Identification of Student- and Teacher-Level Variables in Modeling Variation of Mathematics Achievement Data

James E. Tarr, *University of Missouri*
Daniel J. Ross, *University of Missouri*
Melissa D. McNaught, *University of Iowa*
Óscar Chávez, *University of Missouri*
Douglas A. Grouws, *University of Missouri*
Robert E. Reys, *University of Missouri*
Ruthmae Sears, *University of Missouri*
R. Didem Taylan, *University of Missouri*

Please address all correspondence to:
James E. Tarr
University of Missouri
Department of Learning, Teaching & Curriculum
121E Townsend Hall
Columbia, MO 65211-2400
578-882-4034
tarrj@missouri.edu

Paper presented at the
Annual Meeting of the American Educational Research Association

Denver, April 30–May 4, 2010

Identification of Student- and Teacher-Level Variables in Modeling Variation of Mathematics Achievement Data

James E. Tarr
Daniel J. Ross
University of Missouri

Melissa D. McNaught
University of Iowa

Óscar Chávez
Douglas A. Grouws
Robert E. Reys
Ruthmae Sears
R. Didem Taylan
University of Missouri

BACKGROUND

Perspective

Based on a long series of international comparisons of student mathematics achievement (e.g., TIMSS, PISA) it is clear that US students are not achieving to their potential. The reasons for this are obviously complex, but "the TIMSS curricular reports suggested that at least part of the problem resided in American curricula, which were seen as more skills oriented, more repetitive, and less conceptually deep than those of nations that scored better on TIMSS (Schmidt, McKnight, & Raizen, 1997)" (Schoenfeld, 2006, p. 14).

Disappointing performance of US students in international studies spurred the National Science Foundation to invest in the development of "reform" curricula that embodied tenets of the National Council of Teachers of Mathematics' *Curriculum and Evaluation Standards for School Mathematics* (1989). These NSF-funded curricular materials differed from "traditional" mathematics textbooks by integrating several branches of mathematics, focusing on the development of mathematical thinking and problem solving, and deemphasizing skills and symbol manipulation (see Nathan, Long, & Alibali, 2002). Because these new approaches to curriculum organization have only recently entered the mainstream, "relatively small numbers of students have worked their way through a full reform curriculum… (and) there are scant data regarding the effectiveness of these curricula—either on their own merits or in comparison with traditional curricula" (Schoenfeld, 2006, p. 15).

Paper presented at the Annual Meeting of the American Education Research Association, Denver, May 2010. The authors wish to thank Michael Harwell for his methodological expertise. This paper is based on research conducted as part of the Comparing Options in Secondary Mathematics: Investigating Curriculum (COSMIC) project, a research study supported by the National Science Foundation under grant number REC-0532214. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The NSF-funded curricula are indeed controversial, having both outspoken advocates and detractors. For more than a decade, mathematics education has endured "math wars" in which traditionalists have argued that "standards-based" curricula are "superficial and undermine classical mathematical values; reformers claim that such curricula reflect a deeper, richer view of mathematics than the traditional curriculum" (Schoenfeld, 2006, p. 15) (for a more detailed history of the "math wars," see Schoenfeld, 2004). To placate teachers, administrators, students, and parents, some school districts recently began to offer parallel curriculum paths in which students are presumably "free" to study mathematics using one of two organizational schemes, an *integrated* approach or a (traditional) *subject-specific* approach. It is in the special context of parallel curricular paths that we examine curricular effectiveness.

The COSMIC Project

Funded by the National Science Foundation, *Comparing Options in Secondary Mathematics: Investigating Curriculum* (COSMIC) is a research project that involves a three-year longitudinal comparative study of integrated mathematics curricula and subject-specific mathematics curricula on mathematical learning in schools that offer parallel curricular paths. The primary goal of the COSMIC Project is to evaluate secondary school students' mathematics learning using multiple measures of student achievement while carefully attending to curriculum implementation via classroom observations, opportunity-to-learn (OTL) data, teacher surveys and interviews. Preliminary work for the COSMIC project began in 2005 with data collection starting in the Fall of 2006 and continuing through the 2008-2009 school year.

RESEARCH QUESTIONS

Given the large federal investment in NSF-funded curricular materials, their infusion into US mathematics classrooms, and the corresponding response by teachers, administrators, students and parents, the COSMIC Project sought to answer the following research questions:

1. Are there differential effects on high school students' mathematics learning when they study from an integrated approach textbook and when students study from a subject-specific textbook? In particular, are there differential curricular effects with respect to student performance on assessments of:
    - Common objectives;
    - Mathematical reasoning;
    - Mathematics concepts and problem solving.
2. What are the relationships among curriculum type, fidelity of implementation, and student learning? In particular,
    - What curriculum implementation factors are associated with high school students' mathematics learning?
    - What teacher characteristics are associated with high school students' mathematics learning?

OBJECTIVES OF THIS MANUSCRIPT

In this paper we address key issues in the design of longitudinal studies of curricular effectiveness with particular emphasis on data collection, reduction, and coherence in

modeling student achievement in year 1 of the COSMIC Project. Reduction and coherence are important in the context of large data sets such as ours that include extensive information about teachers as well as detailed information about their teaching practice. In particular, there is a need to strike a balance between the collection of massive amounts of student and teacher data to help explain and understand the student learning that takes place and moving forward with developing parsimonious models of the important variables that relate to student learning. In this paper, we describe our approach to balancing these competing demands and share insights related to year 1 of the COSMIC Project.

SIGNIFICANCE OF THE STUDY

Developing a clear understanding of the dynamics of parallel curriculum use and a comprehension of the factors associated with how and what students learn under different curricular approaches is imperative for many reasons. For example, understanding a parallel-use context is essential for future curriculum change because parallel programs can likely be an intermediate step in curricular change in many schools. Furthermore, if and when, a scaling up of an integrated content approach occurs in US schools then the findings from this study will provide valuable information for the decision making that will need to take place as part of such a movement. The improved understanding that this study provides (albeit in a special context) concerning the relationships among curricular organization, curriculum implementation factors, and gains in student learning will be useful to the field in theory building, curriculum writing, professional development, and decision making by school administrators.

THEORETICAL PERSPECTIVES:
On Evaluating Curricular Effectiveness

In designing the COSMIC project research we took account of the comprehensive framework for evaluating curriculum effectiveness developed by the National Research Council (2004) (see Figure 1). As the figure shows, the first step in developing a research design is to attend to Program Theory, which essentially means determining program components, identifying implementation strategies including processes and contextual influences, and deciding on student outcomes to be taken into account. In the COSMIC project, the program components of mathematical content and curriculum design elements were characterized using a comprehensive content analysis of each of the two curriculum types studied. Implementation components, in particular implementation resources and processes, were ascertained by curriculum type in two ways: examination of teacher's editions of textbooks and textbook author interviews. Our careful attention to teachers' implementation of curricular materials was necessary in order to draw causal inferences between curriculum and student learning; the National Research Council advocates that studies of curricular effectiveness account for *treatment integrity*, or what we refer to as *fidelity of implementation.* Student outcomes were carefully considered for inclusion in this study, including multiple assessments, enrollment patterns, attendance, and attrition. Because there are multiple important student outcomes but limits on how many assessments can be administered in one study, we decided to focus on the most important outcome in our view, namely student learning. We annually measured student learning in

three distinct ways in the study by using what we call a *fair* test, a mathematical reasoning test, and a standardized achievement test.

