Empirical Education
EMPOWERING EDUCATORS FOR EVIDENCE-BASED DECISIONS

The Unvarnished Report of Conducting an RCT of a Statewide STEM Program in the Process of Scaling Up[1]

Denis Newman

Andrew P. Jaciw

Empirical Education Inc.

October 31, 2014

We were responsible for a large scale randomized control trial to evaluate the effectiveness of the Alabama Math Science and Technology Initiative (AMSTI). We conducted this as part of our contract with the Regional Education Laboratory, Southeast, which is funded by the U.S. Department of Education's Institute of Education Sciences (IES), which funds the regional labs (RELs). The official report (Newman et al., 2012[2]), released by IES after considerable technical review, editing, and formatting, is what we call the varnished version. This chapter delves into a number of the nuts and bolts issues and lessons-learned in conducting this study and provides an unvarnished version of the story.

Our study has provided a strong example of taking advantage of a planned scale-up of a program by an education agency to conduct an RCT.[3] Alabama was expanding AMSTI to new schools and we were able to work with them to randomize applicants. We start with a short summary of the varnished version. We then provide five lessons we learned from processes that

---

[1] Chapter in preparation for in R. Maynard, L.V. Hedges, & S. Wachter (Eds.), *Teaching Cases on Designing and Fielding Experimental Design Evaluations in Education* (provisional title).

[2] Available at http://www.ies.ed.gov/pubsearch/pubsinfo.asp?pubid=REL20124008. The Empirical Education team led the design, operations, analysis, and reporting of the impact and related exploratory analyses. We want in particular to recognize Boya Ma who conducted, and oversaw other analysts, in conducting all the analyses under Jaciw's direction. Steve Bell provided guidance for the analysis especially of the two-year impact and co-wrote the chapter on that topic. A team from AED was responsible for data collection and reporting of program implementation in a chapter of the report. We are grateful to Herb Turner, Fatih Unlu, and, particularly to Rob Olsen, for technical review and advice. SERVE Center at UNC Greensboro, particularly Pam Finney, provided a communication channel with IES reviewers and program officer.

[3] The study is used as the exemplar of a "Scale-up Study" in the IES/NSF document "Common Guidelines for Education Research and Development" (http://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf).

arose behind the scenes and worth considering by others when conducting similar studies. These include how the research questions evolved, how the approach to randomization can help working with a local agency, reasons for difficulty in maintaining a control group embargo when working in the context of a locally-funded scale-up, issues in avoiding bias in the impact model selection, and what the local agency should make of the results of an RCT.

Short Summary of the AMSTI study

We start with a very brief summary of the varnished report.

Study Characteristics

The Alabama Math, Science, Technology Initiative (AMSTI) was developed and implemented by the Alabama State Department of Education (ALSDE), which also hosted this multi-year RCT. The research was able to take advantage of the further rollout of the program, which for the previous four years had already been implemented in 99 schools in the state. This study was unusual in representing the scale-up of a locally developed intervention for which the local decision-makers (in this case ALSDE and the state legislators) were the primary audience for the study.

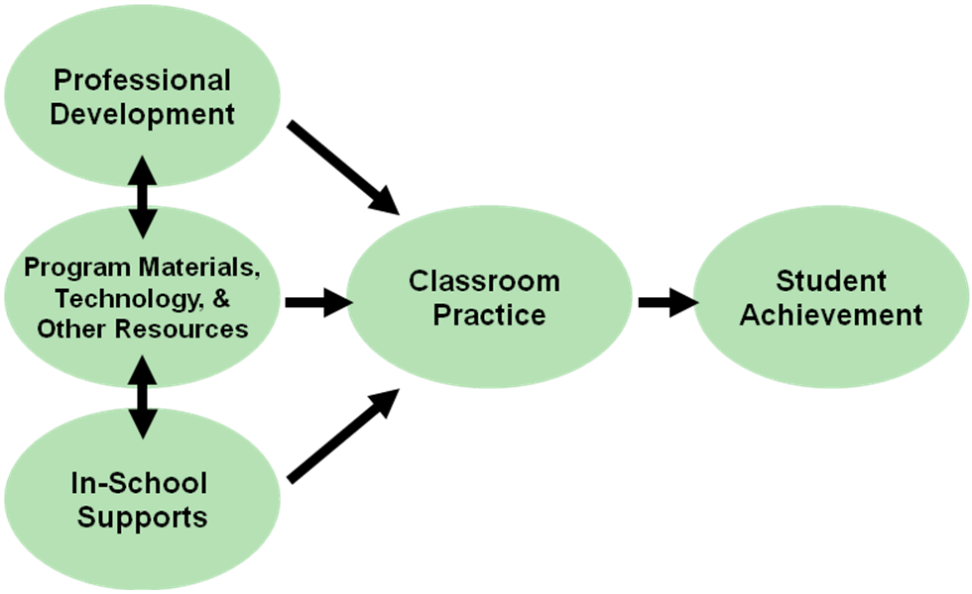

FIGURE 1: AMSTI LOGIC MODEL

The AMSTI theory of action, illustrated in Figure 1, posits that in order to improve student achievement, teacher "classroom practices" should include inquiry-based hands-on activities that promote higher-order thinking skills. The three components of the program that are intended to foster this type of instruction are (1) comprehensive professional development delivered through a 10-day summer institute (which is grade and content specific) and follow-up training during the school year; (2) access to program materials, manipulatives, and technology needed to deliver hands-on, inquiry-based instruction; and (3) in-school support by

AMSTI lead teachers and site specialists who offer mentoring and coaching for instruction. The full program is delivered over the course of two years. In each region, the AMSTI staff at ALSDE worked with managers and trainers at technical assistance sites housed at local universities and colleges.

Like the treatment schools, the control schools had applied for participation in AMSTI but were randomly assigned to a one year delay. In the first year, these schools used their regular instruction programs for mathematics and science. Programs in use met the objectives specified in the Alabama Course of Study and were based on comprehensive and supplementary texts recommended by the Alabama State Textbook Committee. More than 50 sets of curriculum materials have been approved for mathematics and science. In their second year, the control schools joined AMSTI providing a comparison between schools in their first year of implementation with those in their second year.

The evaluation was conducted in 82 schools within five regions in Alabama. A majority of the schools in the sample were in rural areas. Students in the study were in grades 4-8. Approximately 49% of the student sample consisted of racial/ethnic minority, 98% were proficient in English, and 64% were eligible for free or reduced priced lunch.

