JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 50, NO. 9, PP. 1127–1141 (2013)

*Research Article*

# Conducting Causal Effects Studies in Science Education: Considering Methodological Trade-offs in the Context of Policies Affecting Research in Schools

Joseph Taylor, Susan Kowalski, Christopher Wilson, Stephen Getty, and Janet Carlson

*BSCS, 5415 Mark Dabling Blvd., Colorado Springs, Colorado 80918*

*Received 1 June 2011; Accepted 12 August 2013*

Abstract: This paper focuses on the trade-offs that lie at the intersection of methodological requirements for causal effect studies and policies that affect how and to what extent schools engage in such studies. More specifically, current federal funding priorities encourage large-scale randomized studies of interventions in authentic settings. At the same time, policies that require schools to make adequate yearly progress (AYP) on state achievement tests make it risky for them to participate in trials of unproven interventions. Researchers attempting to balance rigorous study designs with the pressures felt by school districts must thoughtfully balance validity threats that are introduced in the negotiations with those districts. In this article, we draw on our experience conducting several randomized trials to discuss how these factors can be balanced. Specifically, we discuss necessary trade-offs in causal effects studies related to statistical power, measurement sensitivity, research ethics, minimizing bias, and addressing the interests of all stakeholders. © 2013 Wiley Periodicals, Inc. J Res Sci Teach 50: 1127–1141, 2013
**Keywords:** policy; accountability; validity/reliability; program evaluation

Recent education policies and their effects on school systems can create formidable challenges to those engaged in rigorous education research that tightly aligns research questions and methods. This alignment transcends research traditions; in fact, rigorous research can take on a variety of foci and methodologies concomitant with their associated epistemologies, all making important contributions to our knowledge base in science education. As such, this article focuses on causal effects studies as just one important pursuit in science education research and comes from the perspective of their unique challenges, not from a position of advocacy.

The National Research Council (NRC, 2002) categorized education research questions into those addressing *what is happening*, *why or how is it happening*, and *is there a systematic effect?* Each of these categories is normally associated with a specific family of research designs or analytic methods, and each has a vital role to play in advancing our understanding of teaching and learning science. This said, federal funding initiatives over the past decade tended to emphasize studies of systematic or *causal* effects. For example, the research solicitations of the U.S. Department of Education's (DoEd) Institute of Education Sciences (IES) have two dedicated categories of funding for studies that seek causal effects. The National Science Foundation's

Contract grant sponsor: Institute of Education Sciences; Contract grant sponsor: U.S. Department of Education; Contract grant number: R305K060142.
*Correspondence to*: J. Taylor; E-mail: jtaylor@bscs.org
DOI 10.1002/tea.21110
Published online 1 October 2013 in Wiley Online Library (wileyonlinelibrary.com).

(NSF) Discovery Research Program (DR K-12) includes a *Cycle of Research and Development* in its funding solicitation (e.g., NSF 11-588, 2011) that clearly situates causal effects studies in a larger sequence of research and development for STEM education interventions (e.g., instructional approaches, curriculum materials, or professional development). Similarly, the perspectives of federal entities such as the *What Works Clearinghouse* (WWC) favor designs where strong evidence of causal effects of interventions can be obtained (IES, 2011a). From the perspective of groups like the WWC, the rationale for doing a causal effects study is to isolate, to the extent possible, the unique effects of education interventions from the array of other variables that affect student outcomes. If the design of a study is such that effects on students or teachers can be confidently attributed to an education intervention, this information is quite valuable to decision-makers charged with choosing education programs to implement in their home context.

Current thinking about intervention studies suggests that certain designs are preferable when the goal is to isolate causal effects. Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson (2007) and Shadish, Cook, and Campbell (2002) describe general conditions that, when met to a sufficient extent, allow for confidence in causal inferences. A subset of these conditions affects research design and scale, particularly that variation in the cause is reliably associated with variation in the outcome, and that plausible alternative explanations be eliminated (to the extent possible). Without meeting these criteria, research findings face threats to statistical conclusion validity and internal validity (Shadish et al., 2002).

In theory, a randomized experiment directly addresses the need to eliminate alternative explanations because it is designed to distribute unobserved variables equally across groups so that those variables do not confound the outcome comparisons of interest (Shadish et al., 2002). Alternative explanations may be also be eliminated using a quasi-experimental design (QED) with closely matched groups. This said, experiments and quasi-experiments may tell only part of the story as without probability sampling, treatment effects may not generalize beyond the sample and, at the surface, these designs tell the researcher if there is a noteworthy systematic effect, not necessarily why or how the effect occurred.

Measures of association between cause and outcome are often expressed with inferential (parametric) statistics under the premise that the association is larger than one might expect to occur by chance. To increase the probability that noteworthy effects will be detected (i.e., avoid Type II error, increase statistical power), researchers often feel compelled to conduct studies with large sample sizes. These two factors, in particular, result in a portfolio of causal effect studies in education that are large in scale and more often than not feature random assignment.

Policies supporting the funding of educational research are encouraging large-scale randomized studies of interventions in authentic settings. At the same time, policies affecting schools tend to work against the implementation of causal effects research. For example, the increased focus on high-stakes student assessments and pressures for schools to make adequate yearly progress (AYP) makes it risky for schools to participate in trials of unproven interventions. The current emphasis on testing can make recruitment more difficult and thus, may have negative implications for statistical power. Therefore, policy-driven shifts in school priorities introduce validity threats that researchers must balance thoughtfully.

