the entire sample.



Note. Two studies are not included because they are still in the recruitment phase. An additional study is not included because the sample sizes at implementation are unknown.

Figure 2. Frequency of changes in the number of clusters randomized by funding agency.



Note. Two studies are not included because they are still in the recruitment phase. An additional study is not included because the sample sizes at implementation are still unknown.

Figure 3. A comparison of the MDES in the planning phase and the implementation phase.

Table 1. Number of studies with and within a planning year categorized by percent of change in overall sample size.

	Planning Year	No Planning Year
Greater than 10 percent loss	1 (0.25)	3 (0.75)
Between 0 and 10 percent loss	2 (0.33)	4 (0.67)
No change	4 (0.40)	6 (0.60)
Between 0 and 10 percent gain	3 (0.33)	6 (0.67)
Greater than 10 percent gain	6 (0.67)	3 (0.33)