Study Design and Analysis

To participate in the study, a school had to house at least one grade from grades 4 through 8, and at least 80 percent of a school's mathematics and science teachers had to agree to participate. The power analysis called for 66 schools to detect impacts on student achievement of magnitude .20 standard deviations or higher with 80% power. Eighty-two schools were recruited to allow for possible attrition. Approximately 780 teachers and 30,000 students participated in the study. (Impacts on math and reading were assessed in grades 4-8 combined, and on science, in grades 5 and 7 combined.) From the eligible schools that applied to the program, 41 matched pairs were purposively selected so that on the whole the sample resembled the population of schools in the regions involved.

The study took advantage of ALSDE's rollout of AMSTI to specific regions during the study years. Because in the first year there were fewer schools that met eligibility criteria than required, the experiment combined two "sub-experiments", one starting in 2006 and the other starting in 2007. In sub-experiment 1, the first set of 40 schools (within three regional AMSTI sites) was randomized in 2006. In sub-experiment 2, a second set of 42 schools (within two regional AMSTI sites) was randomized in 2007. To estimate the average impacts of AMSTI on achievement, outcomes data from both sub-experiments were pooled and analyzed together.

Outcome Measures

The primary outcome measure for students was the Stanford Achievement Test Tenth Edition (SAT 10) in mathematics problem solving, science, and reading. The SAT 10 is a norm-referenced test. During the study years, all schools in Alabama administered the SAT 10 each April to comply with state accountability requirements. The mathematics and reading subtests are required in grades 3–8. The science subtest is required in grades 5 and 7.

For teachers, the main outcomes were responses to questions asked through four web-based surveys concerning the use of active learning instructional strategies, content knowledge and student engagement in math and science. These data were collected each study year from teachers in both the AMSTI and control conditions. Response rates ranged from 83 to 96 percent. For the measure of active learning instruction, teachers reported time spent on 1) inquiry-based instruction; 2) hands-on instruction; and 3) instruction engaging students in higher-order thinking skills. For the scale measuring instructional strategies for active learning, Cronbach's alpha was .89 for AMSTI schools and .78 for control schools in mathematics and .88 for AMSTI schools and .92 for control schools in science.

Analytic Approach

Two-level hierarchical linear models (Raudenbush & Bryk, 2002) were used to estimate regression-adjusted impacts of AMSTI on student and teacher outcomes. Students (or teachers) were modeled at Level 1 and schools at Level 2. Dummy variables were used to model effects of matched pairs, and a school-level random effect accounted for deviations in individual school performance from the grand mean of performance. Covariates were modeled at both levels.

Covariates, including the pretest, were used in the impact models to increase the precision of effect estimates. Cases with missing posttests were listwise deleted. Missing values for covariates were addressed using the dummy variable method (Puma, Olsen, Bell, & Price, 2009).

For analyses involving student outcomes after one year, balance between conditions was achieved on all teacher-level characteristics, and all student-level characteristics for math and all but one characteristic (gender) for science. The pretests were balanced for both subject areas.

Levels of attrition in the study were low. Only one control school left the study but the district was able to continue providing student-level data for analysis. No schools were lost in the analysis of one-year impacts on achievement in math. Only one school was lost in the analysis of one-year impacts in science. Rates of overall attrition at the student level were 4.7% in math and 7.8% in science. Rates of differential attrition at the student level were 0.1% and 2.3% in math and science, respectively.

Results

The findings of the RCT, as described in the final report are summarized as follows:

The standardized effect sizes for impacts on use of active learning instructional strategies after one year were .47 ($p < .01$) for mathematics classrooms and .32 ($p < .01$) for science classrooms. Other findings for teachers included: AMSTI mathematics and science teachers were not more likely to report higher levels of content knowledge than were their control counterparts; and AMSTI mathematics and science teachers were more likely to report higher levels of student engagement than their control counterparts.

The standardized effect sizes for impacts on achievement *after one year* were .05 ($p < .01$) for math problem solving, .05 ($p = .09$) for science, and .06 for reading ($p < .01$). The standardized effect sizes for impacts on achievement after two years were .10 ($p = .03$) for math problem solving and .13 ($p = .04$) for science. A differential impact after one year favoring White students of .08

effect size units was observed for reading. Minority status did not moderate impacts after one year on math or science. None of the other moderating characteristics (pretest achievement, socioeconomic status, gender) were associated with differential impact after one year.

Introduction: The AMSTI Program and Their Research Questions

We now turn to the unvarnished story, addressing issues that were not part of the final report but were essential to the construction of the scientific findings.

The experiment we describe was undertaken in the context of a Regional Education Laboratory (REL) that had the mission of addressing questions of relevance to the states and districts in the southeast region. In this case, the Alabama State Department of Education (ALSDE) had for the previous four years been rolling out AMSTI and annually had to make the case to the legislature for funding to continue the expansion, or even the maintenance of the program.

TABLE 1: TOTAL SCHOOLS ADDED EACH YEAR BY REGION

	Years												Total
Regional Center:	02	03	04	05	06	07	08	09	10	11	12	13	
Athens State U					4	2	13	10	0	0	7	2	32
Auburn U						10	30	7	0	0	7	6	60
Jacksonville State U						32	13	11	0	0	11	8	64
Troy U					8	18	21	11	0	0	0	3	61
U of Alabama in Birmingham					3	10	25	1	3	0	32	4	75
U of Alabama in Huntsville	20	12	11	9	13	25	20	1	0	1	0	1	113
U of Alabama and U of West Alabama					17	17	31	3	1	4	1	0	74
U of Montevallo					13	10	16	0	5	0	5	2	51
U of North Alabama		10	12	12	16	11	13	5	0	2	0	7	88
U of South Alabama			10	12	19	10	16	7	0	0	1	0	75
Wallace C C-Selma and Alabama State U						23	17	6	3	0	0	0	49
Total (Does not include schools closed due to merger)	20	22	33	33	86	155	215	62	12	7	64	33	742

Table 1 shows that when our experiment began in the 2006-2007 school year, it was the beginning of a period of substantial scale-up of AMSTI. ALSDE had commissioned some studies in previous years, but while those studies showed positive results, they were not conducted with the rigor and imprimatur of the federal research agency that the REL program was promising. The staff of AMSTI was confident that the program was having an impact but getting some solid evidence would make it easier to convince the lawmakers to continue expansion of the program.