In this article, we draw on our experience conducting several randomized trials (Getty, Wilson, Taylor, & Kowalski, 2011; Wilson, Roth, Taylor, Landes, & Stuhlsatz, 2012; Wilson, Taylor, Kowalski, & Carlson, 2010) to discuss trade-offs among different strategies that may help future researchers address the challenges of conducting studies of causal effects in the current policy climate. Other similar articles have been useful in helping the field in this area. For example, McDonald, Keesler, Kauffman, and Schneider (2006) provided practical guidelines for the implementation of scale-up studies, often featuring a randomized control trial design. Further, in

2012, JRST dedicated a special issue to large-scale intervention research and some of the issues discussed in this article were introduced in that issue. For example, Lee and Krajcik (2012) discussed the challenges of implementing efficacy trials of scaled interventions, with particular emphasis on challenges in urban settings. Penuel and Fishman (2012) described the limitations of efficacy trials for informing teaching and learning, suggesting that design-based research be conducted in concert with efficacy research. Finally, Marx (2012) summarized the articles in the special issue and highlighted the implications of conflicting policies for education research. The purpose of this article is to (1) extend and integrate the ideas presented in these important contributions by situating the methodological requirements of causal effects studies within the current policy climate, and (2) provide a detailed case study that illustrates the trade-offs encountered in one study when threats to validity arose.

## Conducting Causal Effects Research in a Climate of Accountability

### *Policies Affecting Research*

Over the past 15 years, U.S. agencies that fund educational research have advocated for and supported efforts that align with what has been called *evidence-based reform*. This movement promotes the use of programs or practices for which there exists causal evidence of effectiveness. Evidence-based reform in education has often been described as a response to (a) a general lack of rigorous evidence regarding what programs or practices are effective; (b) the adoption of programs or practices based on ideology or popular trends instead of evidence; and consequently (c) the absence of coherent knowledge growth about effective programs or practices (Slavin, 2008). As a result, some in the education community have looked to research approaches from the medical and agricultural fields in an attempt to address questions associated with causal effects and to better inform policy and practice. A range of legislation heralded this effort, most notably the No Child Left Behind Act (NCLB), which famously mentions scientifically based research more than 100 times (U.S. Congress, 2001), the Education Sciences Reform Act (U.S. Congress, 2002) which established the IES, and the National Center for Education Research (NCER).

Subsequently, researchers and government agencies both made efforts to synthesize evidence on effective programs, under the argument that the effectiveness of programs or practices could not be established from individual studies, but that such evidence must be replicated to build a convincing case (Slavin, 2008). Such syntheses include the WWC, funded by the U.S. Department of Education and the Best Evidence Encyclopedia (BEE) at Johns Hopkins University. Each synthesis effort has its own set of methodological criteria a study must meet to gain entry, but each favors randomized experiments. This preference is evident in the hierarchy of research designs for causal inference described in the evidence standards established by the WWC for reviewing studies. Specifically, the WWC Standards of Evidence document states, "Currently, only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence, while QEDs with equating may only meet standards with reservations. . . (IES, 2011a, p. 11)."

In response to these policies, some argued that the criteria were too methodologically restrictive and did the field a disservice by sending the message that only certain types of evidence counted, invoking the paradigm wars of the 1970s and 1980s (Schwandt, 2005). Others noted that evidence-based reform did not and should not exalt any particular methodological approach (Feuer, Towne, & Shavelson, 2002). In any case, the place of causal effects studies in science education research has certainly been elevated, and today retains a prominent place in many research policies. Because of this elevation, whether it is merited, it is important to carefully consider the methodological tradeoffs of causal effects studies.

*Policies Affecting Teaching and Learning*

Just as federal legislation has had a significant impact on considerations of research evidence and research design, these same documents also have had a significant impact on what happens in schools via parallel accountability mandates. The accountability mandates affected how students are assessed, what gets taught (and what does not), and how teachers and schools are evaluated. These outcomes can create a climate that is challenging to research studies focused on causal effects.

*How Students Are Assessed.* One of the primary mandates of NCLB was that students would be routinely tested in math, reading, writing, and eventually science, and would be expected to meet certain proficiencies. This resulted in an increased emphasis on preparing students for state-level standardized tests, including the loss of instructional time to test preparation activities (Bill and Melinda Gates Foundation, 2012; Zellmer, Frontier & Pheifer, 2006). Consequently, researchers wishing to study programs or practices within schools may struggle to find willing study participants because many school personnel proceed with caution if there are no clear and direct links between the intervention and achievement on high-stakes tests, which, of course, there often are not—that is the point of conducting the research.