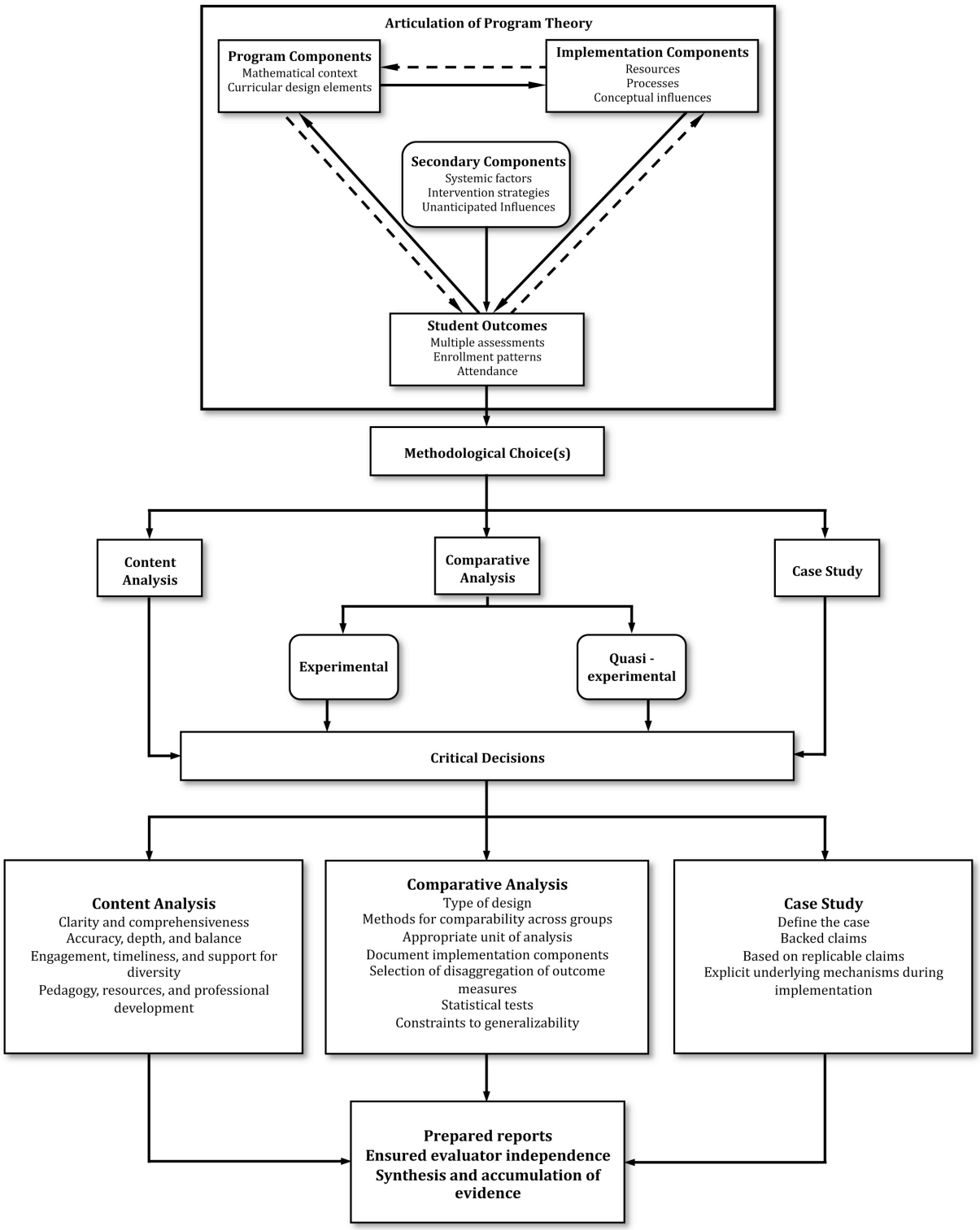

Figure 1. Framework for evaluating curricular effectiveness (National Research Council, 2004).

SAMPLE

Curriculum Types

The COSMIC project studied two curriculum types where the organization of the mathematics content differed, namely *subject-specific* organization and *integrated* organization. Commercially-developed, traditional mathematics textbooks exemplify subject-specific curricula and these widely-used textbooks focus on a particular strand of mathematical content each year, such as algebra or geometry. Textbook series of Holt, Prentice Hall, Glencoe, McDougal Littell, and HRW constitute the sample of subject-specific curricula; subject-specific courses include common titles such as Algebra 1, Geometry, Algebra 2, and Precalculus. By way of contrast, in the integrated curriculum organization multiple strands of mathematical content (geometry, algebra, discrete mathematics and statistics) are coalesced. During the 1990s, the National Science Foundation invested heavily in the integrated approach to mathematics curricula. Among the curriculum development projects at the high school level, *Core-Plus* emerged as the most popular and maintains the greatest market share. In this study, the *Core-Plus* textbook series was selected as the representative for the integrated curricula; students studying from the integrated curriculum took Course 1, Course 2, and Course 3.

A primary difference between the subject-specific curricula and the integrated curricula is how lessons are structured. Each subject-specific lesson usually has a *Lesson Preview, Teach* (containing numerous worked examples), *Practice and Apply*, and closure component. Teachers enacting a subject-specific curriculum generally facilitate student learning using teacher-led, whole-class discussions. The integrated curriculum is structured such that, following a relatively brief *Launch*, students work in small-group setting to *Explore* mathematical ideas while the teacher serves as facilitator; subsequently students participate in a *Share and Summarize* component in which they share their thinking in a whole-class discussion, and discuss the important mathematical ideas of the lesson. In the integrated curricula, a lesson *Launch* occurs at the beginning of a unit, so there may be multiple days in which students engage in *Explore* and *Share and Summarize* without a *Launch* component of the lesson. Notwithstanding the preceding, for the integrated curricula, the textbook publisher recommends that closure be included in all lessons.

Schools

Selection method. To identify an appropriate sample the COSMIC project did an extensive search for high schools throughout the US that offered parallel curriculum paths to their secondary students, and students were free to choose between either path. In particular, we searched for schools that offered both integrated mathematics and a subject-specific (Algebra 1, Geometry, Algebra 2, Pre-calculus) curriculum organization. This requirement narrowed the field of possible high schools significantly, but it was a crucial requirement for the design of this research study. Satisfying this requirement helped ensure that there would be a balance between curriculum types with regard to the number of days of instruction, and controlled for many other contextual factors such as homework and technology policies, organization and length of class periods, professional development provided during the study, SES make-up of the student body, and so forth. Moreover, we stipulated that schools were eligible for participation only if students were not *tracked*;

that is, schools were ineligible if policies channeled high-performing students into one curriculum type while directing lower-performing students into another curriculum.

We also selected schools that would provide diversity in our sample with regard to geographic region, race/ethnicity, and social economic levels. Furthermore, we selected only schools that were not using either of their mathematics textbook series for the first time. This requirement ensured that most teachers were familiar and had experience using the textbooks in our sample.

Once schools that met our criteria were identified, we visited the schools to talk with school representatives. This always included mathematics teachers, mathematics chair/coordinator (where they existed) and the school principal. In the majority of cases it also required meeting with the district superintendent, and in all of these meetings the researchers described the nature of the research and what commitments were required by the district (e.g., providing prior achievement test data, allowing researchers to observe classes, and committing three days for assessments). We also discussed the benefits the district would receive from this research effort, including results from the additional assessments, modest honoraria for teachers, and the findings from the research. The latter point was particularly persuasive, as all of the schools were interested in research data regarding the impact of these two curricular paths on the performance of students. The process described above resulted in choosing 11 schools in six school districts that were located in five geographically dispersed US states.

Demographic data. Consistent with the selection criteria, there was diversity in the student sample for year 1. As depicted in Table 1, data were collected from 2,621 students, with slightly more females comprising the sample than males. While the majority of students were White (77.56%), the sample represented a relatively diverse ethnic population with the proportion of White students ranging substantially from 50.45% in District R to 94.02% in District C. A larger percentage of Black students were reported in District W (20.44%) than in other districts; while District R reported nearly 40% of its student population as Hispanic. Other races—including Asian/Pacific Islander, Native American/Alaskan Native, Mixed Race, and Unclassified—comprised 4.36% of the sample but accounted for nearly 7% in District R.

With regard to characteristics that qualify students for school services and/or resources, there was similar diversity across districts. For example, the portion of students with Individual Educational Plans (IEP) ranged from 1.66% in District W to 10.12% in District B. The percentage of students classified as Limited English Proficiency (LEP) was as high as 4.89 in District C while District I reported none. The use of Free/Reduced Lunch (FRL) is commonly used in educational research despite its limitations (Harwell & LeBeau, 2010) as a measure of SES. In this study, there was a wide range in the percent of students qualifying for FRL, from 19.09% in District R to 53.27% of District I.

Table 1

Demographic Data for Each School District in the COSMIC Project, Year 1

		Gender		Race/Ethnicity				Qualifications		
District	**Students**	**Male**	**Female**	**Black**	**Hispanic**	**White**	**Other[1]**	**IEP**	**LEP**	**FRL**
B	257	46.70	53.30	1.56	6.23	90.27	1.94	10.12	3.11	28.79
C	184	50.00	50.00	0.00	4.89	94.02	1.09	9.78	4.89	46.20
I	336	47.62	52.38	10.12	8.04	77.98	3.86	7.74	0.00	53.27
R	440	47.05	52.73	2.95	39.77	50.45	6.83	4.32	3.86	19.09
T	802	50.00	50.00	0.62	2.99	92.77	3.62	8.60	1.12	25.31
W	602	46.35	53.65	20.44	7.31	66.44	5.81	1.66	1.66	25.75
Totals	**2,621**	**48.04**	**51.93**	**6.83**	**11.25**	**77.56**	**4.36**	**6.41**	**2.02**	**29.76**

As depicted in Table 2, the 2,621 students comprising the year 1 sample were largely evenly distributed across the two curriculum types, with 48% enrolled in integrated and 52% enrolled in subject-specific curriculum. Although the number of students enrolled in each curriculum type was similar overall, far more students enrolled in the integrated path in District I while the opposite was true for District R. Correspondingly, there were slightly more teachers of the subject-specific curriculum than taught the integrated curriculum. Of the 43 teachers who participated in the COSMIC project, 20 taught the integrated curriculum while the remaining 23 taught the subject-specific curriculum; this includes a few teachers who taught both curriculum types.

