Designing the Impact Evaluation Component of a Multi-Arm Trial of Enhanced Middle-school Science Curricula[1]

Morgan S. Polikoff, Rebecca Maynard, & Bob Boruch
University of Pennsylvania

Presented at the 2010 AERA Research Conference, Denver, CO
May 3, 2010

The Institute of Education Sciences, created under the Education Sciences Reform Act of 2002 (PL107-279; http://ies.ed.gov/pdf/PL107-279) raised the bar for creating, disseminating, and applying scientific evidence regarding the effectiveness of educational policies and practices supported through public funds (Whitehurst 2008). The Institute also supports a number of research centers and individual research and development initiatives aimed at improving curriculum and instruction, particularly in Science, Technology, Engineering, and Mathematics (STEM) and literacy (see, for example, a list of currently funded projects in math and science at http://ies.ed.gov/ncer/projects/program.asp?ProgID=12). The first major research projects launched under *The Twenty-first Century Research and Development Center on Cognition and Science Instruction* (henceforth, referred to as the *21st Century Science Project*) focus on developing and testing new middle school science curricula that apply theoretically-based cognitive science principles in the design of the curriculum and in the preparation of teachers for delivering the curriculum. The goals of the curriculum modifications and accompanying teacher professional development are both to improve substantially student learning of the specific scientific principles reflected in the curriculum modules that are modified and to improve students' ability to master scientific principles taught subsequently using conventional methods. The primary goal of the impact evaluation component of the multi-method study of the

[1] The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305C080009 and R305C050041. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

demonstration curriculum modules and teacher professional development is to estimate their impacts on student learning and to provide evidence to gauge whether or not the principles that under-girded their development are likely to be applicable to science curricula at all levels. A secondary goal is to estimate impacts on mediators of student achievement—curriculum and teacher knowledge.

This paper begins by laying out the foundational conditions that shaped the impact evaluation design and discussing their implications for designing the study sample and data collection plan. It then proceeds to a discussion of the sample and timeline of the parallel randomized trials. Finally, the planned analytical strategies are discussed.

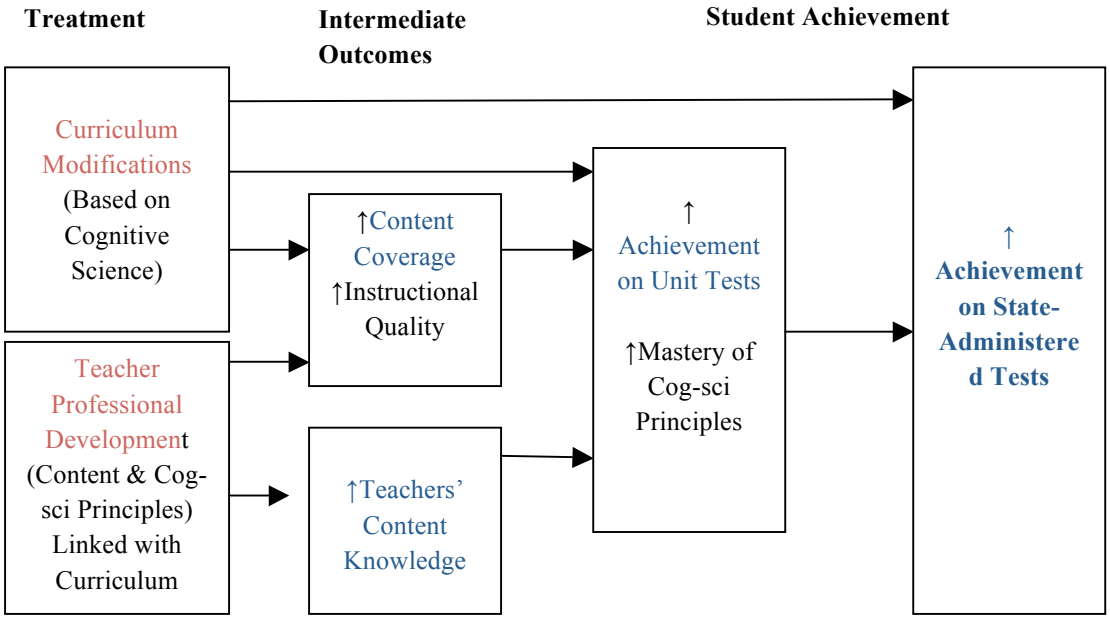
Conditions that Shaped the Study Design

Two factors were especially important in establishing key design parameters: (1) the nature of the interventions being tested; and (2) the imperative that the study results provide strong evidence regarding whether or not the interventions *caused* changes in instruction and student learning.