The Empirical Education team that took this project on was at the same time conducting research under an IES grant[4] exploring how local education agencies can use small-scale randomized control trials to help them in decisions about new programs they are piloting (Newman, 2008). While the RCTs conducted under the grant were smaller in scale (we typically worked with a group of teachers volunteering to pilot a new program within a single district) we recognized many of the same characteristics in the study of AMSTI where the scale was much larger—over multiple years we would have 82 schools throughout the state in the study. The key similarity was that the education agency, whether at the district or state level, was looking for rigorous evidence for the program as it was being rolled out in the jurisdiction—the "subjects" were drawn from the same local setting that had the policy questions to be answered.

From this perspective, the local agency itself was also the primary stakeholder and this led to some of the distinctive characteristics and challenges we address in this chapter. For example, recruiting into the RCT was inherent in the scaling up of the program—schools weren't volunteering to be in the study in the normal sense; rather, they were interested in being in the program and were randomized to receive AMSTI earlier or later. Since there were more schools interested than slots available in a given year (of the 101 applicants in the regions where the RCT was to take place, the first year only 20 were admitted to AMSTI in the three regions that agreed to be part of the experiment) we were able to use a wait-list design and randomize half the schools to serve as controls until the following year. Importantly, because the primary stakeholders who could most directly use evidence from the RCT were from the agency participating in the experiment, getting buy-in to the rigorous methods necessary for valid findings was more a matter of explaining the value to them of the method than of offering extrinsic incentives. The reward to the participants if valid positive results were found, whether AMSTI program staff, district administrators, or teachers, would be positive votes in the legislature to keep AMSTI growing.

ALSDE's direct interest in the research findings and the embedding of the randomized trial into the multi-year scale-up of AMSTI, had ramifications for many aspects of the research process, as we discuss in this chapter. In some cases we found that the frame of reference of a research project that addressed the needs of a local agency testing its own program were different from the frame of reference of the IES program, aimed at producing rigorous research for its national audience. The final, "varnished" report released by IES (Newman et al., 2012) can be viewed as

[4] Grant # R305E040031 to Empirical Education Inc. from the National Center for Education Research (IES)

representing a message to the national audience, rather than a report to ALSDE. We find lessons to be learned from a tension between the perspective and concerns of the federal agency funding the work, and the local participants who had a more direct stake in the outcomes. We examine how the research questions, the approach to engaging participants, some of the analytical methods, and the choices about reporting are conditioned by this tension.

Lesson 1: Research Questions Evolve

A major agenda for IES during this period was to develop technical approaches that would standardize a level of rigor across a large number of RCTs. These standards were developed and applied to the AMSTI RCT already in progress. Taking advantage of these technical advances made our analyses much stronger than they otherwise would have been. The version of the research questions as they appear in the final report reflects the evolution of technical knowledge as well as the formation of policy regarding standards for reporting federally funded education research. As a result, the emphasis of the report may have had less value for the local participant stakeholders for whom the original framing of the research questions were more balanced. The final framing of the research questions, as emphasized in the report, represents a specific prioritization of methods and results. The questions are shifted somewhat from their original framing that placed a more even emphasis on a wider set of issues. Applications of standards for rigor in research narrowed their scope but also may have increased their specificity.

The research questions that were listed in the initial proposal were different from those listed in the final IES report. The questions in the final report represented the culmination of lengthy dialogue between the research team and a variety of stakeholders including the state officials who eventually might be using the evidence, expert advisors concerned with scientific standards (which evolved over the course of the trial) and staff at IES who had methodological and policy priorities and who were responsible for securing reliable results within the timeframe of the contract period. We find a nuts and bolts issue in a tension between the goals of the local stakeholders and the federal agency funding the study. For both players there was a desire for rigorous evidence of overall efficacy but for the stakeholders there was a need for the evidence sooner rather than later and, we argue, a use for information that could support formative improvements of AMSTI.

Table 2 presents the earliest version of the research questions. This is taken from a draft of the proposal to IES that was eventually funded to conduct the study. The proposal was prepared in consultation with AMSTI staff and the early drafts of the study plan were based on dialogue with ALSDE. The discussions took into consideration the nature of the intervention, its implementation, the mechanism for randomization, the availability of data for analysis, and importantly, what ALSDE needed to know from the study in order to make the case for getting the program reauthorized by the state legislature.

TABLE 2: INITIAL VERSION OF THE RESEARCH QUESTIONS

What is AMSTI's impact on:
math and science achievement
proficiency in other subjects, such as reading and writing
specific subgroups of students (e.g., gender, race/ethnicity, economic background)
teacher classroom practices
teacher professional outcomes (i.e., retention, promotion)
the ways in which schools utilized their technological resources

The original study plan did not stress a particular priority for the research questions. While they were numbered and began with impact on students, there was no strict division between primary and secondary. Rather, we considered the whole set of research questions as important for understanding the efficacy of the program, the conditions for observing impact, and the mechanisms leading to effects. For example, we understood that differential effectiveness on policy relevant variables like minority status, poverty or gender, could be important for understanding the impact in terms of achievement gaps in math and science as well as potentially providing pointers to the developers on ways to improve the program. The decision to look at the impact on reading and writing was based on the AMSTI program's use of reading materials. Teacher outcomes were also interesting to AMSTI since a major thrust of the intervention was professional development. We considered the priority issues holistically and in a balanced way, reflecting the theory of action and situational factors that ALSDE drew our attention to.

With the AMSTI study lasting several years, including more than a year of analysis with rounds of technical review, it was reasonable to expect that technical methods would evolve over that time. IES had commissioned a series of reports to assist with technical questions that commonly arise when conducting experimental research. An example of the technical advice was to separate research questions from an impact evaluation into confirmatory and exploratory (Tukey, 1980) (or impact and exploratory) types as part of a strategy for implementing multiple comparison adjustments in order to avoid Type-1 error (Burghardt, Deke, Kiser, Puma, & Schochet, 2009; Schochet, 2008; Silverberg, 2010). The recommendation was to identify the main questions as confirmatory and to apply multiple comparison adjustments to results addressing those questions only. The primary rationale was to keep low the probability that any given confirmatory contrast would result in a false rejection of the null hypothesis. This set a high bar for statistical conclusions validity for the confirmatory analyses. We note, however, that IES added another dimension to the scheme based on different domains of inquiry. In this case, confirmatory questions were divided between two domains of inquiry: primary (associated

with student outcomes) versus secondary (associated with teacher outcomes). A separate multiple comparison adjustment was performed within each domain.