*What Gets Taught.* It was inevitable that the content of these high-stakes student assessments would, in many ways, drive what gets taught in schools. While many states developed science standards that encouraged the development of higher order thinking, scientific reasoning, or inquiry skills, the assessments that followed often emphasized students' ability to recall factual information and vocabulary (NRC, 2012). When inquiry skills were assessed, it was primarily limited to measuring whether students could design controlled experiments. Consequently, many administrators and teachers were reluctant to begin or continue enacting many of the recommendations of major reform documents in science education (e.g., AAAS, 1993; NRC, 1996, 2000, 2012), including teaching science as inquiry, engaging students in argumentation, organizing the content around unifying ideas, or connecting science with other content areas (Getty et al., 2011). Similarly, many administrators and teachers are hesitant to participate in research studies that investigate these practices and approaches to content. Ironically, although science education reform documents (AAAS, 1993; Achieve, 2013; NRC, 1996, 2012) favor engaging students higher order thinking and problem solving, causal effects research around programs that embody these outcomes is largely lacking.

*How Teachers and Schools Are Evaluated.* Under NCLB, the goals of the state-level student assessments not only included evaluating students, but also informing objective judgments and comparisons of the quality of educators and educational systems around the country. Schools are required to make AYP on these tests, and there are various consequences for those that do not. These measures present recruiting obstacles to researchers conducting studies of causal effects in schools. For example, if a school is making AYP, the administrators within that district or school may consider participation in a study too risky in terms of threatening their progress. On the other hand, in a school that is not making AYP, the administrators and teachers may also be less likely to participate in research studies of unproven interventions because they need "proven" solutions.

As a result, researchers have a dilemma. While federal policies are advocating for causal effects research on science education programs and practices, other policies implemented at the same time can create barriers to conducting such research studies within schools that may have justifiable higher priorities. In the following section, we will explore the requirements for conducting rigorous studies of causal effects and the associated implications, including the challenges and trade-offs.

## Requirements of Studies of Causal Effects and Associated Implications

Studies of causal effects have multiple requirements, some stemming from the need to establish statistical conclusion and internal validity as mentioned above (Shadish et al., 2002). Other requirements stem from economy of scale. In this section, we address several key requirements of studies of causal effects and discuss the challenges and implications associated with attending to those requirements in the context of current policy and funding limitations at district, state, and national levels.

### *Description of Requirement 1: The Need for Many Clusters of Students*

Studies of causal effects in schools are particularly challenging to execute simply because of their considerable scale. The scale of causal effects research conducted in schools is necessarily large due to the need to account for the nested nature of the data. When students are clustered within classrooms or within schools, observations made of those students cannot be said to be independent of one another (Raudenbush & Bryk, 2002). Multilevel designs and analysis can attend to the nested structure of the data, but as a result, randomization must typically occur at the cluster (e.g., either the teacher or school) level. Further, power calculations in multilevel analyses depend heavily on the number of clusters, and much less on the number of individual students within a cluster (Spybrook et al., 2011). It is not uncommon to require 25–50 *schools* in a study, depending on the nature of the design and the characteristics of a power analysis.

*Many Clusters: Methodological Implications.* Within the current funding climate, conducting adequately powered studies of causal effects within the confines of typical award parameters is more feasible when the travel between sites is not extensive. Specifically, researchers often need to find large numbers of participating schools *within the same geographic region*. If a research question requires including a large number of rural schools, it is likely that the research methods must be adjusted to fit the budget such that the researchers do not need to travel to the schools for frequent data collection or professional development. In other words, adequately powered studies of causal effects may be possible with rural schools provided that the intervention and data collection can be done remotely. Ultimately, however, funding constraints place limitations on either the types of schools recruited to participate in studies of causal effects or on the nature of the interventions or data collection techniques possible.

*Many Clusters: Implications Within Current Policy Contexts.* The recruitment effort generally required for studies of causal effects is substantial. Various policy factors can influence recruitment. School districts may be reluctant to try an intervention that they deem to be unproven in an environment where standardized test scores can make or break careers. Studies focused at the elementary-school level face an additional constraint. If science is not tested at the elementary school level in a given state, districts may eschew participation in science education interventions in favor of placing greater emphasis on reading, writing, and mathematics.

In causal effects studies at the elementary level, researchers might bolster recruiting and statistical power by expanding their target outcomes to include reading, writing, and mathematics along with science achievement outcomes. If a reasonable argument can be made that a science intervention would also improve student reading, writing, and mathematics abilities, then elementary schools in districts under pressure to produce high test scores might be more inclined to participate. A possible downside of this approach is the additional monetary costs associated with the increased data collection and analysis.

Further difficulties may arise with causal effects studies of curriculum interventions. Many districts have lengthy curriculum adoption procedures lasting 1–2 years. Districts often view the

decision to participate in a curriculum-based intervention study as akin to agreeing to adopt the curriculum. As a result, recruitment of districts for a curriculum intervention may be impossible within the time frame allotted for most research projects. One solution is to partner with districts when the research proposal is developed so the timelines of the research study and the curriculum adoption process can be coordinated.

A further consideration related to the need for many clusters and their geographic distribution arises from the need in many causal effects studies to measure program implementation—a factor that can significantly mediate or moderate the effects of an intervention on student achievement. While classroom observation techniques can be potentially objective and capable of measuring a range of classroom practices, they can be extremely expensive, require multiple observations of each classroom, and the presence of the observer or video camera can influence teacher practice and student behavior. On the other hand, teacher self-report methods (e.g., online surveys or implementation logs) might be more economical, but are subject to social desirability bias and limited by teacher's awareness of their own practice. Current research (Hill, Charalambous, & Kraft, 2012) is exploring trade-offs between efficiency, precision, and reliability in different approaches to measuring classroom practice and implementation fidelity.