Table 2

Number of Teacher and Student Participants in Year 1, by School District and Curriculum Type

	Teachers		Students	
District	**Integrated**	**Subject-Specific**	**Integrated**	**Subject-Specific**
B	3	2	127	130
C	1	2	97	87
I	4	2	286	50
R	1	5	47	393
T	6	4	462	340
W	5	8	237	365
Total	**20**	**23**	**1,256**	**1,365**

Student participation necessitated the writing of three end-of-year exams, administered during the final six weeks of year 1. Although discussion of the outcome measures is offered subsequently, it is worth noting that 2,615 of 2,621 students took *at least one test* in year 1 of the COSMIC Project, an astonishing participation rate of 99.77%.

LITERATURE REVIEW

The notion that NSF-funded curricula are more effective than traditional curricula in yielding student mathematical learning is highly controversial (Senk & Thompson, 2003).

[1] Includes Asian/Pacific Islander, American Indian/Alaskan Native, Mixed Race, and Unavailable.

In a comprehensive review of research on curricular effectiveness, the National Research Council (2004) identified numerous methodological limitations of studies on the impact of mathematics curriculum on student learning. Among these, the NRC noted that few studies utilized an experimental design, included multiple measures of student learning outcomes, or were sensitive to *treatment integrity* (or fidelity of implementation). Moreover, they acknowledge a dearth of longitudinal studies of curricular effectiveness. Although the NRC advocates additional studies, Cai and Moyer (2006) argue that selecting the optimal way to conduct research on the effects of different curricula on student learning is as equally controversial.

Most studies measuring the effectiveness of NSF-funded curricula have been conducted in the context of field-tests (Senk & Thompson, 2003). Considering the potential bias that field tests conducted by the curriculum developers might carry into the studies, Harwell et al. (2007) and Post et al. (2008) based their curriculum research on district-wide curricula adoptions, where teachers were required to teach the adopted curriculum, as opposed to the case of field-test versions of the curricular material. Harwell et al. (2007) examined mathematics achievement of secondary students while comparing different types curricula; Post et al. (2008) studied achievement models of middle school students who were enrolled in a *Standards*-based curriculum. Both studies utilized hierarchical linear modeling (HLM) to differentiate the effects of student- and classroom-level variables that offer predictive power in modeling student achievement. Descriptive data suggest that low socioeconomic status (SES), African American, nonnative English speakers and special education students were consistently outperformed by their peers.

In both Harwell et al. (2007) and Post et al. (2008), student-level variables included prior mathematics achievement, SES (i.e., students qualifying for free or reduced-price school lunch [FRL]), gender, and attendance. Classroom-level variables included class SES level, percent ethnic minority (Black, Asian, and Hispanic), English Language Learners, special education students, and female students as well as attendance and school district affiliation. Harwell et al. (2007) added curriculum type as a classroom-level predictor. HLM analyses results revealed that SES level and prior mathematics achievement consistently and strongly predicted mathematics performance at both student- and classroom-levels whereas gender and attendance were not found to explain significant variability in students' mathematics performance. Post et al. (2008) found that suburban classrooms outscored urban classrooms in mathematics achievement, indicating how school location may impact student performance. A key finding by Harwell et al. (2007) was that when all variables were taken into account, there was no statistically significant difference among different types of curricula. In both of these studies, the teacher-level variable "professional development hours" was not a significant predictor of student achievement. However, neither of the previous studies was able to carefully assess the extent of curriculum implementation in the classroom.

Schoen et al. (2003) and McCaffrey et al. (2001) investigated effects of teacher variables on student achievement. The former was a field test study that examined teachers' preparation, practices and concerns related to students' mathematics achievement in the implementation of the *Core-Plus* curriculum. This study measured the *teacher achievement index*, which was defined as the mean of each teacher's students' *adjusted mean posttest score* (posttest scores after removing the variance due to the pretest). Although pretest results revealed that the percentage of free or reduced-price

lunch (FRL) and the sum of percentages of African American, Native American and Hispanic students were strongly and negatively correlated with the pretest achievement, the adjusted mean posttest scores were not statistically significantly correlated with any of these variables. Using regression analysis, the authors proposed a model of student achievement. In contrast to the findings of Harwell et al. (2007), Schoen et al. determined that teachers' completion of a developer-sponsored summer workshop was the most significant variable predicting their students' achievement. Moreover, they identified several other variables that were positively and significantly associated with adjusted student achievement, including (a) cooperation with other teachers and having confidence in teaching, (b) using group work instead of teacher presentation and whole group discussion, (c) spending less time on non-academic matters, (d) using a variety of assessment methods, (e) not replacing the curriculum materials with the ones that are less open-ended and more skill-oriented, (f) high expectations on homework and grading, and (g) high observer rating based on the criteria for effective reform teaching.

Investigating the relationship between teachers' use of reform-based instructional practices and student achievement, McCaffrey et al. (2001) based their study on self-reported data collected from NSF-funded and traditional curriculum teachers while also taking into account of different student variables. Results revealed that teachers' reported use of reform teaching practices were positively correlated to the achievement of the students in the integrated mathematics classes while no significant correlation was observed for the achievement of students in the traditional mathematics classes. In general, students whose teachers had a graduate degree in mathematics or mathematics education tended to score higher in achievement tests. Nevertheless, teacher's level of training was excluded from the model due to its possible interaction with teaching practices in classroom. Teacher background variables defined as their degree, certification status, coursework in mathematics, gender, ethnicity and years of teaching experience were not found to have significant predictive power on reported reform-practices.

DESIGN AND DATA SOURCES

Hierarchical Linear Modeling

Because students experience the school mathematics curriculum in groups, not as individuals, it is not appropriate to use *student* as the unit of analysis in curriculum evaluation studies (National Research Council, 2004; Osborne, 2000). The recognition of this fact warrants the use of group means (e.g., class averages, scores aggregated by teacher) or multi-level modeling, in which students are nested in hierarchical structures. For example, students experience mathematics as a *class*; several sections of the same class are taught by the same *teacher*; several teachers are nested within the same *school*; and (public) schools are held accountable to the same state curriculum framework. In principle, one could argue that students represent the first of many levels in a nested, hierarchical educational system. However, modeling student achievement across many levels is extraordinarily complex, necessitates a sufficient number of cases in each level, and interpretation of results is particularly challenging.

Although students experience curriculum as a *class*, we argue that several classes taught by the same teacher are *not independent* because it is likely many aspects of instruction do not vary within the school day. For example, throughout a given school day, a high school Algebra 1 teacher is likely to cover the same mathematics content in a single

lesson for each class period he teaches Algebra 1 that day. Moreover, the same Algebra 1 teacher is likely to emphasize the same mathematics (e.g., procedural fluency), spend approximately the same amount of time on particular lesson components, and assign the same homework from the Algebra 1 textbook. Because the independence of cases is fundamental in hypothesis testing, we use *teachers* (not classes) as the unit of analysis. Our design is a two-level model, students nested within teachers, and we seek to identify student-level and teacher-level variables for inclusion in models of student achievement.

Independent Variables

Student-level. Recent studies of curricular effectiveness provide insight into what data are essential to collect. At the student-level, it seems requisite to collect data regarding prior achievement, gender, race/ethnicity, and other designations such as Individual Education Plan (IEP) and Limited English Proficiency (LEP). A complete list of the student variables to which we attended appears in Table 3. In studies of curricular effectiveness, it is imperative to include measures of student prior achievement in order to (a) establish the equivalence of treatment groups, or (b) control for non-equivalence of treatment groups. In a subsequent section, we provide detailed narrative on how we generated a common prior achievement score across districts in our sample, and data that show that mean student prior achievement scores across our two curriculum types are not significantly different.

Table 3
Student-level Control Variables

Control Variable	Data Type	Data Source
COSMIC Prior Mathematics Achievement	Interval	Transformation of scores on state-mandated tests
Gender	Dichotomous	Student Records
Race/Ethnicity	Polytomous	Student Records
Individual Education Plan	Dichotomous	Student Records
Limited English Proficiency	Dichotomous	Student Records

Teacher-level. At the teacher-level, it is essential to collect information regarding characteristics such as experience, hours of professional development, knowledge and beliefs. Moreover, in response to the NRC (2004) stipulation that *treatment integrity* be documented in studies of curricular effectiveness, it is clearly necessary to collect data on teachers' implementation of curricular materials, including how the curriculum was enacted and what opportunity-to-learn mathematics students were afforded. As reported in another paper (see McNaught et al. 2010), we examined the *fidelity of implementation* of curricular materials through two lenses: content fidelity and presentation fidelity. We used multiple data sources to gauge teachers' implementation of curricular materials including Table of Contents Records, Textbook-Use Diaries, Initial Teacher Survey, and observations using a Classroom Visit Protocol.