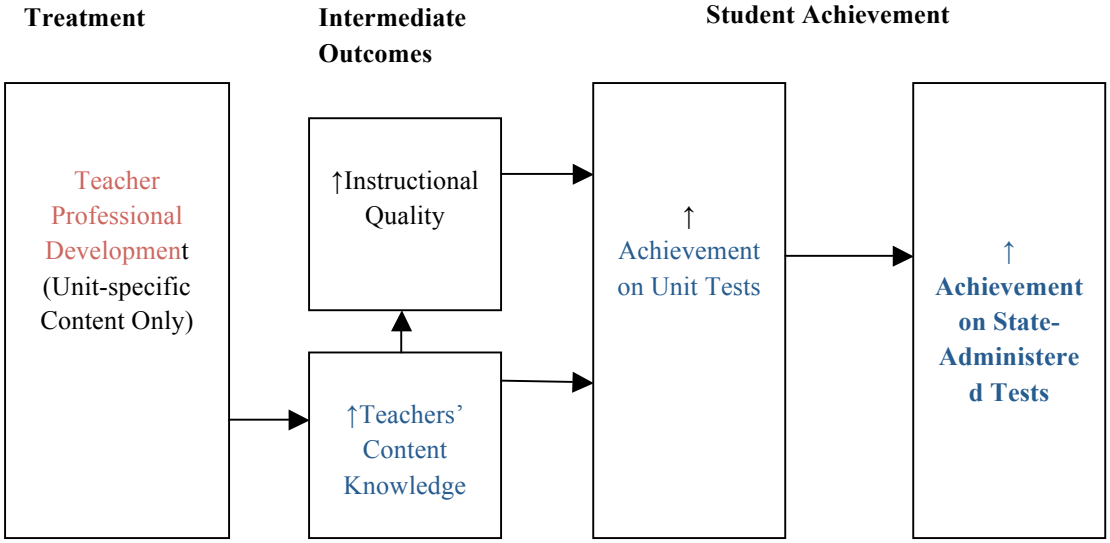
The interventions entailed curricula reforms and teacher professional development targeted at the middle school grades. In so far as smaller schools often have only one middle school science teacher and in other cases there is *team* or coordinated instruction, the interventions are best considered to be *school-level interventions*. The underlying logic of the planned curriculum reforms and accompanying teacher professional development is depicted in Figure 1.

Figure 1: Program Logic

a. Full Treatment (T1 = Curriculum Modifications Plus Teacher Professional Development



b. Limited Treatment (T2 = Unit-specific Content Professional Development Only)



The requirement that the study produce strong (i.e., unbiased and reasonably precise) evidence regarding the impacts (or lack thereof) of the interventions leads to designing the evaluation using an *experimental design* (Boruch 1997; Orr 1999; Rossi et al. 2004; Shadish et al. 2002). Alternative designs, such as comparisons of instruction and student outcomes before and after the changes in curriculum or comparisons of outcomes for students in schools that do and do not adopt the curriculum modifications (and, in some cases the enhanced teacher professional development) would leave open questions regarding whether observed differences between the "treated" and "untreated" groups are attributable entirely to the intervention.

These two requirements together dictate that the study rely on a *cluster random assignment design* in which impacts are estimated by comparing teachers/classrooms and students in schools that were *randomly assigned* to the treatment group and schools that were randomly assigned to "business as usual." Other details of the study design depend on a complex set of factors, including the following: (1) the specific outcomes that will be used to judge whether or not the curriculum modifications and professional development *worked* as hypothesized and properties of the constructs that will be used as indicators of these outcomes; (2) the lag time until meaningful impacts are expected to be observed, which may differ by outcome and the units to which they apply; and (3) minimum size impact of the interventions that would have relevance for policy or practice. Notably, the cognitive scientists responsible for developing the curricula modifications and complementary professional development programs, the education researchers seeking to identify strategies for promoting better science education outcomes among students, and district leaders who were especially concerned with strengthening their science education offerings such that school level performance on state-tests improve would frame priorities for the impact evaluation somewhat differently. In cases where differences are

relevant for study design decisions, it was necessary to carefully weigh the trade-offs and preserve the ability of the study to address the questions that were most central to the interests of each constituent group.