*Description of Requirement 2: Economy of Scale on Outcome Measures*

Most studies of causal effects seek to estimate the effect of the intervention on student achievement. Indeed, many funding agencies *require* a student achievement outcome measure to be part of funded research. As such, researchers must find economical ways to measure student outcomes within the current funding parameters.

*Economy of Scale for Outcome Measures: Methodological Implications.* If, for example, 25–50 schools are recruited to participate in a study of causal effects, it is not uncommon to have 5,000–10,000 students in the study. Obtaining outcome measures on each of these students within the current funding constraints has important methodological implications. For example, objectively scored assessment items (e.g., true/false or multiple choice) administered in electronic format are usually fairly economical because the tests can be scored electronically. However, it can be difficult to assess important practices of science (argumentation, explanation development, etc.) using such items (NRC, 2012, p. 262; Pellegrino, Chudowsky, & Glaser, 2001, p. 194). Furthermore, it is time-consuming to develop assessment instruments that are sufficiently sensitive to detect treatment effects. Depending on the budget and timeline for a particular study, the researchers may default to using extant instruments, but by and large, the instruments are seldom valid for purposes of the research into which they have been imported (Wilson et al., 2012). As such, researchers may find that they have traded away instrument sensitivity for economy.

Additional challenges arise as researchers attempt to mitigate threats to validity associated with experimenter bias. In many experimental designs, funding agencies prefer the researchers distance themselves from the development of the outcome measures to ensure that the instruments do not favor the intervention group. Indeed, the criteria for inclusion in major syntheses of effective interventions (IES, 2011a; Johns Hopkins University, 2011) include specific requirements on the extent to which an outcome measure is fair to participants in all treatment conditions. Since the pool of high-quality assessments of students learning in science education is quite limited, one solution to this bias concern is to use an outside group to develop the outcome measures in a research study. However, external organizations that are not intimately familiar with the intervention may not be capable of designing instruments that would be sensitive to systematic differences in student achievement on targeted learning outcomes. Oftentimes, it is the researchers

themselves who would know best how to make instruments sufficiently sensitive to the treatment effect. One approach to reconciling this conflict is to employ a series of assessments or subscales that range in distance from the students' experiences with the intervention (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002, Ruiz-Primo et al., 2012). That is, the assessments include a set of "close" items that are very closely aligned with students' experiences with the intervention; "proximal" items that are aligned with the learning objectives of the intervention; and distal items that are aligned with national or state standards. Experimenter bias can be limited in this approach through collaboration with an external assessment developer to ensure appropriate item proximity while maintaining a bias firewall, or through using extant external instruments as the distal measures.

There has been some work with computer-based assessments or computer-based scoring to determine students' abilities to perform complex tasks such as designing controlled experiments (Sao Pedro, Gobert, Heffernan, & Beck, 2009) and explanation development (Nehm, Ha, & Mayfield, 2012; Nehm & Haertig, 2012), but the design of these scoring tools is expensive and the tools are generally applicable to narrow and specific content outcomes. As an alternative to the use of objective items or computer scoring, researchers may opt for investing a large portion of research dollars in manual scoring of a limited number of open-ended student responses. Using this approach, researchers would be limited to a small number of questions for each student. The researcher is thus faced with a trade-off between enhancing construct validity by assessing the desired complex scientific practices using a small number of items and risking the statistical conclusions of the study due to the potential unreliability of the measures.

*Economy of Scale for Outcome Measures: Implications Within Current Policy Contexts.* In addition to the multitude of methodological challenges associated with the development and use of sensitive, unbiased, and economical outcome measures, current state and federal policies can add additional obstacles. Understandably, both districts and funding agencies tend to be very interested in student outcomes on state-level measures of achievement. However, because such assessments frequently focus on lower order cognitive skills (NRC, 2012, p. 262; Pellegrino et al., 2001, p. 194), there can exist a significant disconnect between the goals of reform-based interventions and the high-stakes outcome measures used to evaluate them. This misalignment extends to the types of items that are used on the assessment. If an intervention is primarily designed to improve student abilities, such as argumentation or the development of explanations, state tests that are made up of primarily objective items are unlikely to be sensitive enough to measure differences in these higher level student outcomes over time or between treatments. While these issues highlight our concern that accountability policies associated with measuring student achievement may lead to studies of promising interventions failing to find effects, we should be equally concerned that in the current policy environment, demonstrating "What Works" is a lot easier for interventions that are merely effective at improving recall of vocabulary or factual scientific information.

### *Description of Requirement 3: Designs That Permit Attribution of Cause and Effect*

As described earlier, researchers can establish confidence in causal inferences by ruling out plausible alternative explanations for associations between key variables. Ruling out alternative explanations has direct implications for study design. In addition, attending to those design issues can create complications within school district policy contexts.

*Designs Conducive to Causal Claims: Implications for Methodology.* To rule out alternative explanations and promote high confidence in causal claims, researchers must typically include

random assignment of participants to groups, matching of groups on key variables, or both. Often, participating groups can be matched or "blocked" on key variables prior to randomization to enhance the power and face validity of the research and address problems associated with biased or unbalanced groups. This approach can capitalize on the positive features of both matching and randomization.