Collectively, nearly 30 variables were measured and these are listed in Table 4. Because of the large number of teacher variables, it was hypothesized that several attributes might be highly correlated, and hence essentially measuring the same construct. For example, consider the variables Seating and Collaboration, from our Classroom Visit Protocols. The extent to which students worked collaboratively during observed lessons

(Collaboration) is likely related to how likely students were seated in groups (Seating); students tend not to work collaboratively if desks are arranged in rows. It follows that subsequent analyses were employed to reduce the number of teacher variables to a more manageable number. Without doing so, we run the risk of introducing bias and explaining far more variance that can be attributable to student- and teacher-level variables.

Table 4

Teacher-level Variables, by Data Source

TABLE OF CONTENTS RECORDS	
OTL Index	Opportunity to Learn Index represents the percentage of textbook lessons taught
ETI Index	Extent of Textbook Implementation index represents the extent to which teachers followed their textbook using weighted averages
TCT Index	Textbook Content Taught index represents the extent to which teachers, *when teaching textbook content*, followed their textbook, supplemented their textbook lessons, or used altogether alternative curricular materials
CLASSROOM VISIT PROTOCOLS	
Pres Fidelity	Global rating of presentation fidelity of textbook in observed lessons
Content Fidelity	Global rating of content fidelity of textbook in observed lessons
Tech_Teacher	Likelihood that teacher utilized graphing calculators in instruction
Tech_Students	Likelihood that most students utilized graphing calculators during instruction
Reasoning	Classroom Learning Environment: Reasoning about Mathematics
Students' Thinking	Classroom Learning Environment: Students' Thinking in Instruction
Sense-Making	Classroom Learning Environment: Sense-Making about Mathematics
Closure	Relative frequency that teacher brought closure to the observed lessons.
Engage	Extent to which most students were engaged (on-task) during observed lesson
Seating	Relative frequency that students were seated in groups during observed lessons.
Collaboration	Relative frequency that students worked collaboratively during observed lessons.
Time_LD	Percent of class period devoted to lesson development
Time_NI	Percent of class period devoted to non-instructional time
Time_PA	Percent of class period devoted to practice and apply (homework)
INITIAL TEACHER SURVEY	
Teach_Exp	Number of years teaching
Math_Exp	Number of years teaching mathematics
Belief 1	Teacher beliefs about reform-oriented practices
Belief 2	Teacher beliefs about didactic approaches
Belief 3	Teacher beliefs about students' self-efficacy
PD_12	Number of hours of professional development in the last 12 month
PD_3	Number of hours of professional development in the last 3 years
Familiar	Familiarity with *Principles and Standards for School Mathematics* (NCTM 2000)
Agreement	Agreement with *Principles and Standards for School Mathematics* (NCTM 2000)
Implement	Implementation of *Principles and Standards for School Mathematics* (NCTM 2000)
Text	Number of years teaching from the district-adopted textbook
Preparation	Preparedness to teach the district-adopted textbook
Rating	Rating of satisfaction with the district-adopted textbook
TEXTBOOK-USE DIARIES	
Days	Number of days spent on target content

It is worth further noting that we use FRL as a proxy for SES despite its limitations (see Harwell & LeBeau, 2010). For at least two reasons, we aggregated the percentage of FRL students for each teacher rather than use FRL as a student-level variable. First, FRL as a student-level variable often yields extraordinary—arguably unwieldy—slopes that seem implausible. Second, District W was unwilling to provide FRL status for individual students (but was willing to provide it at the class-level), and this decision introduced methodological challenges. Accordingly, FRL is treated as a teacher-level covariate.

Dependent Variables

Project-developed tests. Following the recommendations of the NRC (2004), we developed assessment instruments using items written around topics common to both curriculum programs with the deliberate goal of not being biased towards either of the two curriculum programs. For the three years of the study, we developed five project tests. Each test was developed following a cycle of curriculum analyses and several rounds of external and internal reviews, pilots, and revisions. For each of the first two years of the study, two tests were developed: the first test, Test A, was comprised of items that focused on common topics (the *fair* test) across the two curriculum types; the second test, Test B, assessed students' mathematical reasoning and problem solving. The items in the reasoning tests were based on topics that were appropriate to the grade level, according to the content in the textbooks, and as identified during our internal and external reviews. For a detailed discussion of the test development process, see Chavez et al. (2010).

The majority of the items in the *fair* test (Test A) used in year 1 deal with linear relationships, a topic holding a central position in both Algebra 1 and integrated textbooks. The mathematical reasoning test (Test B) included problems on data analysis, algebra, and geometry. Although analyses are not reported here, in year 2, the fair test included some items on algebraic topics, although it was focused primarily on geometric topics and concepts common to the two curriculum types (e.g., coordinate geometry, perimeter and area, and trigonometry). The mathematical reasoning test for year 2 included geometric items and an algebraic item. In year 3, we developed only one test that focused on functions as the central mathematical idea. The items in these tests were constructed response with but one or two exceptions.

The scoring rubrics were refined in an iterative manner, following a process parallel to the development of the tests. We examined the reliability of our scoring process and the results were excellent, with an inter-rater reliability above 94% for all five tests. After the tests were administered to 2,621 students in year 1, analyses of scores revealed that the rubrics we developed were applied in a highly reliable manner.

Standardized test. The standardized measure of achievement we selected was the Iowa Test of Educational Development [ITED]: Mathematics: Concepts and Problem Solving. It received high ratings in the Buros Mental Measurements Yearbook (Schafer, 2005) with regard to reliability and validity and it has been described as "among the best general-purpose assessments of high school students' educational development available" (p. 10). Furthermore, it is nationally-normed, which makes it particularly useful in a comparative study. Naturally, there are important differences between the ITED and our project-developed tests. The ITED is a multiple-choice test. Large-scale assessments that rely on multiple-choice items permit only indirect inferences about students' thinking

(Silver, Alacaci, & Stylianou, 2000). The items in our project-developed tests are constructed-response items. Our scoring rubrics were designed to score separately each item's answer and the work done to get that answer. In this way we have collected ample direct evidence regarding students' performance on more complex problem-solving tasks.

DATA ANALYSIS

Student Data

COSMIC Prior Achievement (CPA). Consistent with our theoretical framework (NRC, 2004) our quasi-experimental design necessitated that comparability be established by matching samples or making statistical adjustments using, among other factors, prior achievement measures. Because a pre-test administered to all students is rarely feasible in large-scale studies of curricular effectivness across multiple states such as ours, we opted for a reasonable alternative, namely the utilization of scores on state-mandated grade 8 tests, typically administered during the 2004-05 school year. These high-stakes tests generally purport to measure student achievement in mathematics at a common point in time (grade 8), and so they provided useful information in characterizing student knoweldge prior to curricular treatments in the COSMIC Project. Nevertheless, state tests are usually not nationally-normed, and are scored using different scales. Consequently, it was necessary to put the scores on a common scale that would take into account differences *across* states, as average National Assessment of Educational Progress (NAEP) scores vary considerably across states. In particular, because participating school districts were located in five US states, it was important to acknowledge and subsequently adjust for differences in student achievement across each state. For example, given that grade 8 mathematics students in State X scored above the US average on NAEP while students in State B scored below the US average, we mapped each student's grade 8 state test score in mathematics onto the NAEP scale score for grade 8 mathematics.

In grade 8, some students in District B were assessed using a nationally-normed mathematics achievement test. In these cases, we simply converted their scores to a national *z*-score, which we then mapped onto an NAEP scale score. Therefore, a grade 8 student in State X scoring at the mean ($z = 0$) was assigned to the mean NAEP scale score for State X. A student scoring 1 standard deviation above the mean was assigned a NAEP scale score that corresponded to the mean NAEP scale score plus 1 standard deviation.

For the vast majority of students in COSMIC, grade 8 scores on state-mandated tests were not nationally-normed. In these cases, we converted students' scores in each state to *z*-scores before mapping these scores onto a NAEP scale score (see National Center for Educational Statistics, 2007). We called the resulting score *COSMIC Prior Achievement Score* (CPA Score). The diagram in Figure 2 illustrates the process.

