Abstract Title Page

Title: The Challenge of Authenticity in Scale-Up Effectiveness Trials

Author(s): John F. Pane, Beth Ann Griffin, Daniel F. McCaffrey, and Rita Karam
          RAND Corporation
Background
A scale-up effectiveness trial seeks to determine whether an intervention with strong prior evidence of efficacy is effective at a broad scale. Two key features of effectiveness studies are 1) implementation in a diverse sample of schools, classrooms, or students, to ensure appropriate generalizability of findings, and 2) implementation under conditions that would typically exist if schools or districts were adopting the intervention on their own – that is, not as part of a research study (Institute of Education Sciences, 2010, p. 72). Addressing generalizability is relatively straightforward—recruit from a sufficient variety of sites. The main difficulties or limitations in this endeavor are the willingness of schools to participate, and the costs of including them. Some decisions about authenticity are also straightforward. For example, randomization of schools is likely to be more authentic than randomization within schools, because within-school randomization can be very disruptive to the operations that are normally under the control of schools, such as the assignments of teachers and students to curricula and classrooms. However, in many studies, there are additional tensions between the demands of experimental methods and authenticity. In essence, the challenge is for researchers to retain control over the sample for the duration of the study, without forcing schools to make implementation decisions they would not have made under an analogous realistic situation such as the adoption of a new curriculum. These tensions around authentic implementation are the focus of this paper.

Purpose
In the context of a randomized controlled trial effectiveness study of an algebra curriculum, this paper discusses the tension between designing a rigorous experiment and enabling authentic implementation. We then focus on one particular research design choice, made in the name of authenticity, which did not exert sufficient control over the samples in the study, resulting in pretreatment group differences that threaten the validity of the experiment and require the application of quasi-experimental analytic techniques to correct for potential bias in the estimation of the effectiveness of the intervention.

Setting
The effectiveness study was conducted in 147 schools in 52 school districts in seven states. The sites include Houston, Texas, Bridgeport, Connecticut, Trenton, New Jersey, Mobile, Alabama, suburban districts near Detroit, Michigan, rural districts in Kentucky, and districts throughout Louisiana. Participating schools include urban, suburban, and rural public schools, and some Catholic Diocese parochial schools. Each school participated for two years. The study includes both middle schools and high schools, and the study was designed to analyze these separately because the populations of students taking algebra in middle schools (grade 8 or earlier) is generally higher-achieving than the population of students taking algebra in high schools (grades 9-12) and the curriculum might have different effects in these populations. All sites except Mobile participated in both the middle school and high school arms of the study.

Participants
Table 1 shows the number of participating students by state and grade level. In total, more than 24,000 students in grades 6 through 12 participated in the study, with the majority (72%) in the high school arm of the study. Among the high school sample, 89% of the participants were 9th-graders. Among the middle school sample, more than 99% were in 8th grade.

(Please insert Table 1 here)
Significance
This paper offers practical advice on the design of large-scale randomized controlled trials of the effectiveness educational interventions, where authenticity is a concern. To date, there have been few such studies conducted, and the experiences from the current study can offer useful guidance to designers and funders of future studies. The specifics observed in this study of algebra may be typical for a broad range of middle and high school courses.

In particular, Table 1 demonstrates a consistent pattern across states that students who enroll in algebra come from a broad range of grade levels. Underlying this pattern is the fact that schools typically do not have firm rules about algebra course taking. Generally, the majority of students take algebra in 9th grade, more advanced students take algebra in 8th grade or earlier, and lower achieving students might enroll in 10th grade or later. Schools also sometimes spread the curriculum over two years for low-performing students. Moreover, the definitions of these groups (low achieving, mainstream, or advanced/honors/gifted) are not precisely specified and may drift over time, and the placement of individuals into these groups is often left to the discretion of teachers, principals, or guidance counselors.

To achieve authenticity in the study, we sought to allow schools to retain their significant discretion regarding when to enroll students in algebra and the resulting population of students taking algebra at any particular time. However, this discretion, left unchecked, would enable schools to change algebra enrollment patterns in response to the their randomly assigned experimental condition, i.e., after the treatment is assigned. For instance, the treatment group of a study might view the adoption of the new algebra curriculum as an opportunity to change algebra enrollment patterns. Unfortunately, such enrollment changes in response to assigned experimental condition subvert the control that the researcher is trying to establish through randomization. Randomization is designed to ensure that the pre-existing characteristics of the students in the treatment and control groups are unrelated to experimental conditions, but the authentic discretion held by schools could lead to systematic differences between groups.