On the other hand, challenges can arise when attempting to implement randomized block designs. For example, blocking participating schools or districts prior to randomization (on a socioeconomic variable, for instance) requires that all participating schools be recruited and identified prior to randomization so that the blocking strata are unbiased. However, it can be difficult to recruit all sites in the study in time to match them on key variables prior to randomization. Once a school or district has "signed up," they typically are eager to know their random assignment, particularly in interventions that require the use of new curriculum materials or professional development. Schools and districts simply do not have the luxury of waiting until recruitment is finished to begin making plans for the next school year. At one extreme, a researcher can lose schools or districts from the pool altogether because they simply cannot afford to wait to learn their assignment. As a result, blocking may not always be practical, even if it is methodologically advisable. If blocking is essential, researchers might be able to address the methodological requirements by building in extra time for recruitment so that the recruitment phase can be concluded well before the study begins. Researchers essentially must weigh the gain in power by randomizing late and foregoing blocking (to include all late coming applicants) against the power gained by blocking when a suitable variable is available.

*Designs Conducive to Causal Claims: Implications Within Current Policy Contexts.* Policies associated with school accountability often create additional challenges for researchers attempting to carry out studies of causal effects. For example, it may be a school district policy that all schools in the district receive comparable resources, including professional development, curriculum materials, and lab supplies. These policies are in place so that districts can argue that no students are disadvantaged in preparation for state tests. By their very nature, studies of causal effects generally require that schools (or classes within schools) receive different treatments. Thus, districts may view the assignment of some schools to a treatment condition and others to a comparison condition as inequitable, particularly for longer interventions (those that may span a semester or more). The perceived inequities of the experimental design may dissuade districts from participating in the study, increasing the challenges associated with recruitment. In addition, large districts often spend extensive time and resources trying to get all teachers within a discipline teaching the same content using similar pedagogical approaches on a common timeline. Thus, a district's participation in a random assignment process can disrupt a district's efforts toward program consistency.

Research teams might be able to address the perceived advantages of the treatment condition by offering the comparison group viable program alternatives. In this case, the treatment contrast can be explicitly described but there may be a small treatment effect between the two promising interventions. If smaller treatment effects are anticipated, this inherently increases the recruitment demands on researchers, because more schools are needed to detect smaller effects (Spybrook et al., 2011).

Another option for encouraging schools or teachers to participate in a randomized control trial is to offer the intervention to the comparison group at a later date (after causal effects data have been collected). This solution has consequences of requiring additional time and funding to offer the intervention incentives at a later time.

To summarize, researchers conducting studies of causal claims face numerous challenges. There is an array of methodological implications for these studies. In addition, the climate in school districts as a result of current accountability policies can create additional challenges for researchers as they attempt to satisfy the methodological requirements of causal effects studies. In the next section, we present a case study that illustrates such trade-offs in a causal effects study of a curriculum and professional development intervention.

## Case Study: Identifying and Addressing Validity Threats in a Causal Effects Study

Mitigating threats to validity is essential in causal effects studies. When considering the threats, researchers may also find themselves challenged to do so within a specific budget and project timeline. In this section, we demonstrate the interplay of these forces using our experiences from a recent study conducted by the authors. We describe key categories of challenges, and how the research team balanced the threats to validity inherent within those challenges.

### *Context of the Case Study*

The case is an efficacy trial of research-based curriculum materials for multidisciplinary science in grades 9 and 10 (IES Grant #R305K060142). Teachers received seven days of professional development to support the implementation of the curriculum materials. The design was a cluster-randomized trial, where schools were the clusters randomly assigned to treatment or comparison conditions. Students at comparison schools continued with grade 9 and 10 science in a business-as-usual (BaU) manner. Because BaU science curricula differed on a school-by-school basis, the experiment did not have a single control condition *per se*. As an incentive to participate, BaU schools were offered the benefits of the treatment after the cohort of students in the experiment had matriculated to the next grade level. Most schools took advantage of this incentive. A key requirement for interpreting the findings was the ability to characterize what occurred in the BaU classrooms. BaU teachers were provided with a stipend to assist in these data collection activities. The study included 23 schools in two cohorts spanning 15 districts in a western state with a concentrated suburban area and many rural areas. Within these schools, 67 teachers, and 5,358 students participated in the study. Below, we describe main categories of research challenges and the associated threats to validity that we had to balance in this large trial.

### *Categories of Challenges*

*Choosing a Unit of Randomization: Trade-Offs Among Statistical Power, Treatment Group Contamination, and Other Threats.* The decision about unit of randomization is usually driven by the nature of the intervention. For example, while some software/online interventions can be manipulated for assignment at the student level, most curriculum materials must be assigned to entire classes of students because schools and/or districts approve certain curriculum programs and mandate their use for all students. It is also often impractical for a teacher to use multiple programs with the same class of students. Further, random assignment can happen at the class level (within a teacher) and though some teacher variables would be controlled in this design, such a design also introduces the threat that the teacher cannot keep key treatment elements from entering into comparison group instruction.