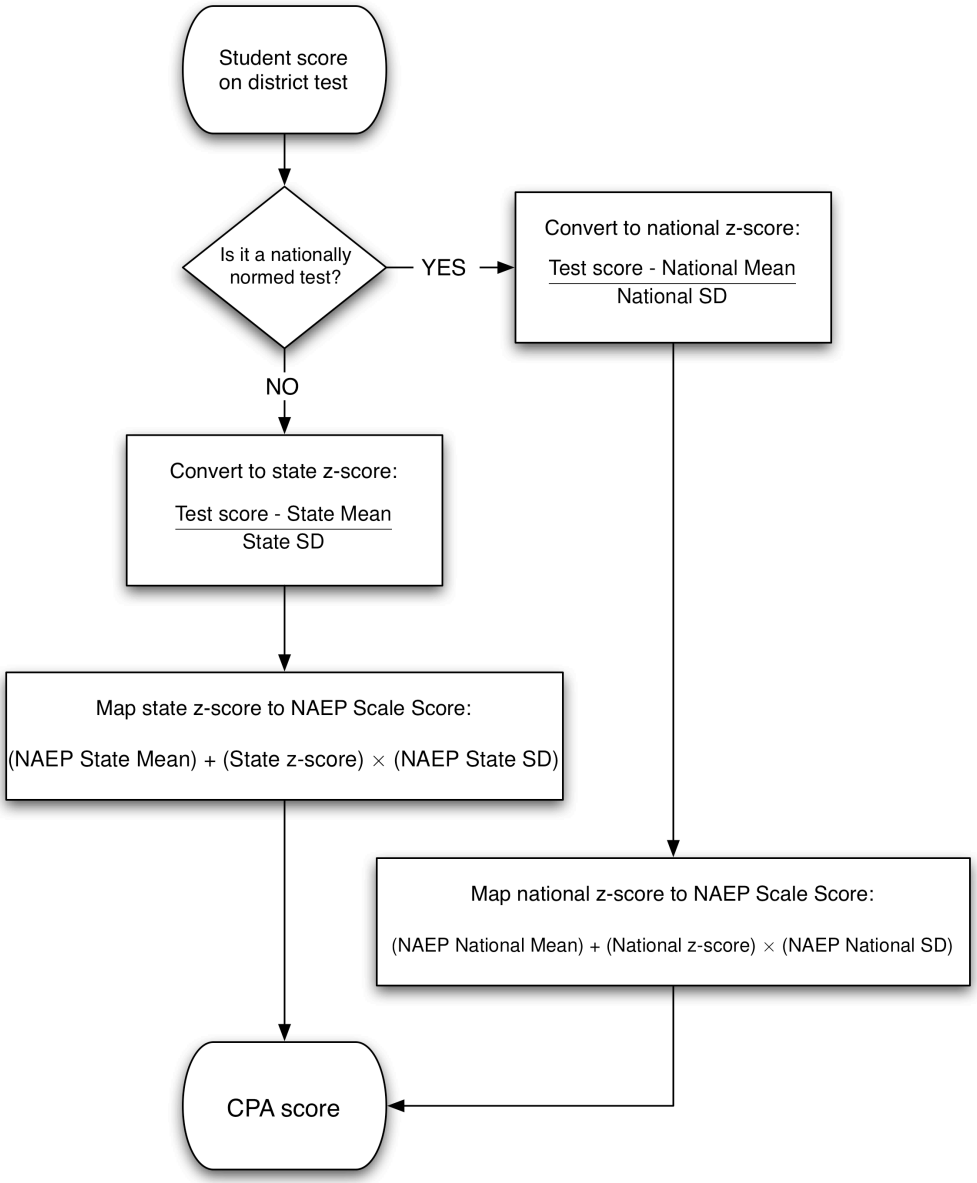

Figure 2. Algorithm for generation of the COSMIC Prior Achievement (CPA) score.

Consider a second illustrative example in which Student A has a scale score of 709 on the 2005 grade 8 test mandated in State Y . Because the assessment for State Y is *not* a nationally normed test, we converted this student's scale score to a state *z*-score using descriptive statistics for the 2005 State Y test: mean score of 682 and a standard deviation of 35. As depicted in Figure 3, this state *z*-score was then mapped onto the NAEP Scale Score: State Y had an average NAEP scale score of 263 and a standard deviation of 34, yielding a CPA score of 289. Thus, although Student A scores 0.77 standard deviations above the mean relative to grade 8 students in State Y, Student A scored approximately

0.28 standard deviations above the mean relative to grade 8 students in the US (see Figure 4).

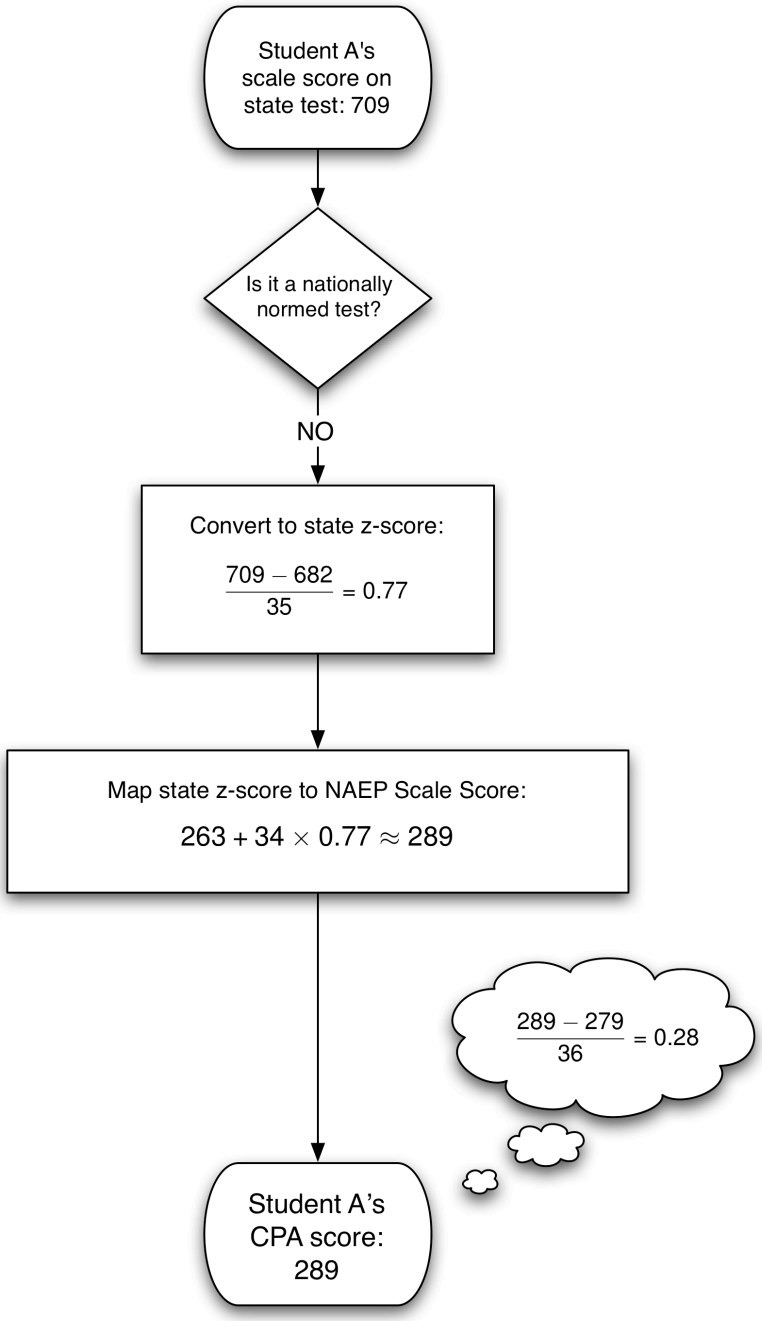

Figure 3. Generating Student A's COSMIC Prior Achievement (CPA) score.

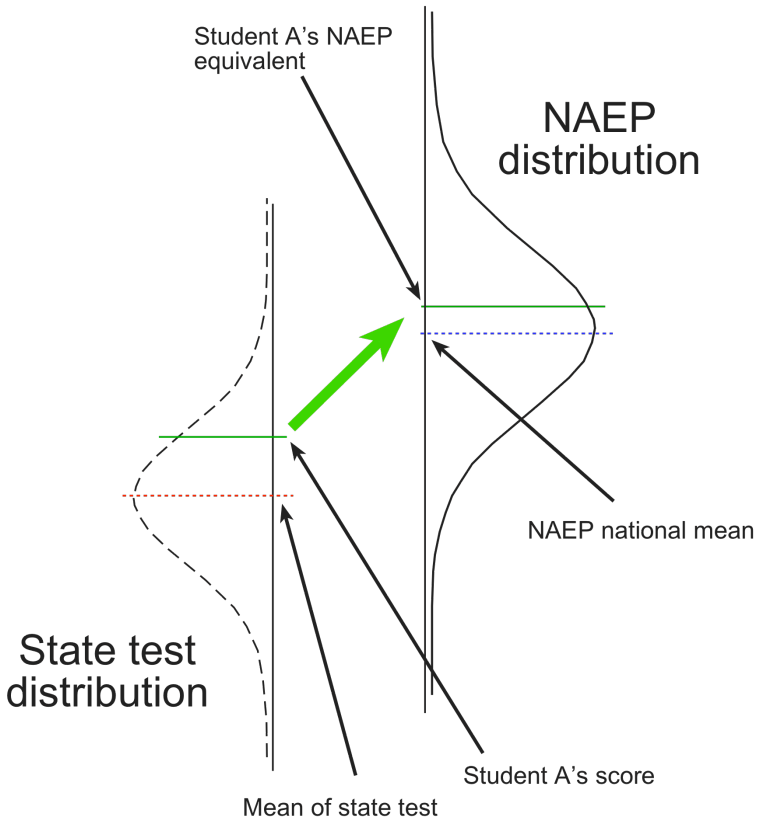

Figure 4. Mapping Student A's state score onto the grade 8 NAEP scale.