Background

As specified in the call for proposals for the *21st Century Science Center*, the curricula and professional development strategies being evaluated were intended to improve student learning in science by developing a set of theory-based modifications to existing curricula that, if found to be effective, could be applied more broadly. In addition, the call for proposals laid out two additional requirements for the project. The first was that the modifications should apply to entire curricula units. The second was that the study evaluating the effectiveness of the modifications should include both a standardized test that is not specifically aligned with the content of the modified curricula and an aligned measure. In addition, the call for proposals stipulated that the study should examine impacts on mediators, such as curriculum and teacher knowledge. Specific evaluation requirements included the following:

- Modify large chunks of curricula in commonly-used science curricula based on theoretically-grounded cognitive science principles shown in prior research to improve student learning;
- Convey the modifications to teachers in a randomly assigned treatment group, enable teachers to deliver the modified curricula to students, measure the fidelity with which the curricula were implemented;
- Measure student learning using both a standardized assessment not specifically aligned to the units and a researcher-developed assessment aligned with the content of the modified units.

To these requirements, the evaluation team added the goal that the study also should examine the implications of the curricula modifications and teacher professional development enhancements on the instructional content, climate, and quality and on teacher knowledge of the scientific principles undergirding the focal science units and the theory-grounded modifications.

Critical Features of the Impact Evaluation Design

The overall integrity of the impact evaluation necessarily rests on the quality of the cognitive science that is the basis for the curricula modifications and the strategy for their implementation. To this end, the project team includes group of cognitive science experts—Dr. Chris Schunn from the Learning Research and Development Center (LRDC) at the University of Pittsburgh; Dr. Christine Massey, Director of the Institute for Research in Cognitive Science (IRCS) at the University of Pennsylvania; and Dr. Jennifer Crowley from the NSF-funded Spatial Intelligence and Learning Center at Temple University. The impact evaluation team is charged with designing and conducting a rigorous evaluation of the curricula modifications and related professional development designed and implemented by that team. We prioritized four qualities of the impact evaluation: (1) using quality measures of the relevant outcomes; (2) ensuring the resulting estimates of program impacts had internal validity (i.e., they are unbiased estimates of the true effects of the curricula modifications and professional development); (3) having reasonable levels of statistical precision in the impact estimates (i.e., confidence intervals around impact estimates are sufficiently small that meaningfully-sized impacts will not be overlooked due to sampling error); and (4) reasonable levels of external validity (i.e., the findings will be applicable to a reasonably broad, identifiable group of schools and school settings).

The Intervention

The request for applications from IES specified that the project should investigate the role of cognitive science in improving science curricula, but it was not prescriptive with respect to the focal cognitive science principles that should guide those modifications. The research team chose to focus on curricula modifications that addressed analogical reasoning, diagrammatic reasoning, and the role of background knowledge in learning, each of which has been shown to be related to understanding, knowledge growth, and transfer (for example, see Silk, Schunn, & Carey 2007; Cromley et al. 2007; Chi 2005; Chinn and Malhotra 2002; Ross and Kennedy 1990; VanLehn 1998; Gentner 2001). For a more thorough description of these principles and the reasons for their selection, see the first paper in the symposium by Schunn and colleagues.

These modifications were applied to selected units of both the Holt and Full Option Science System (FOSS) science curricula (Table 1). The underlying theory of change suggests that three particular types of modifications to commonly-used curricula like Holt and FOSS would lead to improved science achievement by students. Those modifications would be ones that (1) build *analogic reasoning* (for example, through contrasting case instruction and exercises based on the pioneering work of Bransford & Schwartz 1999), (2) build *diagrammatic reasoning* (for example, to address difficulties pointed out by Hegarty et al. 2003 that students have with visual representations in the Holt and FOSS materials), and (3) address the limited background knowledge and misconceptions students often face (Chi 2005). A competing theory suggest that improved student outcomes can be achieved using the current curricula, but by providing teachers with deeper understanding of the key science concepts in the science units they are teaching. Thus, the study focuses on two treatment variations: (1) teacher professional development to improve their understanding of the content they are teaching; and (2) curriculum modifications plus teacher professional development focused on content knowledge and the

cognitive science principles undergirding the curriculum modifications. The former group is referred to as the *limited treatment group* (content professional development only) and the latter is referred to as the *full treatment group* (curriculum modifications and professional development on the cognitive science principles). The impact evaluation has been designed to measure differences in instructional and student outcomes between these two treatment groups, but also between each of these treatment groups and a *business as usual* control group (neither curriculum modifications nor special professional development).