In this case study efficacy trial, a prospective power analysis suggested that we needed 22–24 clusters to detect the expected effect size. Recruiting 22–24 teachers as clusters is usually quite a bit easier than recruiting 22–24 schools as clusters. However, distinct validity threats were identified for random assignment at the teacher level. Several of these could be eliminated by switching to assignment at the school level. For our study, the benefits of school-level assignment

outweighed the drawbacks of increased recruiting efforts and expenditure of resources. The ultimate benefits of school-level assignment included the following:

- For treatment schools, it enabled teachers within a school to support each other and implement the intervention as a team.
- It reduced the chance that the BaU teachers might be contaminated by the treatment teachers' experiences with the intervention.
- It substantially reduced possible biases such as BaU teachers trying to outperform building-level colleagues in the treatment condition (compensatory rivalry).
- It helped in longitudinal scheduling of students from grade 9 to grade 10 in the curricular intervention.
- It allowed the researchers use a school-based manager to support researchers in collecting data from multiple teachers.

*Flux in State Standards and Assessments: Trade-Offs Among Fidelity, Measurement Sensitivity, and Project Costs.* In an age of increasing accountability, both content standards and assessment frameworks at the state level are subject to change with relatively little notice. This study experienced challenges associated with state-level policy changes in two states. In the first state, policy changes required us to abandon our work and start over in a second state. We experienced policy changes again during the course of the project in the second state. In each case, a state standardized test in multidisciplinary science initially served as an important outcome measure in the research design; however, both states changed to an end-of-course content exam in one subject (biology) as their state standardized assessment. Introducing a state biology test in the middle of the study had significant negative implications for expected levels of fidelity and for measurement sensitivity to detect an effect of a multidisciplinary science intervention.

To safeguard the study against changes in tests at the state level, the research team contracted to have a project-specific achievement measure developed. The test developer designed this achievement measure to be unbiased to both the treatment and BaU conditions. The team was able to administer this test to both cohorts, but developing, administering, scoring, and analyzing those data came with considerable monetary costs. Furthermore, the decision to use an external organization to develop the test had implications for instrument sensitivity.

*Obtaining and Tracking Consent: Trade-Offs Between Treatment Interactions With Populations and Treatment Interactions With Outcomes.* Several districts and schools wished to participate yet also had very stringent requirements for informed parental consent, particularly when the data collection plan included student surveys or feedback instruments. In the pilot phase of the research, this understandable requirement also posed considerable challenges for recruitment (due to low return rates of the consent forms). This was particularly challenging in schools with overall low socioeconomic status. Despite an array of incentives, and efforts throughout the pilot year, the research team was not able to exceed 10–15% return rate for some classes. As a result, we lost power to detect important treatment interactions (including analyses of achievement gaps).

The research team made an important, subtle adjustment for later phases of recruitment. Specifically, we worked with prospective districts to ensure that all elements and measures of the project were consistent with normal classroom instruction and assessment (measures of achievement only) and we provided research data that teachers could use in a formative fashion. After we tightened this alignment with normal school practices, the districts no longer required parental informed consent. A benefit of this trade-off was that the sample now had more

participants from underrepresented populations, increasing the generalizability of the treatment effect. On the other hand, the data set now included only achievement data, narrowing our ability to detect the full scope of outcomes that the treatment might affect.

*Working With External Researchers: Trade-Offs Among Experimenter Bias, Measurement Error, and Project Costs.* Reducing experimenter bias is a key part of any efficacy trial where the researchers work for the same organization that developed the intervention. In this case study, we used external experts in the following three ways:

- to analyze the curriculum materials in both treatment groups with a common set of curricular measures;
- to develop items for the project-specific assessment that were intended to be unbiased to both treatment groups and aligned with more general state science standards; and
- to conduct classroom observations of instructional practices for both treatment groups.

Use of external raters for curriculum materials and teacher practice helped protect the study from inadvertent experimenter bias. The external development of the achievement assessment helped blind the internal researchers and professional development providers to a key outcome measure. The trade-off for this bias protection was additional financial costs and increased uncertainty that the measure was sufficiently sensitive to detect treatment effects. That is, an external contractor is less likely to have intimate knowledge of the nuances of an intervention. Often these nuances are exactly what set the intervention apart from other programs; and therefore, should be the focus of the assessment.

*An Array of Stakeholder Interests: Trade-Offs Among Statistical Power, Generalizability, and Fidelity of Implementation.* Many stakeholders within and outside the schools have a variety of interests in educational outcomes. In this case study, such stakeholders ranged from district- and school-level stakeholders, to parties who were largely independent of the research. While the stakeholders all had very legitimate interests and concerns, those interests sometimes placed important limitations on the research.

*School Administrators*

Several schools instituted review days each Friday for standardized testing. This had an impact on dosage and fidelity to the intervention. That is, teachers' implementation of the intervention was 80% of what it might have been without the required test review days.

*Teacher Unions*

Some unions expressed concern that participation in the study would violate contractual agreements negotiated between the district and its teachers. Specifically, they objected to the extra (non-contract) time required for teachers to participate in professional development after school (even though teachers received a stipend). This created scheduling challenges for professional development and posed a threat to implementation fidelity.

*Athletic Directors and School Counselors*

These groups voiced concerns about participating, citing that the science intervention in the treatment group might not be accepted and approved by the NCAA, the service academies, or selective colleges and universities.

*District Personnel*

An assistant principal was concerned that there was a 25% chance (through one outcome of the random assignment) that the high SES school would get new books and professional

development a year ahead of the low SES school. After considerable discussion they decided not to participate.