Equivalence of treatment groups: Students. The transformation of student prior achievement scores on state-mandated tests to COSMIC Prior Achievement (CPA) scale scores yielded a relatively normal distribution across the year 1 sample (see figure 5). A preliminary analysis revealed there was no significant difference in mean CPA scores across curriculum types. Stated differently, while there was substantial variation in prior achievement within the year 1 sample, there was a comparable distribution in student achievement across curriculum types.

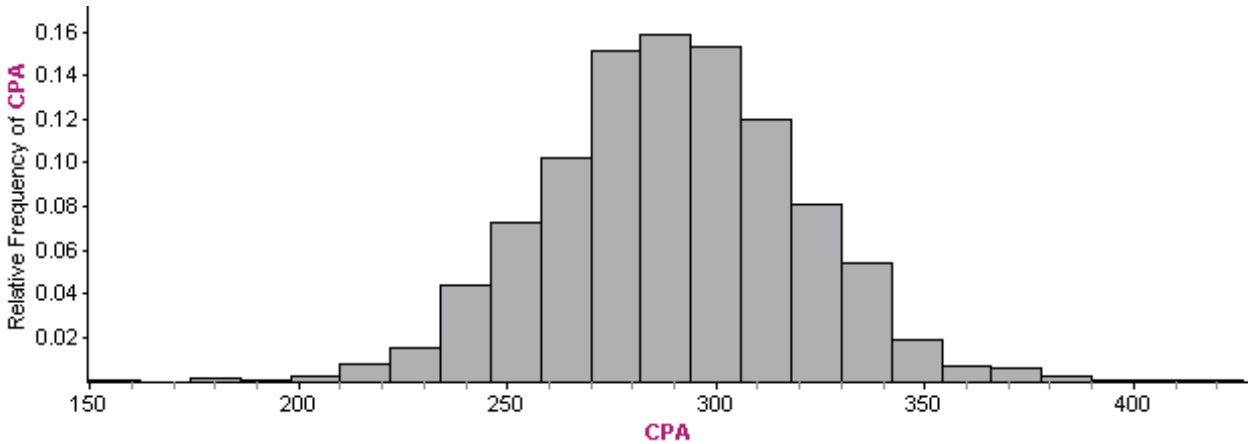

Figure 5. Distribution of COSMIC Prior Achievement scores, year 1.

Scaling student outcomes on constructed-response tests. For each student, scale scores were generated from raw scores on each of the project-developed exams, Test A (common objectives) and Test B (mathematical reasoning). By using Item Response Theory scale scores, additional information related to item difficulty and item discrimination can be incorporated in the determination of student scores. More specifically, two-parameter IRT was employed, using *item difficulty* and *item discrimination* indices to generate theta scores for *each* student on *each* test. The large number of students, sufficient number of test items, and sound psychometric properties of the assessment allowed the models to converge, and checks were done to ensure proper model fitting.

Teacher Data

Equivalence of treatment groups: Teachers. All teachers who participated in the COSMIC study completed an Initial Teacher Survey (ITS) in which they reported numerous background characteristics, including: number of years of experience teaching mathematics; beliefs about teaching and learning mathematics; familiarity and agreement with NCTM *Standards*; implementation of NCTM *Standards*; the amount of time allotted for and the focus of professional development; the impact professional development had on instructional practices; and the use of technological tools in mathematics instruction (for a more comprehensive set of attributes on the ITS, see Table 4).

A comparison of group means, using a one-way analysis of variance (ANOVA), of responses from teachers of subject-specific curricula and of integrated curricula revealed that the differences between the two groups were not significantly different for most of the variables assessed. For example, the two groups of teachers were remarkably comparable in terms of teaching experience, teaching licensure, and mathematics background. Nevertheless, a few differences were discerned. For example, teachers of integrated curricula were significantly more inclined to report familiarity with the NCTM *Standards* ($F = 13.126$, $p = .000$), agreement with the *Standards* ($F = 9.394$, $p = .003$), and (self-reported) implementation of the *Standards* ($F = 5.161$, $p = .025$) than teachers of subject-specific curricula. Furthermore, teachers of the integrated curriculum reported approximately one more year of experience using the integrated textbook than teachers of using subject-specific curricula, a significant difference ($F = 7.107$, $p = .009$). No other differences were detected across the two groups of teachers.

Teacher beliefs. Teachers responded to 32 five-point Likert scale items on the ITS to report their beliefs about the teaching and learning of mathematics. Factor analysis of teachers' belief responses extracted three factors of teachers' beliefs: (1) *reform practices*[2] (Ellis, Malloy, Meece, & Sylvester, 2007), (2) *didactic approaches*[3] (Nie & Lau, in press), and (3) *self-efficacy*[4] (Hoffman, in press). A scale score for each factor was generated for each

[2] "For the purposes of this study, 'reform-oriented' refers to a teacher's use of instructional practices aligned with the curriculum and teaching Standards of the National Council of Teachers of Mathematics (NCTM, 1989, 1991) and *Principles and Standards for School Mathematics* (NCTM, 2000), collectively referred to as the NCTM Standards" (Ellis, Malloy, Meece, & Sylvester, 2007, p. 2).

[3] "Didactic instruction focuses on the transmission of knowledge as represented in curriculum and textbooks and student learning focuses on the passive receipt of knowledge reinforced through drill and practices" (Nie & Lau, in press, p. 2).

[4] "Self-efficacy, the belief in one's ability to execute courses of action to achieve desired results (Bandura, 1986), is related to superior performance and may moderate the influence of anxiety on mathematics (Hackett,1985; Jain & Dowson, 2009;

teacher. Although the two sets of teachers were remarkably comparable on most teacher characteristics and self-reported practices, our analysis revealed that teachers' beliefs were markedly different. With regard to Belief 1 *reform practices,* the mean scale score for teachers of integrated curricula was significantly higher (F = 11.648, p = .001) than the mean for teachers of the subject-specific curriculum. In contrast, for Belief 2 *didactic approaches,* means for teachers of subject-specific curricula were significantly higher (F = 9.009, p = .003) than for integrated. These differences suggest that teachers of integrated curricula held beliefs more consistent with a reform approach while teachers of subject-specific curricula held beliefs that embodied a more didactical instruction orientation. No differences were found with regard to Belief 3, teachers' *self-efficacy* toward the discipline of mathematics.

Principal Components Analysis (PCA). Given the large number of teacher variables from a number of sources (including Initial Teacher Surveys, Table of Content Records, Midcourse Teacher Surveys, and Classroom Visit Protocols), Principal Components Analysis (PCA) was used to reduce the large number of teacher variables (Table 4) from these sources. The initial extraction was conducted with 27 variables and 151 cases; data from the TUD were excluded and are not reported. Each case represented one teacher of a given curriculum type in a given year. Teachers who taught multiple years or both curriculum types had multiple cases. Although classroom visits were not conducted for teachers in the third year, their data from all other sources were used in the PCA.

Throughout the process, parallel analysis was used to determine the number of factors that should be extracted. This technique simulates a factor analysis with the same number of cases and variables on a random set of data. The eigenvalues for each factor indicate the amount of variance that the factor explains, and only factors that explain more variance than the factor based on random data should be kept. Each parallel analysis suggested that seven factors should be kept.

A number of tests were conducted to examine the strength of the model and fit to the variables. For each extraction the Kaiser-Meyer-Olkin Measure and Bartlett's Test of Sphericity were used to confirm that the sample was adequate and that the factor model produced was not inappropriate. Residuals from reproduced correlations were also examined to identify specific variables that were not being modeled well. In addition, the communalities of each variable were examined to determine how well the factor model explained its variance. From the initial extraction, two variables from the ITS (Belief 3, and Preparation – see Table 4), and one variable from the CVP (Closure – see Table 4) were dropped because of low communalities. The final extractions were based on 24 variables, each with communalities over 0.450.

Extracted factors were rotated using the Varimax method. The rotation converged in seven iterations. These seven factors explained 69.6% of the variance in the original set of data. An Anderson-Rubin technique was used with the final model to generate factor scores for each case without missing data. This technique produces scores with a mean of approximately zero and standard deviation of about one.