Research Design
The study used a pair-matched cluster randomized design. Schools within states and school type (middle or high school) were matched into pairs on a number of criteria, including school-level variables and the achievement profile of participating students, as specified by schools when they enrolled in the study. Schools were randomized in the spring prior the first year of implementation, and implementation continued for two academic years. The gap between randomization and implementation was needed for treatment schools to prepare for the new curriculum, including the training of teachers. Moreover, two-year participation was an important aspect of the study, in order to examine treatment effects among teachers who were more experienced with the curriculum in their second year of use.

To explore the design tradeoffs between authenticity and experimental control over the treatment and control group samples, we considered a variety of options. For each, we assessed how the study might evolve from the time of randomization through the duration of the study. Table 2 summarizes some of the options we considered, along with the level of control each afforded to the research, the authenticity of the design, and other feasibility concerns.

(Please insert Table 2 here)

Ultimately, we selected the first option listed and described it in our grant proposal to IES. This option required schools, prior to randomization, to specify schema for selecting the students who would participate for both years. The schema identified a set of criteria schools would apply to determine the students who would participate each year, such as all algebra classes or the
inclusion or exclusion of specific classes by ability level or the inclusion or exclusion of specific classes taught by certain teachers. Figure 1 shows an example of the portion of the study enrollment form where schools specified the schema. This method grants schools discretion over algebra enrollment, which has good authenticity. During the design phase of the study, we judged the level of control afforded by the schema sufficient because schools would specify the exact rules for selecting the types of students who would participate even though they would not specify the exact students. Moreover, we could include the schema as one of the characteristics used to create the matched pairs for randomization, helping to ensure that similar types of students would be participating in the treatment and control schools.

(Please insert Figure 1 here)

However, in hindsight, we now understand that the authenticity afforded by the schema gave schools more than the desired amount of flexibility in enrollment, reducing our control over the sample. For example, schools designating that all algebra classes would participate still had discretion to enroll more or fewer students in algebra than originally intended, changing the nature of the population of students taking algebra.

Other options considered, shown in Table 2, included the possibility of requiring schools to specify the precise student sample for both years of the study prior to randomization. This option affords the researcher a higher level of control over the sample than the option we selected; however, the level of authenticity is lower than the option we adopted. Although schools retain the usual discretion over student enrollment in algebra, forcing them to make the determination for the second-year cohort of students 18 months in advance would not be authentic. In addition, this design could result in very high levels of attrition because schools may not have accurate knowledge of their enrollment so far in advance. Students they specify to be in the study might move or drop out of school during the intervening period. Similarly, we also considered having the study require schools to enroll all students in a certain grade level to take algebra and be included in the study. This would have afforded the study a high level of control over the sample, but would have had very poor authenticity. Few schools would have been willing to participate in the study under these conditions. A final option we considered which would have granted both good control and excellent authenticity was to allow the schools to assign students to algebra classes at their discretion but to administer algebra pre- and posttests to all students in the school each year, regardless of whether they were enrolled in algebra at the time. This design would include all students in the school in the sample, and the study sample would not be affected by any discretion exercised by schools regarding students enrolled in algebra. Changes in algebra achievement through changes in enrollment would be incorporated in estimates of schoolwide treatment effects. Thus, the authenticity is excellent in this design. Unfortunately, the cost of administering algebra tests to the entire school populations can be prohibitive, and schools might have been less willing to participate if such large-scale testing were necessary, particularly because it may seem incongruent that the study would need to test students not enrolled in algebra.

Data Collection and Analysis
About three weeks after the start of each school year (or semester), the study administered as a pretest the CTB/McGraw-Hill Acuity Algebra Readiness Exam, a 40-item assessment designed for students who have completed grades 6 through 11 to assess the skills necessary for successful performance in algebra. Scores on this assessment were standardized to have a mean of zero and standard deviation of one, and used to assess group balance in the experiment. We used a hierarchical linear model that accounted for the clustering of students in classes and schools, and which included fixed effects for randomization pairs.
Findings
This analysis estimated that the treatment group scored 0.24 standard deviation units lower than the control group \((p < 0.001)\) on the pretest. The difference was larger in middle schools than in high schools \((0.35 \text{ vs. } 0.15)\), was present in both the first and second years of school participation \((0.28 \text{ vs. } 0.20)\), and was consistently present across all seven states, although not always statistically significant. Figure 2 shows the distributions of the standardized pretest scores for the treatment and control groups. Similar pretreatment differences were observed on prior-year state mathematics assessments. These pretreatment differences are large enough to threaten the validity of the study. For example, for studies to be classified as providing rigorous evidence about an intervention, the Institute of Education Sciences’ What Works Clearinghouse requires pretreatment differences between the treatment and control group to be no more than 0.25 standard deviation units in absolute value (What Works Clearinghouse, 2008).