*Teachers*

Teachers were concerned about how teaching the new 9th grade course (the unproven treatment intervention) would affect their performance evaluations.

This sampling of concerns illustrates that the ability to conduct causal effects studies can also be affected by local interests, initiatives, and policies and the corresponding threats to validity are serious. Thus, we needed to make a trade-off in this case study between maximizing power and accepting or retaining schools where local contextual factors may not favor high-fidelity implementation. This trade-off was a function of the complex interplay of the various stakeholder interests.

## Discussion

As a result of federal education policies in support of funding research, we have seen a marked increase in the number of causal effects studies funded. For example, in the year 2000, there was only one program evaluation funded by the U.S. Department of Education that used an experimental design (Boruch, DeMoya, & Snyder, 2002). In the decade since NCLB and the Education Sciences Reform Act, the number of funded studies testing causal effects has been on the rise. However, causal effect studies are still somewhat underrepresented in the body of science education research. For example, only 6 of 45 research articles in the 2011 issues of JRST described studies of causal effects (experiments or quasi-experiments). Further, at the time this manuscript was first prepared, just one-third of the studies in the IES Mathematics and Science Education Research portfolio were the types of studies that typically seek causal effects—then called *Efficacy and Replication Studies* and *Scale-up Evaluations* (IES, 2011b). Certainly, this stems from the fact that there are many important, researchable questions in the field, not just those of cause and effect. However, another possible explanation for this modest number is that the typical scale of causal effects studies makes them more time- and resource-intensive than studies exploring other research questions. In addition, researchers are often reluctant to propose them because they require extensive resources and expertise to balance the trade-offs among threats to validity. Further, some researchers do not see the research designs associated with causal effects research as feasible for studying certain types of interventions or for comparing certain types of outcomes, even if causal inferences are desired. As a result, the time period we should expect for the development of a rich and comprehensive portfolio of causal effects studies in science will likely be measured in decades, not years. To make more rapid progress in building a body of evidence about effective programs, we suggest that a multifaceted shift in perspective is needed.

One type of shift would be placing more emphasis on the structures within the science education community intended to increase opportunities for researchers to grow and collaborate. Some of the reluctance of researchers for conducting causal effects studies may come from being isolated from colleagues with similar research questions. Successfully implementing causal effects studies usually requires expertise beyond that developed in science education graduate programs. It is also unusual for a single researcher to have all of the necessary expertise, making a team approach to these projects most effective. The community should be continually looking for ways to provide methodological training and networking for all types of research methods (quantitative or qualitative), and to promote a more open dialogue on these issues. An example would be the IES-sponsored institute on designing and implementing cluster randomized trials (IES, 2013). NARST might consider a dedicated strand for its annual conference that focuses on research methodology, in general, and research design and measurement, in particular.

The policy and research fields could also benefit from a more pragmatic view of differing research designs, including stronger bridges between qualitative and quantitative research traditions. A truly powerful body of evidence that can inform policy decisions on effective programs and practices must include findings from both causal studies on the magnitude of effects as well as in-depth qualitative findings on why and how interventions may work. By bringing together a broad body of research traditions, the field may be better positioned to evolve and ultimately improve science education for all students.

At the state level, we suggest a shift in emphasis from a set of achievement tests that, more often than not, focus on factual recall, to measures that emphasize higher order cognitive outcomes such as scientific reasoning and argumentation. This will allow us to target the types of outcomes that the science education community values, as illustrated by the Next Generation Science Standards (NGSS Lead States, 2013). When schools and districts find that interventions with reform-based aims could plausibly increase student scores on state assessments, recruitment and hence statistical power for causal effects studies will improve. There are, of course, budgetary considerations. NCLB was implemented during a time when many states had to cut education budgets. High-stakes assessments often focused on the types of student thinking that were affordable to measure—most commonly factual recall via multiple-choice items (Bracey, 2005). The additional resources required to measure higher order thinking were not available. Our increasing ability to score open-response items with more economical approaches such as lexical analysis and scoring holds promise for positive change (Attali, Powers, Freedman, Harrison, & Obetz, 2008; Nehm & Haertig, 2012).

Finally, we need a collaborative community of school district personnel and researchers that truly embrace a spirit of experimentation and scientific inquiry about effective education interventions. That said, it is naïve to think a spirit of experimentation should be the first priority of school district personnel. Thus, the conversation should address the urgent needs of school districts as much as the merits of long-term intervention research. The apparent misalignment between the goals of researchers and practitioners can be improved with respectful dialog.

A spirit of experimentation can only thrive in the absence of punitive uses of state achievement tests. Using state achievement test scores for accountability purposes has its benefits when implemented as a carrot, not a stick. No school district should receive a permanent reprieve from accountability for student progress toward outcomes. However, state- and federal-level decision makers could consider plans where selected diverse and reform-minded school districts are designated as research sites with temporary dispensation from state accountability testing. Some related federal efforts are already underway and these could make a significant contribution (The White House, Office of the Press Secretary, 2012). When the fear of the stick disappears, true research collaboration and educational innovation will ensue; as a result, we will be able to test and learn about what really works.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305K060142 to BSCS. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors would like to acknowledge the invaluable contributions of Karen M. Askinas toward this article.