Table 3 displays the research questions as they appear in the final report. Exploratory questions addressed impacts that the research team chose to exclude from the multiple comparison adjustment. These included questions concerning moderator effects, a topic that we address in a later section of this chapter. The interpretation of the rules of evidence also put the two year impact of AMSTI into the exploratory category. This was because the impact estimate was obtained after the treatment embargo against the control group had been lifted so that to be unbiased, additional assumptions had to be met for this "extra-experimental" result. We return to this issue later in this chapter.

The IES technical standards calling for separating research questions into confirmatory and exploratory required us to put our research questions into priority order. The structure of the report would be consistent with this, with the confirmatory questions addressed early on and the exploratory ones addressed in later chapters. The reframing of the research questions and the ongoing development of the study plan also reflected the pragmatics of getting the reports written, reviewed, edited and delivered on time. The work had to meet the basic practical need for results to be reviewed and published within contract resources and timelines. The IES technical standards policy prioritized the confirmatory questions concerning average impact. This policy also drove the sequence of analytical work since analyses of these questions had to clear the review process first to make sure they would meet the publication deadline.

TABLE 3: RESEARCH QUESTIONS FROM THE IES REPORT (EDITED TO MAKE MORE COMPACT)

Primary confirmatory research questions:	
What is the effect of AMSTI on student achievement:	in mathematics problem solving after one year? in science after one year?
Secondary confirmatory research questions:	
What is the effect of AMSTI on the use of active learning instructional strategies:	by mathematics teachers after one year? by science teachers after one year?
Exploratory research questions:	
What is the effect of AMSTI on student achievement:	in mathematics problem solving after two years? in science after two years?
What is the effect of AMSTI on student achievement:	in reading after one year?
What is the effect of AMSTI on reported levels of content knowledge:	by mathematics teachers after one year? by science teachers after one year?

What is the effect of AMSTI on reported levels of student engagement:	by mathematics teachers after one year? by science teachers after one year?
Do the one-year effects of AMSTI on student achievement in (a) mathematics problem solving, (b) science, and (c) reading vary by:	student pretest scores? low-income status, proxied by enrollment in the free or reduced-price lunch program (as part of the National School Lunch Program)? racial/ethnic minority status? gender?
What are the effect of AMSTI on student achievement in (a) mathematics problem solving, (b) science, and (c) reading for categories of:	student pretest performance? low-income status? racial/ethnic minority status? gender?

The term exploratory came to apply to any of the analyses and findings that were not prioritized for inclusion into the multiple comparison adjustments or were not direct experimental impact results. The IES guidance document prepared by Schochet (2008) considered these exploratory findings to be preliminary and used to generate hypotheses for future study, that is, they would have to be replicated in future research before they could be considered reliable.

While there was flexibility in reprioritizing the research questions and in designating certain questions as confirmatory, technical standards had to be adhered to. For example, while we designated the questions as confirmatory or exploratory after randomization, it was critical to make the determination before seeing the results of the analysis; doing otherwise would introduce the possibility of bias from a post-hoc selection of the findings that really matter.

The lesson here is that the research questions in the 'varnished' version of the study plan and final report, were not static entities. Changes to the study plan—based on new technical standards, pragmatics or issues of feasibility, or priorities set by the funder—have to be considered relative to the basic needs of the stakeholders, who are the subjects of the research and who are participants in the evaluation. It is important to not lose sight of their interests. For example, while technical standards may call for corroboration of exploratory results through replication before they can be considered dependable, it may be impossible to replicate an experimental finding in the time required by a stakeholder to make a decision. The impact evaluator faces the challenge to produce results using high standards of research within the timeframe of the research contract; however, the stake-holder also needs reliable information, perhaps including answers to exploratory questions, on which to base decisions about implementation and program improvements, while working against a different timeline. The tension we describe here between local stakeholder needs and centralized standards for research

is not new. We refer the reader to Cronbach et al. (1980) and Cronbach (1982) for useful discussions about challenges of conducting evaluations under real-world constraints, where timeliness and relevance of findings figures into the validity of the findings.

Lesson 2: The Value of Paired Randomization for Working within AMSTI's Scale-up

From our work in conducting RCTs in school districts on local questions (Newman, 2008), we had developed a workshop format for conducting randomization that had the goal to increase teacher participation in, understanding of, and commitment to randomization. Getting buy-in from participants can be critical for the success of the study, including reducing attrition, avoiding contamination, and providing useful input on surveys. Especially, as was the case with AMSTI where the participants are also the main stakeholders, it is important that they have an appreciation of the advantages of randomization to provide them with valid results. Although using a matched pairs design is not essential to achieve the benefits of randomization, we found that stakeholder involvement in identifying the criteria to block on, and randomizing in matched pairs, further increased the sense of commitment—participants felt that they were part of a larger unit that was important to keep intact to maintain the integrity of the design. We also found that pairing improved precision of impact estimates, even when modeling covariates, and so adopted the strategy (Jaciw, Wei, Ma, & Newman, 2008).

Of the 144 schools that volunteered and met criteria for participation in the study we were able to select as many as AMSTI had available resources to support. Within each region, we worked with the AMSTI staff and regional site directors to select the schools and identify the pairs for participation. Sampling was of the matched pairs, so sampling and pairing were conducted simultaneously. The process was described in the report as follows.

> "To select the pairs that would then be randomized within each region, researchers paired schools first on the basis of similarity of grade configuration, then on mathematics scores, on percentage of racial/ethnic minority students, and finally (when possible) on percentage of students from low-income households (students enrolled in the National School Lunch Program). AMSTI regional directors provided input on the appropriateness of the pairings, based on their local knowledge of similarities that went beyond those captured in the formal criteria, and corrected or updated data. As pairs were selected, researchers used a spreadsheet to maintain a running tally of the demographics of the combined pairs in each region to ensure they were similar to the demographics of that region's schools. In some cases, this consideration was the deciding factor in choosing a school from two closely-matched pairs" (Newman et al., 2012, p. 17).

Involvement of regional staff in the school selection and pair formation process had several advantages. The process increased understanding of, interest in, and trust of, the research methods for the AMSTI staff and regional site directors. The process utilized the local knowledge of the regional directors to select pairs that were similar on factors that were not captured by the available demographics, thereby potentially yielding an advantage for

precision. In several cases, for example, the participants identified a pair within the same district, perhaps with understanding of local conditions. This had an additional advantage in avoiding giving the district leadership two control schools, which would discourage participation—we needed the support of district leaders to carry out the research, and having some schools in each district randomly assigned to treatment was useful for getting district involvement and support. To further cement the regional commitment to the experiment, a regional official (under close scrutiny of the researchers) ceremonially tossed the coin implementing the randomization.