Table 1: Curricula and Units Modified

	Curriculum	
Units	Holt Middle School Science Short Course Series	Full Option Science System (**FOSS**)
1	Cells (Grade 7)	Weather and Water (Grade 6)
2	Introduction to Matter (Grade 8)	Earth History (Grade 6)
3	Inside the Restless Earth (Grade 8)	Diversity of Life (Grade 7)

Expected Outcomes

The impact evaluation includes four sets of outcome measures (Table 2). Two pertain to the ultimate outcomes of interest—student achievement—and two pertain to intermediate outcomes—instruction and teacher knowledge. One set of the student achievement measures was developed specifically for the study. This consists of constructed tests that are aligned with the learning goals of the six curriculum units that were modified using the cognitive science principles. These tests are administered to students immediately following their completion of the related curricula units. The other set of student achievement measures include prevailing state administered science achievement tests. These tests are administered is spring of the eighth

grade by all schools in the study sample as part of the state assessment system under No Child Left Behind (NCLB).

Table 2: Outcome Measures

Outcome Measure	Frequency	Timing
Student outcomes		
Aligned student achievement tests (evaluator developed)	6, once for each of three unit in each of two curricula (Holt and Foss)	Immediately after instruction in the target unit
State administered tests	2, once for each state	Spring of 8th grade
Intermediate outcomes/mediators		
Curriculum-aligned teacher knowledge tests (evaluator-developed)	6, once for each unit	After first year's content professional development
Survey of Enacted Curriculum (Porter 2002)	2-3 times per year per teacher	Immediately before and after modified curriculum units and at end of the year

In order to measure impacts on the intermediate outcomes (expected mediators), the evaluation will gather data on the teachers' knowledge of the science content in the focal units using teacher knowledge tests that are aligned with the content of the content professional development offered to the limited treatment teachers (which is itself aligned with the content of the modified curriculum units). These tests are administered to teachers in all treatment conditions at the end of the professional development cycle for each unit. The nature and quality of the curriculum as enacted by classroom teachers will be measured using the *Survey of Enacted Curriculum* (Porter 2002).[2] The SEC will be administered immediately before and after the

[2] Alignment on the SEC is defined at the intersection of topics and cognitive demands. Two sources (e.g., the content of instruction, the content of an assessment, the content of a curriculum) to the extent that the sources agree with one another in terms of the proportions of their total content associated with each particular topic/cognitive demand combination.

completion of instruction on the focal curriculum units for the study and again at the very end of the school year.

Both the state administered achievement tests and the Survey of Enacted Curriculum have well-established psychometric properties that are consistent with the planned use of them in this study. It is incumbent upon the research team to establish the psychometric properties of those measures that are being newly developed for this study. Tests of student achievement and teacher knowledge will be analyzed using classical test theory and will include estimation of scale reliability and scale development using factor analysis. We will use data on the content of teachers' instruction (from the SEC) to investigate the sensitivity of the assessments and items to content coverage.

Internal Validity, External Validity, and Precision

Our approach to ensuring internal validity, precision, and external validity of the study findings rests on the sample design for the study. To ensure *internal validity*, the study was designed as a cluster randomized control trial whereby school are recruited into the study and, subsequently randomly assigned to one of the three conditions: (1) *full treatment*, which includes curriculum modifications and professional development support for implementing them; (2) *limited treatment*, which consists of teacher professional development aimed at improving teacher content knowledge relevant to the curricula units that are modified for the full treatment (but not for this group); and (3) a *business-as-usual* (control condition). Assignment to condition was accomplished by assigning a random number to each participating school then, within strata (see below) assigning successive triplets to random permutations of the three treatment conditions (T1, T2, and C, respectively).