## References

American Association for the Advancement of Science (AAAS). (1993). Benchmarks for science literacy. New York, NY: Oxford University Press.

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). Automated scoring of short-answer open-ended GRE subject test items, GRE Board Research Report No. 04-02 (ETS RR-08-20).

Bill and Melinda Gates Foundation. (2012). Primary sources: 2012 America's teachers on the teaching profession. Retrieved from http://www.scholastic.com/primarysources/pdfs/Gates2012_full.pdf

Boruch, R. F., DeMoya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. Boruch (Eds.), Evidence matters: Randomized field trials in education research (pp. 50–79). Washington, DC: Brookings Institution Press.

Bracey, G. W. (2005). No child left behind: Where does the money go? Education Policy Research Unit, Arizona State University. Retrieved from http://nepc.colorado.edu/files/EPSL-0506-114-EPRU-exec.pdf

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. Educational Researcher, 31(8), 4–14.

Getty, S. R., Wilson, C. D., Taylor, J. A., & Kowalski, S. M. (2011 , April). Managing threats to validity in experimental tests of education interventions: Data and evidence from a large, cluster-randomized trial (CRT) of a high school science intervention. Paper presented at the National Association of Research in Science Teaching (NARST), Orlando, FL.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. Educational Researcher, 41(2), 56–64.

Institute for Education Sciences (IES). (2011a). What Works Clearinghouse procedures and standards handbook (v. 2.1). Retrieved from http://www.ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

Institute for Education Sciences (IES). (2011b). Education research grants request for application (CFDA Number: 84.305A). Retrieved from http://ies.ed.gov/funding/pdf/2012_84305A.pdf

Institute for Education Sciences (IES). (2013). Summer research training institute: Cluster-randomized trials. Retrieved from http://www.ies.ed.gov/whatsnew/conferences/?id=1049

Johns Hopkins University Center for Data-Driven Reform in Education (CDDRE). (2011). Best evidence encyclopedia. Retrieved from www.bestevidence.org

Lee, O., & Krajcik, J. (2012). Large-scale interventions in science education for diverse student groups in varied educational settings. Journal of Research in Science Teaching, 49(3), 271–280.

Marx, R. W. (2012). Large-scale interventions in science education: The road to utopia? Journal of Research in Science Teaching, 49(3), 420–427.

McDonald, S.-K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. Educational Researcher, 35(3), 15–24.

National Research Council (NRC). (1996). National science education standards. Washington, DC: National Academy Press.

National Research Council (NRC). (2000). Inquiry and the national science education standards. Washington, DC: National Academy Press.

National Research Council (NRC). (2002). Scientific research in education. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.), Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council (NRC). (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences. Washington, DC: National Academy Press.

National Science Foundation. (2011). Discovery research K-12 (DRK-12) program solicitation (Publication No. 11-588).

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. Journal of Science Education and Technology, 21(1), 193–196.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: Testing the efficacy of text analytic software. Journal of Science Education and Technology, 21(1), 56–73.

NGSS Lead States. (2013). Next Generation Science Standards: For states, by states. Washington, DC: The National Academies Press.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Penuel, W. R., & Fishman, B. J. (2012). Large-scale science education intervention research we can use. Journal of Research in Science Teaching, 49(3), 281–304.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M.-C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. Journal of Research in Science Teaching, 49(6), 691–712.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. P. (2002). On the evaluation of systematic science education reform: Search for instructional sensitivity. Journal of Research in Science Teaching, 39 (5), 369–393.

Sao Pedro, M. A., Gobert, J. D., Heffernan, N. T., & Beck, J. E. (2009). Comparing pedagogical approaches for teaching the control of variables strategy. In N. A. Taatgen & H. van Rijn (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Schneider, B. H., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). Estimating causal effects using experimental and observational designs. Washington, DC: American Educational Research Association.

Schwandt, T. A. (2005). A diagnostic reading of scientifically based research for education. Educational Theory, 55(3), 285–305.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin Company.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. Educational Researcher, 37(1), 5–14.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design Software Version 3.0. Retrieved from www.wtgrantfoundation.org.

The White House, Office of the Press Secretary. (2012). President Obama: Our children can't wait for congress to fix no child left behind, announces flexibility in exchange for reform for ten states [Press release]. Retrieved from http://www.whitehouse.gov/the-press-office2012/02/09/no-child-left-behind-announces-flexibility-exchange-reform-ten-states

U.S. Congress. (2001). No Child Left Behind Act of 2001 (Public Law 107–110). Washington, DC: Government Printing Office.

U.S. Congress. (2002). Education Sciences Reform Act of 2002. Washington, DC: Government Printing Office. Retrieved from http://www2.ed.gov/policy/rschstat/leg/PL107-279.pdf

Wilson, C. D., Roth, K. J., Taylor, J. A., Landes, N., & Stuhlsatz, M. (2012 , March). In search of instructional sensitivity: The measurement problem in large scale studies of professional development programs. Paper presented at the annual conference of the National Association of Research in Science Teaching (NARST), Indianapolis, IN.

Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge. Journal of Research in Science Teaching, 47(3), 276–301.

Zellmer, M. B., Frontier, A., & Pheifer, D. (2006). NCLB: Taking stock, looking forward. Educational Leadership, 64(3), 43–46.