During the analysis phase, naturally we were curious about whether local stakeholder-informed criteria for modeling matched pairs resulted in greater precision of the impact findings. That is, in addition to the procedural payoff from securing interest and commitment by the stakeholders, was there a statistical benefit? If we found one, it would attest to stakeholders' implicit knowledge of factors that make schools similar or different, beyond what is captured through basic demographics. The relevant results for the mathematics outcome that were obtained as part of the analysis for the IES report are displayed in Table 4 below. (The results were conducted as part of the trial but did not end up in the IES report.) We examined impact on math achievement after one year using six impact models. Each included a different subset of the covariates. Comparing models 1a to 1b, 2a to 2b, and 3a to 3b, we observe that including matched pairs led to a reduction in the school-level variance component in each case. The difference in deviance statistics reaches significance ($p < .01$) in each of the three comparisons. There is a benefit to including matched pairs even after modeling all the other covariates.

TABLE 4: VARIANCE COMPONENTS AT SCHOOL AND STUDENT LEVELS FOR SEVERAL MODELS OF IMPACT ON MATH OUTCOMES AFTER ONE YEAR

Model	Covariates	School-level variance	Student-level variance	Deviance statistic
1a	No covariates	0.224*	0.776*	189471.8
1b	Matched pairs only	0.053*	0.776*	189085.1
2a	Pretest only	0.020*	0.345*	174213.7
2b	Pretest and matched pairs	0.012*	0.345*	173951.0
3a	All covariates but matched pairs	0.018*	0.338*	173797.6
3b	All covariates including matched pairs	0.009*	0.338*	173534.1

Note. The variance components displayed in the table have been rescaled by dividing each component by the total variance in the outcome measure.

Except for the model without covariates (Model 1a) the variance components represent the residual variance at each level after conditioning on the effects of the covariates.

N (students) = 18,713
N (schools) = 82

*$p < .01$

Lesson 3: Reasons for Difficulty in Maintaining a Control Group Embargo for Two Years

An RCT in the context of an agency's own popular program and its need for evidence greatly simplifies "recruitment" because the participants have already applied to participate. But this context has its own issues and lessons to be learned. We address the problem of the embargo that keeps the control schools out of treatment for the duration of the RCT. AMSTI was a two year program with two summers of training and additional materials and support provided through the second year. It would have been preferable to run the comparison of treatment and control for two years to obtain an estimate of the two year impact. The political context made this difficult to implement because there was no assurance year-to-year that the program would continue to scale up to new schools. For each year a school remained in control, the probability of not getting into the program at all increased because of this uncertainty. There was also a potential problem for a school staying in control for two years if AMSTI did continue to scale up to new schools—other schools would be joining AMSTI ahead of them.

Since the cost of providing AMSTI to the control schools was not built into the research funding (or assured in ALSDE's funding), the research team lacked the leverage often available to researchers with the resources to recruit schools with the promise of receiving the intervention at the end of the control embargo. It may have been possible to persuade the AMSTI team to include a two year embargo for schools randomly assigned to control as a condition for acceptance into AMSTI but they were reluctant to do so. As the research team, we were concerned that perceptions of lost opportunity, could lead to resentful demoralization, compensatory rivalry (Shadish, Cook, & Campbell, 2002) and attrition by controls thereby producing biased effect estimates.

Also influencing our decision to lift the treatment embargo after one year was the availability of an alternative non-experimental approach to estimating two-year impacts. Bell and Bradley (2009) developed a methodology for inferring a 2-year program impact based on a comparison of outcomes between randomized groups after the first year of the intervention (before the control group received AMSTI) and after two years (after the control group received its first year of exposure and the original treatment group received its second.) [We refer the reader to chapter 5 in Newman et al. (2012) for a description of the method.] The validity of the two-year impact results using this method depended on the assumption that the intervention had not substantively changed between the initial group's first year in AMSTI and the "control" group's first year in AMSTI. Since the program was quite stable, having been established four years earlier, this was a reasonable assumption. The research team looked for change in a series of conditions that potentially could moderate the first year impact. For the most part, no such changes were found, giving the team confidence that the assumption of a stable effect of the first year of implementation was satisfied.

AMSTI is a two year program, but because of the additional assumptions required for the Bell and Bradley estimates, IES determined that the analyses of two year impact would be considered exploratory. This meant that only the impact after one year could be included among the confirmatory findings.

Every study is unique in the burdens it places on participants, including the perception of a lost opportunity through being assigned to the control group. While researchers may conduct RCTs under the premise that, until demonstrated, one cannot assume that the treatment offers an advantage—business as usual may be as or more beneficial—this way of thinking is very often not shared by study participants. In the case of AMSTI, all the participants had already gone to considerable effort to apply for AMSTI and the belief they shared with many AMSTI staff was that they were engaged in rigorously generating evidence to demonstrate efficacy. Some did recognize that they were taking a risk in agreeing to a rigorous RCT—the rigorous study may produce negative results— but also recognized that by following the "rules" the results would be credible. We view it as a difficult judgment call: In this case, would the results have been stronger with a two year embargo, where we may have been risking attrition and other potentially biasing behaviors, compared to providing the two year impact estimate under the assumption of a non-changing intervention? In any case, a researcher working with the implementation of a local program does not have the same control as a researcher bringing funding for the intervention and recruiting schools into an established design.

Lesson 4: Avoiding Bias in the Impact Model Selection

The statistical analyses in this project were exceedingly complex and benefited from the ongoing technical work sponsored by IES parallel to the work of research teams such as ours. We learned from our interactions with experts and some of the lessons learned were practical steps, such as building independent duplication of critical analyses into the analysis process, while others were about the nature of statistical problem solving, which we illustrate through the example of how we arrived at the final version of our impact model.

A lesson that came through loud and clear from the multi-year study was that over that period the methods that were available and documented changed and improved. Also, that there was not always agreement among the multiple sets of experts advising the project. This added to the interest and excitement of this work. A lesson for researchers about to embark on an RCT is to not expect analysis of a complex dataset to be cut and dried. There is craft knowledge that needs to be used.

An example of craft knowledge concerning the process of analysis—something not often seen in textbook discussion—is the practice of replicating the results of analysis from start to finish, especially for the most important findings. Most of our work was done in SAS using the MIXED procedure, however, we included a process check whereby after identifying the analytic samples, at the school and student levels, and accounting for and documenting the steps leading from the baseline sample (at time of randomization or student roster formation) to the analysis sample, we handed the data with the documentation and codebook to a Senior Research Scientist well versed in R software, and asked him to replicate the result. We found that in some cases the analyst needed more information to capture the same process. Through this parallel process we obtained greater assurance that the results of the analysis did not just represent the peculiarities of a certain software or estimation algorithm, or a human misstep on the part of one analyst. We can consider this duplication step as a type of sensitivity analysis where we test the robustness of results to changes of analytic software or individuals performing the analysis.