Pajares, 1996; Pajares & Graham, 1999; Pajares & Kranzler, 1995; Pajares & Miller, 1994; Shores & Shannon, 2007)" (Hoffman, in press, p. 1).

Table 5

Factors Related to Curriculum Implementation, Primary Loadings, Data Source.

FACTOR 1: STANDARDS-BASED INSTRUCTION	
Variable: Description (Data Source).	**Loading**
Focus on Sense-Making: Multiple solution strategies were encouraged, the enacted lesson developed procedural knowledge in meaningful ways and conceptual understanding of mathematics, and connections within mathematics were explored. (CVP)	.880
Reasoning about Mathematics: Students were afforded opportunities to make conjectures about mathematical ideas, students' mathematical arguments were challenged by others, mathematical authority rested with students, not with the teacher or textbook. (CVP)	.855
Students' Thinking in Instruction: Formative assessment techniques were used to guide instructional decision making, student statements were used to build a shared understanding, student misconceptions or mistakes were used as a learning site for others. (CVP)	.835
Presentation Fidelity Rating: Global rating of how closely the teacher's enactment of the lesson was consistent with authors' recommendations. (CVP)	.647
Belief 2: Teacher beliefs regarding didactic instruction. (ITS)	-.418

FACTOR 2: IMPLEMENTATION FIDELITY	
Variable: Description (Data Source).	**Loading**
TCT Index: Textbook Content Taught index represents the extent to which teachers, *when teaching textbook content*, followed their textbook, supplemented their textbook lessons, or used altogether alternative curricular materials. (TOC)	.826
Content Fidelity Rating: Global rating of how closely the content of the teacher's enacted lesson reflected the textbook content. (CVP)	.807
Rating: Teacher's reported rating of satisfaction with their textbook. (ITS)	.616
ETI Index: Extent of Textbook Implementation index represents the extent to which teachers followed their textbook. (TOC)	.514

FACTOR 3: TECHNOLOGY & COLLABORATIVE LEARNING	
Variable: Description (Data Source).	**Loading**
Technology by Students: Percent of lessons in which most students used graphing calculators. (CVP)	.783
Technology by Teacher: Percent of lessons in which the teacher used graphing calculator. (CVP)	.739
Belief 1: Teacher beliefs about reform-oriented approaches to teaching and learning mathematics. (ITS)	.529
Text: Number of years teaching from the district-adopted textbook. (ITS)	.492
Seating Arrangement: The seating arrangement of observed lessons. (CVP)	.463
Collaboration: The extent to which most students worked collaboratively during the lesson. (CVP)	.439

FACTOR 4: OPPORTUNITY TO LEARN	
Variable: Description (Data Source).	**Loading**
OTL Index: Percent of textbook lessons taught by the teacher during the school year. (TOC)	.876
ETI Index: Extent of Textbook Implementation index represents the extent to which teachers followed their textbook. (TOC)	.763
Seating Arrangement: The seating arrangement of observed lessons. (CVP)	-.514
Engage: The dominant level of student engagement in observed lessons. (CVP)	.408

Teacher-level factors. The seven factors extracted from our teacher data file generally clustered around two themes: (1) curriculum implementation, and (2) teacher characteristics. More specifically, four factors were related to teachers' implementation of curricular materials: (Factor 1) *Standards*-Based Instruction, (Factor 2) Implementation Fidelity, (Factor 3) Technology and Collaboration, and (Factor 4) Opportunity-to-Learn (see Table 5). Each of these factors embodies coherence in that the variables that "load" on each factor are related conceptually. For example, in Factor 1, the three elements of the Classroom Learning Environment—Focus on Sense-Making, Reasoning about Mathematics, and Students' Thinking in Instruction—have factor loadings between .835 and .880, compelling loadings indeed, and all gathered from observations of classroom practices.

As depicted in Table 6, three factors were related to teacher characteristics: NCTM Standards (Factor 5), Experience (Factor 6), and Professional Development (Factor 7). As was the case with implementation factors, each of the factors related to teacher characteristics reflects coherence in the data. For example, in Factor 5, teacher responses to the three questions about the NCTM Standards (2000)—Familiarity with, Agreement with, and Implementation of NCTM Standards—have factor loadings between .559 and .805, suggesting that these questions are closely related (as one might expect them to be). Moreover, scale scores for Belief 2 *didactic instruction* loaded negatively with Factor 5, suggesting that teachers holding strong beliefs about didactic instruction generally disagreed with NCTM's vision for school mathematics.

Table 6

Factors Related to Teacher Characteristics, Primary Loadings, Data Source

FACTOR 5: NCTM STANDARDS: FAMILIARITY, AGREEMENT, AND IMPLEMENTATION	
Variable: Description (Data Source).	**Loading**
Agreement: Extent to which the teacher agrees with the overall vision of *Principles and Standards for School Mathematics* (NCTM 2000) (ITS)	.805
Familiar: Extent to which the teacher is familiar with *Principles and Standards for School Mathematics* (NCTM 2000) (ITS)	.778
Implementation: Extent to which the teacher has implemented the recommendations of *Principles and Standards for School Mathematics* (NCTM 2000) (ITS)	.559
Belief 2: Teacher beliefs regarding didactic instruction (ITS)	-.453
Belief 1: Teacher beliefs about reform-oriented approaches to teaching and learning mathematics. (ITS)	.405

FACTOR 6: EXPERIENCE	
Variable: Description (Data Source).	**Loading**
Experience: Number of years teaching experience. (ITS)	.943
Mathematics Experience: Number of years mathematics teaching experience. (ITS)	.926
Text: Number of years teaching from the district-adopted textbook. (ITS)	.405

FACTOR 7: PROFESSIONAL DEVELOPMENT	
Variable: Description (Data Source).	**Loading**
PD_12: Number of hours of professional development in last 12 months. (ITS)	.874
PD_3: Number of hours of professional development in last 3 years. (ITS)	.864

The Principal Components Analysis (PCA) allowed us to responsibly and substantially reduce the number of teacher-level variables, resulting in a more manageable and coherent data set. Such results yielded interdependencies that were important for at least two reasons. First, while we suspected that several measures of curriculum implementation were related (e.g., TCT Index and Content Fidelity Rating), the PCA supported this notion empirically. Second, there were additional relationships among teacher variables that we did not anticipate but were discerned only through factor analysis. Moreover, this technique was likewise useful in ascertaining which teacher variables were tenuous. For example, Belief 3 simply did not "perform" well when examined in relation to every other measure; this variable was ultimately excluded from the final PCA because it explained the smallest portion of variance in the set of teachers' beliefs.

In summation, the PCA elucidated the key relationships among a large number of variables, and in so doing helped us to achieve data reduction. The fact that the teacher variables clustered around two themes, curriculum implementation and teacher characteristics, supports the notion that our copious teacher data are, after all, coherently related. Having achieved data reduction and coherence, we proceeded to examine the relationships among student outcome measures and teacher-level variables.

RESULTS

The National Science Foundation (2004) recommends that studies of curricular effectiveness report participation (i.e., testing) rates as well as attrition. Given our exclusive focus on year 1 data only, we do not report attrition rates herein. Nevertheless in this section we offer testing rates, as well as bivariate correlations, and partial correlations.

Testing Rates

During year 1, students were assessed on a fair test (Test A), a reasoning and problem solving test (Test B), and a standardized test (ITED 15). Overall, participation rates were remarkably high on any given test (see Table 7). As reported in Table 7, participation rates on Test A ranged from 94.5% of target students in District R to 98.8% in District W; on Test B, ranged from 92.5% in District R to 99.0% in District W; and on ITED-15, ranged from 93.4% in District R to 99.9% in District W. Not surprisingly, "complete" testing data was available for 84.1% of students in District R but as high as 97.3% in District W. As stated previously, nearly (99.77%) all students took at least one test in year 1 of the COSMIC Project.

Table 7
Participation Rates of Students in Year 1 Testing, by School District

	Test A		Test B		ITED-15		All Three Exams	
District	**Tests**	**Rate**	**Tests**	**Rate**	**Tests**	**Rate**	**Tests**	**Rate**
B	247	96.9	247	96.9	247	96.9	237	92.9
C	182	98.9	180	97.8	181	98.4	177	96.2
I	321	95.5	314	93.5	331	98.5	300	89.3
R	415	94.5	406	92.5	410	93.4	369	84.1
T	773	96.6	765	95.6	778	97.3	730	91.3
W	594	98.8	595	99.0	595	99.0	585	97.3
Total	**2,532**	**96.8**	**2,507**	**95.9**	**2,542**	**97.2**	**2,398**	**91.7**

Correlations between Student Outcome Measures and Teacher-Level Variables

It is important to note that the student outcomes used in the correlations that follow are IRT scale scores that were adjusted for students' prior achievement (CPA) and subsequently aggregated by teacher; that is, they represent mean residualized gain scores for each teacher. Curriculum Type is a dichotomous variable with integrated coded 1 and subject-specific coded 0. The time variable is a continuous variable and the implementation and teacher characteristic factors are centered on the grand mean ($\overline{X} = 0$, SD = 1).