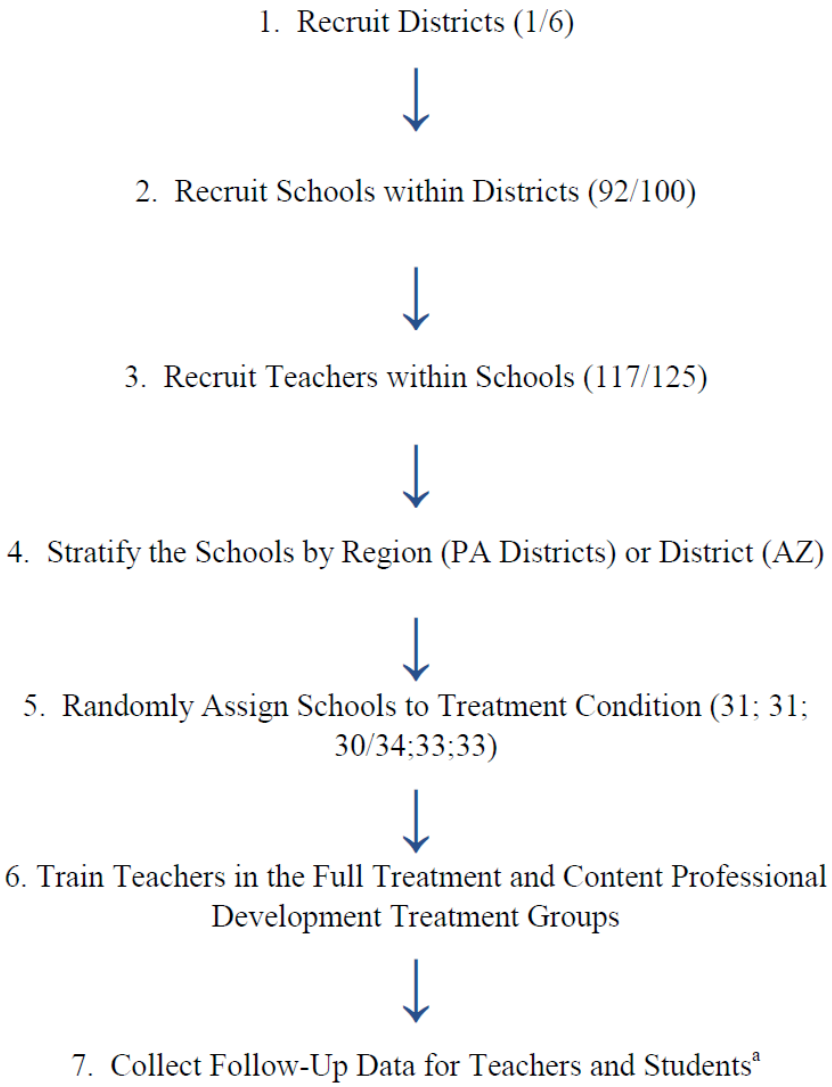
There was a five step process of recruitment and enrollment of the study sample, that began with recruiting partner districts and ended with randomly assignment of schools to treatment condition (Figure 2). Following assignment of schools to condition, the participating teachers in the treatment schools received their assigned professional development and curriculum modifications prior to beginning to teach the focal units. Each teacher is expected ot teach all three curriculum units for two successive cohorts of students, as described further below in the section on data collection.

External *validity of the study* is enhanced by conducting the tests in the context of two different curricula— Holt and FOSS—and by implementing the study in a large and diverse set of schools. The project includes two distinct impact studies—one focused on schools that use the Holt curriculum and one focused on schools that use the FOSS curriculum. The study schools for the study sample are clustered within eight school districts—Philadelphia, which uses the Holt curriculum, and seven districts that use FOSS—Pittsburgh, PA, and six districts in Arizona (Dysart, Deer Valley, Tucson, Sunnyside, Madison, Buckeye). Schools within and across districts vary in terms of size, race/ethnic composition, student achievement levels, and grade configurations.[3] Within each of the test settings (i.e., Holt versus FOSS), schools are stratified by region (Philadelphia) or district, and then randomized to treatment condition within strata. Stratification was helpful in securing district and school-level buy-in in so far as it

[3] The School District of Philadelphia, the eighth largest district in the nation, is divided into regions encompassing diverse neighborhoods, each with its own superintendent. Most schools in Philadelphia containing grades 6-8 are in fact K-8 schools. The district is approximately 64% African American, 16% Hispanic, 13% White, and 6% Asian. The Pittsburgh School District is a moderately sized urban district of approximately 30,000 (60% African American, 35% White) containing roughly half K-8 schools and half 6-8 schools. The districts in Arizona include urban, suburban, and rural/suburban districts in the Phoenix and Tucson areas that vary from 88% Latino to 70% White, depending on the district. Again, these schools vary in terms of their grade makeup, with some districts containing K-8 and some districts containing 5-8 or 6-8 schools.

ensured that the coveted treatment was relatively evenly distributed throughout the partner districts.

Figure 2: Recruitment and Enrollment of the Study Sample (Estimated # for Tests in the Context of the Holt Curriculum/ # for Tests of FOSS Curriculum)



[a]The goal is to test all students are assigned to science classes including the units selected for modification that are taught by teachers in the study sample.

Also, the fact that the curriculum adaptation and professional development are being tested in the context of two widely used curricula—Holt and FOSS—that have differing orientations toward science instruction and content enhances the generalizability of the findings. Holt is a fairly traditional, textbook-centered science curriculum, published by Houghton Mifflin Harcourt. The Full Option Science System (FOSS), created by researchers at UC Berkeley's Lawrence Hall of Science employs an active, inquiry-based approach to science, based on the belief that "The best way for students to appreciate the scientific enterprise, learn important scientific concepts, and develop the ability to think critically is to actively construct ideas through their own inquiries, investigations, and analyses" (http://www.lhsfoss.org).