But what should we do if two reasonable approaches give different results? The following example of a quandary we encountered as a result of conflicting recommendations illustrates the need to engage a 'tie-breaker', an expert who is blind to the outcome and who can guide aspects of analysis objectively and reduce the possibility of bias in the results. While impacts of programs can be estimated from experimental data without the use of statistical models, it is common to use regression adjustments to increase the precision of the estimates. In doing so it is important to select a model in advance of knowing the results to prevent 'shopping' for a model that yields the results one expects or wants. A study plan helps serve this purpose because it allows the analyst to commit to a certain course of analysis and specify the analytic models to be used. While ideally the data would be 'well-behaved' and allow the analyst to stick with the script of the study plan, this does not always happen. Sometimes the planned approach simply does not work (e.g., the estimation algorithm does not converge), or produces results that require further-refinement of the impact model, and the analyst must deviate from the plan. This was the case with the analysis of the AMSTI data and it taught us the importance of enlisting 'tie-breakers'—individuals well-versed in the study but who are kept away from results of specific models so that they can advise on model-selection in the event that several equally-good candidate models emerge. It was interesting to us that the need for blinding procedures could arise very suddenly. It is normal in medical studies to blind patients and researchers to the condition to which individual patients are assigned, but in this instance it seemed just as important to blind the analyst to the impact outcomes resulting from alternative model specifications. Protocols for doing this are not usually built into the research process. Building in this step is a further example of the use of craft knowledge in the course of the analysis.

We describe the AMSTI case in more detail to illustrate how such situations may arise. In our initial approach to analysis of one year impacts on student achievement, we represented the teacher level in our analytic model. Our web-based surveys gave us reliable information concerning which students belonged to which teachers, as well as concerning teacher membership in schools, thereby permitting a three-level model (students nested in teachers nested in schools.) While ignoring the middle level of a three-level design has been shown to not influence program impacts and their standard errors (Zhu, Jacob, Bloom, & Xu, 2012) since the data were available, for the sake of completeness, and to reflect the design in our analytic model as fully as possible, we included the teacher level. A three-level model that included the full set of covariates led to the peculiar result that variance components could be estimated at the student and teacher levels but not at the school level. Discussion with statisticians at SAS indicated that the statistical routine could not estimate a variance component at the school level because there was a trivially small amount of variance to be estimated at that level—the estimation algorithm was reaching a boundary condition for estimating the effect, and stopping, issuing a zero value for the variance estimate with no p value.

We sought technical advice and reached a solution that was reasonable: since there was negligible amount of variance remaining at the school level we would not model a random effect at that level but would retain random effects at the teacher and student levels. A different set of technical reviewers rejected this approach and gave a firm admonition that a random effect must always be modeled at the level of randomization. Our final model reflected this new advice. To

preserve an estimable random effect at the school level, we removed the teacher level from analysis. Eliminating the teacher level and including the school level random effect introduced a very small but estimable amount of variance at the school level (3% of the total variance was redistributed to the school level with removal of the teacher effect.)

We found ourselves in a situation where the planned model produced an unexpected result, requiring additional model refinement and a choice to be made about the final model. We recognized that while the jury was still out concerning which specification would be selected as yielding the benchmark finding, we had to enlist a member of the research team and blind them to the results of the different models in case the different specifications were consequential to the results. That person could help settle the debate concerning the best model using their own expertise and judgment of the benefits of different approaches to modeling, without possibility of bias arising from knowing the results from the different models. In our case, the three-level model and both two-level models (with schools and students, or with teachers and students) produced much the same results for the impact estimate and its standard error. Therefore which model we selected was inconsequential. However, one can imagine situations where a slight re-specification of a model tips the results from being significant to being non-significant or vice versa. Here the temptation may be to swing the result in one direction or the other, and so a blinded decision process is preferable.

This story also illustrate the tiers and stages to the review process and the decision, such as concerning which analytic model is considered yielding benchmark results, must be reached through productive argumentation. In this case, experts disagreed, and open exchange allowed us to put options on the table, select a defensible alternative, and we hoped, avoid bias in the process. This process can be contrasted with the situation that we think is more typical—where the analyst is working alone and may inadvertently introduce bias into the process when forced by the situation to stray from the study plan.

Lesson 5: What Alabama Can Make of (and Do With) These Results?

The researchers must also be concerned with external validity, particularly in understanding the reach of general conclusions that can be drawn from an RCT of this sort. What should ALSDE, the primary stake-holder, make of the results? The positive impact findings for math and science gave ALSDE evidence to make the case for the continuing funding and roll out of AMSTI across the state. The primary message was that AMSTI works and should be continued.

We were very aware that the report, especially the results of confirmatory analyses, were based on the strongest standards for internal and statistical conclusions validity; however, the process of distilling the results also left out the part of the story that would be of potential importance to Alabama. The tension we referred to early in this chapter—that the report represented primarily the message to the national audience as opposed to a local agency testing its own program—made us consider how the results might have been framed to be of additional use for the latter audience.

Framing the findings for greatest use for Alabama involved considering questions of the reach of the confirmatory impact findings with a view to how the program could be improved. We were

particularly mindful of Cronbach's (1975) point that generalizations decay with time. It was not the charge of the impact report to draw generalized conclusions; however, we felt that as evaluators it was our charge to give attention to the possible reach of the impact by considering factors that moderate it. This would point to where program improvements might be made. The project gave us only a modest opportunity to investigate potential sources of heterogeneity in the impact, including factors leading to intensification or diminution of impact with time. We explored (but did not report) whether impact varied across the two sub-experiments. We found that it did not, recognizing the low power of the test to detect such differences. Further, the check of assumptions as part of the Bell and Bradley technique for estimating two year impacts, showed stability of factors that could potentially moderate impacts, suggesting constant impact at least over the short interval of the experiment.

One place for caution in extrapolating the primary results and conclusions of the report concerned the potential impact for the schools that did not meet eligibility criteria—specifically, the schools not admitted because 80% of teachers did not agree to participate. The impact of AMSTI in such schools was unknown.