(Please insert Figure 2 here)

Conclusions
The large and consistent pretest differences between groups suggest that there were systematic shifts in the samples after randomization. Here we consider some possible explanations.

1. The pretest was administered several weeks into the school year or semester. If there was a substantial early treatment effect, the pretest might have captured this. This explanation is unlikely because the group differences are also apparent on prior-year state test scores.

2. Treatment schools may have been enthusiastic about implementing a new curriculum and decided to increase enrollment of lower-achieving students in algebra. However, we find that the treatment group has 14% fewer students than the control group, making this explanation unlikely.

3. As administrators and teachers in treatment schools learned more about the new algebra curriculum, they may have decided it was inappropriate for higher-achieving students. This explanation is consistent with observing fewer students in the treatment group.

The imbalance between groups on pretest scores means that simple treatment versus control group analysis may produce a biased estimate of the treatment’s effectiveness. In light of this, we plan to run propensity score models that attempt to adjust for group imbalances on the pretest. We will attempt to balance pretest scores and other baseline covariates within each randomization pair to achieve overall balance. Results from these different approaches can then be compared to make reasoned judgments about the effect of the new algebra curriculum.

These results suggest that design option 2 in Table 2 may have resulted in an experiment with greater validity while affording excellent authenticity. In that option, algebra pre- and post-tests would have been administered to all students in the schools, regardless whether they were enrolled in algebra. As discussed above, this option would have been substantially more costly, and may have made schools less willing to participate in the research because the testing could be very disruptive. This approach can also have implications for the power of the study, requiring a larger sample and thus another potential driver of increased cost. It is unlikely such a design would have fit within the budgetary limits of the IES grant program that funds this study.

In conclusion, striving for authenticity in effectiveness experiments in education can pose serious challenges to designing the study for strong validity, and satisfying both concerns can drive up the size and cost of the study to the detriment of feasibility.
Appendices

Appendix A. References


Appendix B. Tables and Figures

Table 1: Number of participating students by state and grade level.

<table>
<thead>
<tr>
<th>State</th>
<th>Middle School Study</th>
<th>High School Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade</td>
<td>6</td>
</tr>
<tr>
<td>CT</td>
<td></td>
<td>102</td>
</tr>
<tr>
<td>KY</td>
<td>3</td>
<td>49</td>
</tr>
<tr>
<td>LA</td>
<td>945</td>
<td></td>
</tr>
<tr>
<td>MI</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>NJ</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>TX</td>
<td>1</td>
<td>3,286</td>
</tr>
<tr>
<td>AL</td>
<td></td>
<td>871</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 2: Various options for controlling the study sample, authenticity, and feasibility

<table>
<thead>
<tr>
<th>Control Mechanism</th>
<th>Level of Control</th>
<th>Authenticity</th>
<th>Feasibility Concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Require schools to specify <em>schema</em> for student sample prior to randomization</td>
<td>Moderate</td>
<td>Good</td>
<td>Ability to monitor or enforce is limited</td>
</tr>
<tr>
<td>Require schools to specify the precise student sample for both study years prior to randomization</td>
<td>High</td>
<td>Medium</td>
<td>Unacceptably high levels of attrition are likely</td>
</tr>
<tr>
<td>Specify a grade in which all students take algebra</td>
<td>High</td>
<td>Poor</td>
<td>Few schools might agree to participate</td>
</tr>
<tr>
<td>Administer algebra pre-post tests to all students in school</td>
<td>High</td>
<td>Excellent</td>
<td>High cost; schools might object to disruption of testing</td>
</tr>
</tbody>
</table>
Figure 1: At enrollment in the study, schools specified a schema for determining student participation.

**Designation of Study Participants**
Please select one of the following rules to determine which students, classes, and teachers in your school will participate in the study during the 2007-08 and 2008-09 academic years. This rule will apply in both years. Any other algebra classes in your school will not participate in the study.

- [ ] All algebra classes.

- [ ] All algebra classes except ______________________________ (specify gifted, honors, remedial, etc.)

- [ ] All algebra classes taught by ______________________________ (specify teachers)
  The students in these classes are (mark all that apply):
  - [ ] Low performing
  - [ ] Average
  - [ ] High performing

Please estimate how many classes and students you have designated to participate:

Estimated number of participating classes: ______________

Estimated total number of participating students: ______________
Figure 2: Pretest distributions by group.