Achieving adequate statistical *power* and *precision* of the impact estimates depends on the size and allocation of the study sample across schools, teachers, and students. What constitutes adequate power and precision of estimates depends on several factors. One is the tolerance of the scientific and policy community for mistakenly concluding that there are impacts of the treatments when there are not and for failing to detect true impacts. Standard convention is to use 90 percent confidence intervals (no more than 5 percent chance of concluding, in error, that there was a favorable impact and a similar chance of concluding that there was an unfavorable impact) and to aim for 80 percent power (no more than 20 percent chance of missing a true impact). Another determinant of power and precision is the error variance in the outcome measures and the sampling error, which depends on the variance in the outcome measure within and across sampling units for the study (Hedges and Hedberg, 2006; Bloom, 2005; Spybrook, 2007). The factor that is most in the evaluators' control is the sample size and allocation across treatment conditions.

In this study, because there are multiple outcomes of interest, it is not clear, a priori, which will present the most limiting constraints. However, in general, the most policy relevant and limiting outcomes would be those measured by state administered tests. Thus, this study was designed to have adequate statistical power to detect meaningful-size (i.e., .20 standard deviations) impacts on the state administered achievement tests. Under reasonable assumptions, achieving the target power and precision requires a study sample of about 50 schools in each treatment condition or 150 total (Table 3, column 1).[4] (Note: At the mean, a .2 standard deviation gain in achievement translates into the equivalent of moving from the 50th percentile in achievement to the 58th percentile.) Even fewer schools would be needed to obtain relevant size impacts on aligned tests, since only relatively large impacts (e.g., .3 standard deviations or larger or the equivalent of moving from the 50th to the 62nd percentile) would have practical significance. It is estimated that it would require 39 schools per condition or 117 total to detect impacts of this magnitude (Table 3, column 2). And, assuming that the minimum size impacts of relevance for teacher knowledge was even larger (e.g., .5 standard deviations or moving from the 50th to the 70th percentile), a sample of about 33 schools per condition or 99 total would be required (Table 3, column 3).

Sample sizes, by Curriculum and state

By design, the target study sample is 25 percent larger than required to meet the minimum targets to ensure adequate precision and power to detect that are equal or exceed the size judged to be the minimum relevant size to have practical or policy relevance-- about 200 schools rather than 150, for example (Table 4). However, this larger sample has the advantage of providing "insurance" against adverse events, such as higher than anticipated sample attrition

[4] The assumption in the evaluation design is that it is equally important to estimate differences in outcomes between the two implementation models (with and without teacher professional development focused on the cognitive science principles that were foundational for the curricula modifications.

(particularly at the school level), and increase the strength of moderator analyses and exploratory mediator analyses. Importantly, it will not make sense to pool the study sample for analyses of the unit-aligned student tests.

Table 3: Estimates of Minimum Sample Sizes to Detect Impacts of Minimal Relevant Size

	Outcome Measures		
	State Administered Student Tests	Aligned Student Assessments	Classroom-level Instruction and Teacher Knowledge
Standards of evidence			
1. Significance level for null hypothesis test	10%	10%	10%
2. Tails on the significance tests?	2	2	2
3. Statistical power Ssandard	80%	80%	80%
4. MRI-ES (Minimum relevant impact in standard deviation units)	0.20	0.30	0.50
Properties of the study sample			
5. Nc (average number of analysis units per cluster)	25.0	25.0	1.1
6. Intraclass correlation (ICC)[a]	0.15	0.10	0.10
7. P_T (proportion of the sample allocated to the intervention condition)	50%	50%	50%
8. R^2 (background variables included in student or teacher -level models)	60%	50%	50%
9. R^2 (background control variables used in the school-level models)	20%	10%	0%
10. Number of group level background control varaibles used	5	5	0
11. Sample retention for the analysis sample at follow-up	75%	90%	80%
12. Sample retention rate for the school level observations	100%	100%	100%
Estimated Minimum Required Number of Schools	**109**	**15**	**62**
Estimated Minimum Required Number of Analysis Unit	**2730**	**375**	**68**

Note: Computation of the sample size requirements used the sample size calculator included in Maynard and Orr (2010). The formulas underlying the calculator are based on those reported in Bloom (2006). The Core Analytics of Randomized Experiments for Social Research. MDRC Working Papers on Research Methodology. Available online at:http://www.mdrc.org/publications/437/full.pdf .

[a]For general guidance in reasonable estimates of ICCs for analysis of different population groups and focused on different outcome measures, see Bloom et al. (2007).