In addition to not overgeneralizing the results of the confirmatory analyses, we did not want to underplay specific exploratory findings. Within this cautious interpretive framework, we recognize that there is a distinction between exploratory results based on underpowered analyses that are not declared ahead of time and perhaps used to map possible directions for future research, and results based on analyses that *are* declared in advance and that may have consequences for the present but that are not adjusted for multiple comparisons and so have a higher probability of being false. Both may be considered exploratory, but the latter type may provide results that are a basis for action, especially if there is little risk and a potential benefit to the action.

An example of a declared-in-advance exploratory analysis was that of the moderating effect of minority status on the impact of AMSTI. These showed a consistent trend towards there being less benefit of the program for minorities. Figure 2 represents the exploratory findings in table form in the final report.

Importantly, the question of impact for minorities was not raised post hoc and the statistical power for these analyses could not automatically be considered low—the minimum detectable effect size (MDES) for the sample as a whole was much lower than expected and power for assessing the difference between minorities and non-minorities in the impact may have been higher than expected.[5] While included in the final chapters of the report, the finding both raises serious question about the generalizability of the confirmatory finding to important parts of the population and points to areas of potential program improvement.

---

[5] Works by Bloom (2005) and Jaciw, Lin and Ma (2013) show that where groups are randomized but impacts are being compared for subgroups identified at the individual level, power to detect the differential effects may be larger than usually assumed. The cluster-level deviations in performance—the usual main source of uncertainty in effect estimates from cluster randomized trials—are differenced away. The precision however has to be considered relative to the size of the differential impact, and there is not much documentation concerning average effect sizes for differential impacts involving different kinds of groups.

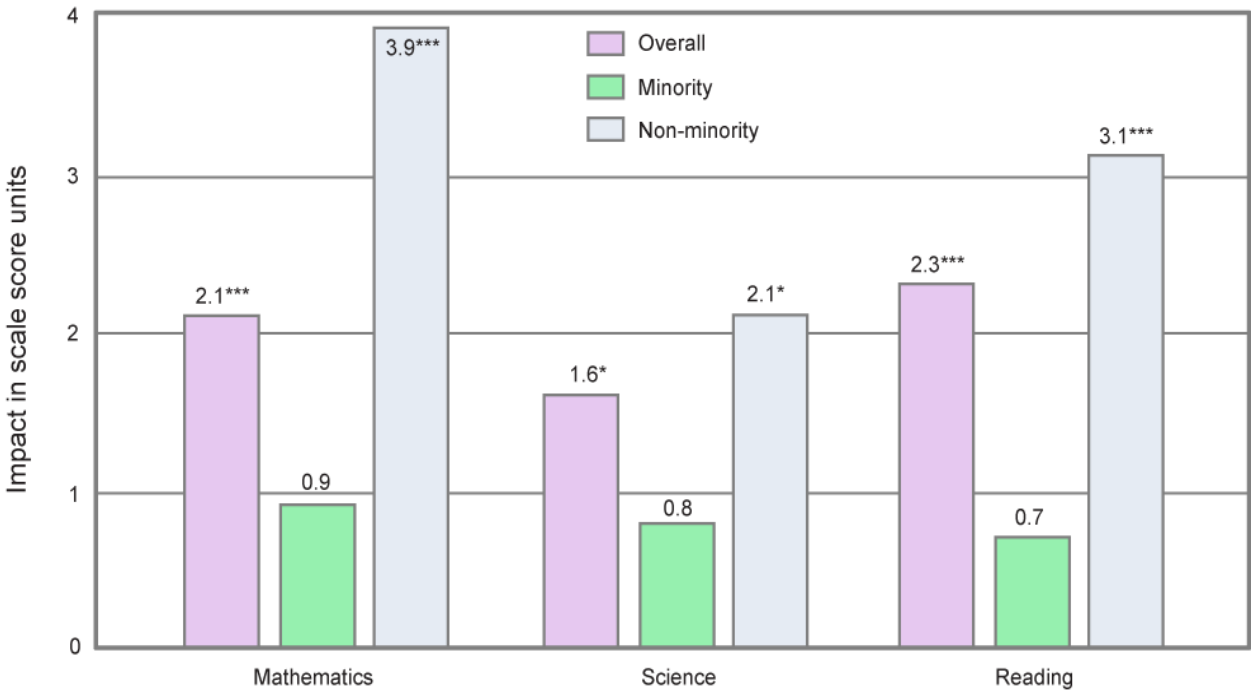

FIGURE 2: RESULTS FOR SUBGROUP ANALYSIS SHOWING MODERATION OF IMPACT BY MINORITY STATUS.

(*** $p < .01$; ** $.01 <= p < .05$; * $.05 <= p < .10$)

Our general lesson was that a strict dichotomy of analyses and results into confirmatory or exploratory is not always appropriate. Specifically in the case of AMSTI, we feel additional analyses involving minority status were merited—there was much more to the story that deserved to be told. In particular, the potential policy significance of the results involving impacts for minorities we felt superseded the higher risk of their being false from not being included in multiplicity adjustments. Unfortunately there was limited opportunity for further investigation within the scope of the contract. Another reason to take the results involving minority status seriously was that we had no assurance that a replication study—what is usually recommended to corroborate results from exploratory analyses—was feasible. Alabama needed information immediately, not a further five years down the road. Also, even if a replication trial could be mounted, Cronbach's reminder of the possibility of time-by-treatment interactions we felt should be taken seriously. If the effects were to decay with time, there would not be an opportunity to corroborate a stable effect even if another trial could be carried out, leaving the exploratory outcomes concerning impacts on minorities as the best available evidence.

While not part of the final report, we did run some additional exploratory analyses to check on alternative explanations for the treatment by minority status interaction (Newman, Ma, & Jaciw, 2012). We wanted to examine whether the interaction was an artifact of socio-economic status or an organizational effect of percent minority at the classroom or school level (it wasn't). The moderating effect of individual minority status on the impact of AMSTI was robust to the inclusion of the interactions of treatment with other covariates at the student and school levels.

We also separated the sample by minority status and found greater variation in the treatment-control difference across randomized blocks for minorities. Further, this variation was reduced, but only for minorities, when we accounted for the interactions between socioeconomic status and treatment and between pretest and treatment. While we were not able to pursue this further, the possibility that under some conditions, and perhaps in some schools, AMSTI was very valuable for minorities, could provide a window through which developers might identify a malleable factor that, once acted on, could help decrease the achievement gap.