Table 4: Projected Study Sample

Curriculum	State/City	# of Districts	# of Schools	# of Teachers	# of students
Holt	PA, Phila	1	92	117	9,200
FOSS	PA, Pitt	1	27	35	1,440
	AZ	5	73	90	6,000
Total		7	192	242	16,640

Note: Sample enrollment is still ongoing in Pittsburgh and Arizona. Numbers of students are projections of the numbers of students who potentially could complete the 8th grade science test.

Analysis

Preliminary analyses will be conducted to assess data quality, including looking for imbalances in the samples assigned randomly to the treatment conditions, comparing the study schools to other district schools that opted out of the study, and examining the extent and implications of sample attrition.

The core analyses will be based on a hierarchical models that account for the nesting of students within classrooms and teachers and the nesting of teachers within schools. In all analyses, the treatment indicator variable will be entered at the school level, since schools were the unit of analysis. Importantly, the impact estimates will reflect the consequences of having been assigned to the treatment group, regardless of whether some teachers fail to participate in their assigned intervention (i.e., the analysis will generate what is commonly referred to as "intent to treat" (ITT) estimates). Student impact analyses will use three level models: (1) students; (2) teachers; and (3) schools. Other analyses (e.g., those focusing on instructional qualities and teacher knowledge) will use two-level models: (1) teachers/classrooms and (2) schools. Table 5 describes the types of moderator and control variables that will be included in the analyses of each of the four types of outcomes—(1) Student performance on state administered achievement tests; (2) student performance on end of unit tests; (3) teachers' knowledge of unit content, and (4) the nature of the curriculum as enacted by teachers.

Table 5: Illustrative Analytic Models

	Outcome Measure				
	State Achievement Test Scores	Curriculum Aligned End of Unit Tests	Teacher Unit Content Knowledge	Enacted Curriculum within Units	Enacted Curriculum Overall
Unit of Analysis	Students within State	Students within Curriculum Groups	Teachers within Curriculum Groups	Classrooms within Curriculum Groups	Classrooms
Moderator Variables					
Curriculum (Holt or FOSS)	X				X
State (PA or AZ)		X	X	X	X
District/Region	X	X	X	X	X
Cohort (1 or 2)	X	X	X	X	X
School-level Control Variables					
Enrollment	X	X		X	X
Grade Configuration	X	X		X	X
% Proficient on 8th Grade Tests	X	X		X	X
% Eligible for Free-Reduced Lunch	X	X		X	X
% English Language Learners	X	X		X	X
% Minority	X	X		X	X
Teacher-level Control Variables					
Years of Experience	X	X	X	X	X
Gender	X	X	X	X	X
Student-level Control Variables					
Grade	X	X			
Age	X	X			
English Language Learner Status	X	X			
Race/Ethnicity	X	X			
State Achievement Test Results	X	X			

In addition to the impact analyses, we also will conduct exploratory correlational analyses to understand the relationship between various qualities of the intervention (e.g., fidelity of implementation or dose) and impacts on intermediate outcomes and student achievement. In cases where multiple indicators of an outcomes are estimated using the same sample, the Benjamini-Hochberg approach to controlling the false discovery rate will be applied (Schochet 2008).

Timeline

This is a five year project, scheduled to conclude in 2013 (Figure 3). The first program year (2008-09 school-year) was devoted to development of the modifications for the Holt curriculum and recruiting sites. Piloting began in spring 2009 for the Holt modifications and in spring 2010 for FOSS. Since the FOSS modifications pertain to units typically taught in sixth and seventh grades, the final data collection will not be completed until spring 2013. However,

prior to that time, we will have preliminary findings based on the Holt modifications and the first cohort of teachers and students enrolled in the FOSS sites.

Figure 2: Schedule of Implementation and Data Collection

	Time Period									
	Fall 08	Spring 09	Fall 09	Spring 10	Fall 10	Spring 11	Fall 11	Spring 12	Fall 12	Spring 13
Teacher PD (T1 and T2) & Curriculum Modifications (T1)										
Holt										
Unit 1	Pilot		Cohort 1		Cohort 2					
Unit 2			Pilot		Cohort 1		Cohort 2			
Unit 3				Pilot		Cohort 1		Cohort 2		
Foss										
Unit 1			Pilot		Cohort 2	Cohort 1				
Unit 2		Pilot		Cohort 1		Cohort 2				
Unit 3					Cohort 1		Cohort 2			
Teacher Content Tests & SEC										
Holt										
Unit 1			Cohort 1		Cohort 2					
Unit 2					Cohort 1		Cohort 2			
Unit 3						Cohort 1		Cohort 2		
Foss										
Unit 1					Cohort 2	Cohort 1				
Unit 2				Cohort 1		Cohort 2				
Unit 3					Cohort 1		Cohort 2			
Student Unit Tests										
Holt										
Unit 1	Pilot		Cohort 1		Cohort 2					
Unit 2					Cohort 1	Cohort 2	Cohort 2			
Unit 3						Cohort 1		Cohort 2		
Foss										
Unit 1			Pilot		Cohort 2	Cohort 1				
Unit 2		Pilot		Cohort 1		Cohort 2				
Unit 3					Cohort 1		Cohort 2			
State Administered Student Tests										
Holt						Cohort 1		Cohort 2		
Foss								Cohort 1		Cohort 2

Conclusion

The 21st Century Science Center project represents an ambitious attempt at modifying science curricula based on cognitive science principles and evaluating the effects of those modifications on student outcomes using experimental design evaluations. In order to maximize the usefulness of the findings, the study is large, longitudinal, and implemented for two quite different science curricula that are typical of curricula commonly found in our public schools.

References

Bloom, H. S. (2006). The Core Analytics of Randomized Experiments for Social Research. MDRC Working Papers on Research Methodology. Available online at: http://www.mdrc.org/publications/437/full.pdf.

Bloom et al. (2007). Randomizing Groups to Evaluate Place-Based Programs, in a *Learning More From Social Experiments: Evolving Analytic Approaches (edited by* Howard S. Bloom). New York: Russell Sage Foundation.

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

Boruch, R. F. (2005). *Place randomized trials : Experimental tests of public policy*. Thousand Oaks: Sage Publications.

Bransford, J. D. & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad and P. D. Pearson (Eds.), *Review of Research in Education*, 24, 61-100. Washington, D.C.: American Educational Research Association.

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences, 14*(2), 161-199.

Chinn, C. & Malhotra, B. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Science Education*, 86(2), 175-218.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99*(2), 311-325.

Gentner, D. (2001). Exhuming similarity. *Behavioral & Brain Sciences, 24*(4), 669.

Hedges, L.V. & E.C. Hedberg (2007). Intraclass Correlation Values for Planning Group-Randomized Trials in Education, *Educational Evaluation and Policy Analysis*, 29(1), 60-87.

Hegarty, M., Kriz, S., & Cate, C. (2003). The role of mental animations and external animations in understanding mechanical systems. *Cognition and Instruction*, 21, 325–360.

Kriz, S. & M. Hegarty (November 2007)Top-down and bottom-up influences on learning from animations. *Inernational Journal of Human-Computer Studies*, 65(11), pp. 911-930.

Maynard, Rebecca and Larry Orr (2010). Social Experiments: Evaluating Education Reforms and Interventions Using Experimental Methods. Philadelphia, PA, University of Pennsylvania.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics & Statistics, 86*(1), 156-179.

Mosteller, F., & Boruch, R. F. (2002). *Evidence matters : Randomized trials in education research*. Washington, D.C.: Brookings Institution Press.

Orr, L. L. (1999). *Social experiments : Evaluating public programs with experimental methods*. Thousand Oaks, Calif.: Sage Publications.

Raudenbush, Stephen. W. , Jessaca Spybrook, Xiao-feng Liu, and Richard Congdon, (First Draft: March 2, 2004). Optimal Design for Longitudinal and Multilevel Research: Documentation for the "Optimal Design" Software, accessed 12_20_08 at http://www.wtgrantfoundation.org/newsletter3039/newsletter_show.htm?doc_id=215108

Raudenbush, Stephen W. , Andres Martinez and Jessaca Spybrook (Revised November 11, 2005). Strategies for Improving Precision in Group-Randomized Experiments, Chicago, IL: University of Chicago.

Ross, B.H.& P.T. Kennedy (1990). Generalizing from the use of earlier examples in problem solving. Journal of Experimental Psychology: Learning, Memory.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation : A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Shadish, W. R., Cook, T. D., Campbell, D. T., & Hazel M. Hussong Fund. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Silk, E., Schunn, C. D., & Carey, M. S. (2007). Evaluating A Design-Based Learning Curriculum in Terms of Students' Science Reasoning Gains. Paper presented at the National Association for Research in Science Teaching. New Orleans, LA, (April, 2007).

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis. 29 (1*): 5-29.

VanLehn, K. (1998). Analogy events: How examples are used during problem solving. *Cognitive Science, 22*, 347-388.

Wan-Chi Wong. (2006). Understanding dialectical thinking from a cultural-historical perspective. *Philosophical Psychology, 19*(2), 239-260.

Whitehurst, G. R. (2008) Institute of Education Sciences, U.S. Department of Education. (2008). *Rigor and Relevance Redux: Director's BiennialReport to Congress* (IES 2009-6010). Washington, DC.