While this kind of further exploration of exploratory findings has limits from the point of view of what can be plausibly concluded from an RCT, we note that the structure of the RCT can provide a very useful framework for exploration that is not available in purely observational data. For example, a systematic exploration of moderators of the treatment effect is more revealing if those moderators are not also confounded with treatment. Our lesson here is that while results of confirmatory analyses give us high standards of assurance about whether or not a program works on average, an RCT also offers a much wider field of exploratory questions that reveal the conditions for efficacy. Where the stakeholder for the RCT findings is also the program developer, these can give guidance about how the program can be improved.

A further anecdote illustrates the 'unvarnished' process of research and its sometimes complex evolution. The treatment-minority interaction in the analysis of impact on reading was present after the end of the first sub-experiment (i.e., for only half the sample). We reported our findings up to that point at a meeting of the technical working group, and one member suggested that the trial should be stopped because it was apparently increasing the minority achievement gap and that this was unethical. We were not prepared to stop in case the initial finding was simply a random aberration; however, continuing to see the effect after combining these data with those from the (independent) second sub-experiment provided confidence that the initial result was not just a fluke—the second sub-experiment can be viewed as a replication. The episode also raised the issue of potentially developing stopping rules for experiments in education in order to catch detrimental effects early on.

We consider it an exciting aspect of educational research to make sense of the heterogeneity of effects. But we believe a disciplined approach is warranted since interaction effects proliferate quickly. Thus it is critical to ensure that model misspecification is not driving observed differences and established strategies such as use of multiple comparisons corrections, continue to apply. Where the stakeholder is also the developer, a value in conducting a rigorous program evaluation arises from the potential for identifying areas of improvement through exploring the reach of the effects. For a study of the next iteration of a program, the results of an RCT on the previous iteration, whether confirmatory or exploratory, can inform working hypotheses about the program and its effects.

Conclusion: Reporting to the Local Agency

ALSDE held a press conference[6] on the day that the results were released, February 21, 2012, over 6 years since the initial randomization. The event at ALSDE celebrated the enormous investment of effort in their program. They were proud to announce the positive results for math, which was the primary confirmatory result. Their announcement also drew on some of the exploratory findings, e.g., the two year impact for math and science as well as reading. The indication that AMSTI was not as successful on average for minority students was also not mentioned. The day the report was released, the Alabama legislature was in session with decisions pending on the AMSTI budget line.

The timeline for reporting the results, including the decisions to not publish interim results and to focus on confirmatory results, meant that some of the opportunity for use of the research for formative improvement and for exploration of puzzling findings may have been lost. Until a few days before the report was released by IES, the stakeholders in Alabama were given no inkling as to the results. If we circle back to the initial discussion in this chapter, where we highlight the way that this RCT was conducted to assist the agency that was implementing its own intervention, we can see that the singular focus on average impact and on only a small number of results limited the overall value of the multi-year effort. This is not to say that federal agencies that have committed to a decade-long agenda of increasing the rigor of education science should be less cautious about declaring programs effective. On the contrary, we might suggest being even more cautious where, when looking beyond the average impact, there are indications that the confirmatory conclusion may mask policy-relevant effects. A lesson from our RCT on AMSTI is that, for researchers working with a local agency, reporting interim results and elaborating on exploratory indications that point toward areas for improvement can increase the value of the investment in research. A program developer is taking a risk in subjecting the program to a rigorous evaluation that involves a long wait to obtain a single up or down result. Providing timely results and analyses that point to areas of potential improvement are ways to reduce the risk and increase their willingness to participate in rigorous research.

[6] The press release was retrieved January 13, 2014 from https://docs.alsde.edu/documents/55/NewsReleases2012/2-21-2012_AMSTI%20study%20results.pdf)

References

Bell, S.H., & Bradley, M.C. (2009). *Potential for Using Experimentally-based Impact Estimates After the Control Group in an Educational RCT Receives the Intervention*. Report submitted to the Institute for Education Sciences, Washington, DC.

Bloom, H.S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning More from Social Experiments*. New York: Russell Sage Foundation.

Burghardt, J., Deke, J., Kiser, E., Puma, M., & Schochet, P. (2009). *Regional Educational Laboratory Rigorous Applied Research Studies: Frequently Asked Analysis Questions*. Princeton, NJ: Mathematica Policy Research, Inc. Unpublished Manuscript.

Cronbach, L.J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 116-127.

Cronbach, L.J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey-Bass.

Cronbach, L.J. and Associates (1980). *Toward Reform of Program Evaluation.* San Francisco, CA: Jossey-Bass.

Jaciw, A.P., Lin, L., & Ma, B. (2013). *An Empirical Study of Design Parameters for Assessing Differential Impacts for Students in Group Randomized Trials*. Palo Alto, CA: Empirical Education Inc. Manuscript in preparation.

Jaciw, A.P., Wei, X., Ma, B. & Newman, D. (2008, March). Matched-pairs designs and standard errors of impact estimates: Lessons from 10 experiments. In F. Ye (Chair), *Matched Samples and Propensity Score Methods*. Symposium conducted in a paper discussion at the annual meeting of the American Education Research Association, New York, NY.

Newman, D. (2008). *Toward school districts conducting their own rigorous program evaluations: Final report on the "Low Cost Experiments to Support Local School District Decisions" project*. Palo Alto, CA: Empirical Education Inc.

Newman, D., Finney, P.B., Bell, S., Turner, H., Jaciw, A.P., Zacamy, J.L., & Feagans Gould, L. (2012). *Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI)*. (NCEE 2012–4008). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Newman, D., Ma, B., & Jaciw, A.P. (2012, March). *Locating Differential Effectiveness of a STEM Initiative through Exploration of Moderators*. Paper presented at the annual research conference of the Society for Research on Educational Effectiveness, Washington, DC.

Puma, M.J., Olsen, R.B., Bell, S.H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Schochet, P.Z. (2008). *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Silverberg, M. (2010). *Exploratory Analysis Chapter* [Memorandum]. Washington, DC: Institute of Education Sciences.

Tukey, J.W. (1980). We need both exploratory and confirmatory. *The American Statistician*, *34*, 23-25.

Zhu, P., Jacob, R., Bloom, H., & Xu. Z. (2012). Issues in the design of group randomized studies: assessing two-level designs for three-level situations, *Educational Evaluation and Policy Analysis*, *34*(1): 45-68.

Reference for this paper: Newman, D., & Jaciw, A.P. (2014). The Unvarnished Report of Conducting an RCT of a Statewide STEM Program in the Process of Scaling Up in R. Maynard, L.V. Hedges, & S. Wachter (Eds.), *Teaching Cases on Designing and Fielding Experimental Design Evaluations in Education (provisional title).* Manuscript in preparation.