The bivariate correlations between student outcomes, curriculum type, %FRL, use of time, implementation factors, and teacher characteristics appear in Table 8. However, it important to note that each correlation reported merely represents the magnitude of an association; the correlations *do not* represent causal relationships nor will they be used to model the variation of student achievement. Notwithstanding these caveats, in year 1, Curriculum Type was found to be significantly correlated with both Test A (.304) and Test B (.518). In particular, mean residualized gain scores (adjusted for prior achievement) were positively correlated in favor of teachers of one curriculum type. Bivariate correlations between Curriculum Type and scores on the ITED-15 were not significantly different than 0.

The %FRL variable was significantly and negatively correlated with scores on both Test A and ITED-15 in year 1. Thus, regardless of Curriculum Type, teachers with larger percentages of FRL students had generally lower scores on the fair test and standardized measure than teachers with lower percentages of FRL students. Teachers with higher scale scores for Technology & Collaboration (Factor 3) tended to have student scores that were higher on Test B and ITED-15. Opportunity to Learn (Factor 4) was significantly and positively correlated with higher performance on all three outcome measures, meaning that more OTL was associated with higher performance. With respect to teacher characteristics, Knowledge of Standards (Factor 5) was strongly correlated with scores on Test A and Test B, but exhibited no significant relationship with ITED-15. Teacher experience and professional development was not significantly related to scores on any student outcome measure.

Table 8

Correlations between Student Outcomes and Ten Potentially Related Variables

					Implementation Factors				Teacher Characteristics		
		Curr Type	**% FRL**	**Time LD**	**FAC1 (SBI)**	**FAC2 (Fidelity)**	**FAC3 (T&C)**	**FAC4 (OTL)**	**FAC5 (KoS)**	**FAC6 (Exp)**	**FAC7 (PD)**
Test A	*r*	.304	-.338	-.029	.157	-.189	.270	.388	.332	.246	.035
	p	.038*	.020*	.846	.293	.204	.067	.007**	.023*	.095	.816
Test B	*r*	.518	-.245	.183	.227	-.085	.387	.370	.331	.186	.089
	p	.000***	.096	.219	.126	.568	.007**	.010*	.023*	.212	.551
ITED-15	*r*	.264	-.318	.217	.231	-.115	.415	.291	.011	.254	-.172
	p	.073	.029*	.144	.117	.442	.004**	.047*	.942	.085	.247

*$p < .05$. **$p < .01$. ***$p < .001$.

We reiterate that associations do not necessarily imply *causal effects* and therefore sweeping generalizations are unwarranted; greater scrutiny of the data is required. It is worth noting that we also examined the association between student outcomes and *pairwise interactions* of teacher variables. Several key interactions were significantly correlated with student outcomes including: (a) Curriculum Type *x* %FRL, (b) Curriculum Type *x* Factor 3, (c) Curriculum Type *x* Factor 4, (d) %FRL *x* Factor 3, (e) %FRL *x* Factor 7, (f) Factor 3 *x* Factor 4, (g) Factor 4 *x* Factor 5, and (h) Factor 4 *x* Factor 7. Each of these interactions was examined in subsequent analyses.

Partial Correlations

Given the potential interdependency among the 10 teacher-level variables in Table 8, we performed additional analyses to examine correlations of individual variables when the effects of another variable was controlled or *partialled out*. Stated alternatively, we used partial correlations to determine the association of one teacher variable when another variable is essentially held constant. For example, when %FRL is partialled out, the correlation of Curriculum Type is significantly correlated with scores on each dependent measure, Test A, Test B, and ITED-15 (Table 9). More specifically, when controlling for %FRL, the magnitude of the correlation between Curriculum Type and student outcomes becomes significantly different than 0, and this was the case for all three tests. When holding %FRL constant, teacher scores for Factor 3 (Technology & Collaboration) and Factor 5 (Knowledge of Standards) become significantly (and positively) correlated with two of the three tests. Moreover, the importance of OTL (Factor 4) is substantially reduced with the partialling out of %FRL, suggesting that %FRL and OTL may be closely related.

Table 9

Correlations between Student Outcomes and Other Variables, Partialling Out %FRL

				Implementation Factors				Teacher Characteristics		
	Curr Type		**Time LD**	**FAC1 (SBI)**	**FAC2 (Fidelity)**	**FAC3 (T&C)**	**FAC4 (OTL)**	**FAC5 (KoS)**	**FAC6 (Exp)**	**FAC7 (PD)**
Test A	.459 .001**		-.017 .910	.141 .349	-.172 .254	.275 .065	.259 .082	.347 .018*	.166 .270	.087 .565
Test B	.646 .000***		.198 .187	.216 .149	-.067 .656	.390 .007**	.291 .050	.337 .022*	.124 .411	.128 .398
ITED-15	.404 .005**		.242 .106	.221 .140	-.094 .535	.426 .003**	.149 .324	.006 .967	.180 .232	-.137 .363

*p < .05. **p < .01. ***p < .001.

As represented in Table 10, when partialling out Factor 3 (Technology & Collaboration), the magnitude of the correlations between Curriculum Type and student outcomes is weakened, with significance found only on Test B. When holding this factor constant, OTL maintains a significant relationship with Test A and Test B, and significance is attained for Knowledge of Standards on Test A and Test B.

Table 10

Correlations between Student Outcomes and Other Variables, Partialling Out Technology & Collaboration

				Implementation Factors				Teacher Characteristics		
	Curr Type	**% FRL**	**Time LD**	**FAC1 (SBI)**	**FAC2 (Fidelity)**	**FAC3 (T&C)**	**FAC4 (OTL)**	**FAC5 (KoS)**	**FAC6 (Exp)**	**FAC7 (PD)**
Test A	.196 .192	-.342 .020*	-.157 .297	.171 .254	-.178 .237		.375 .010*	.379 .009**	.264 .076	.014 .929
Test B	.398 .006**	-.252 .091	.031 .838	.259 .083	-.065 .667		.358 .015*	.409 .005**	.213 .155	.063 .678
ITED-15	.053 .726	-.334 .023*	.058 .700	.268 .071	-.096 .524		.271 .068	.064 .672	.293 .048*	-.227 .129

*p < .05. **p < .01. ***p < .001.

When partialling out the effect of OTL, Curriculum Type become significantly correlated with each of the three outcome measures (see Table 11) and mediates the effect of %FRL. Knowledge of Standards (Factor 2) continues to be significantly (and positively) correlated with scores on the project-developed tests.

Table 11

Correlations between Student Outcomes and Other Variables, Partialling Out OTL

				Implementation Factors				Teacher Characteristics		
	Curr Type	**% FRL**	**Time LD**	**FAC1 (SBI)**	**FAC2 (Fidelity)**	**FAC3 (T&C)**	**FAC4 (OTL)**	**FAC5 (KoS)**	**FAC6 (Exp)**	**FAC7 (PD)**
Test A	.411 .005**	-.165 .274	-.052 .733	.155 .303	-.251 .093	.248 .097		.330 .025*	.286 .054	.122 .421
Test B	.638 .000***	-.057 .707	.178 .237	.230 .124	-.135 .372	.375 .010*		.327 .027*	.217 .147	.176 .241
ITED-15	.335 .023*	-.199 .184	.212 .157	.231 .122	-.153 .310	.403 .006**		-.012 .939	.279 .060	-.124 .413

*p < .05. **p < .01. ***p < .001.

Because the interaction between Curriculum Type and Technology & Collaboration (Factor 4) was significantly correlated with student achievement, we examined its effect by partialling the interaction out of the initial bivariate correlations in Table 8. As reported in Table 12, when holding constant the interaction between Curriculum Type and Factor 3, scores on Test B continue to be significantly correlated with Curriculum Type. When controlling for this same interaction, both Knowledge of Standards (Factor 5) and OTL emerge as significantly correlated with both Test A and Test B.

Table 12

Correlations between Student Outcomes and Other Variables, Partialling Out the Interaction between Curriculum and Technology & Collaboration

				Implementation Factors				Teacher Characteristics		
	Curr Type	**% FRL**	**Time LD**	**FAC1 (SBI)**	**FAC2 (Fidelity)**	**FAC3 (T&C)**	**FAC4 (OTL)**	**FAC5 (KoS)**	**FAC6 (Exp)**	**FAC7 (PD)**
Test A	.202 .178	-.311 .035*	-.083 .584	.146 .334	-.183 .224	.165 .274	.386 .008**	.386 .008**	.220 .142	.070 .644
Test B	.438 .002**	-.209 .162	.140 .353	.220 .141	-.073 .631	.288 .053	.369 .012*	.393 .007**	.152 .312	.133 .379
ITED-15	.149 .323	-.289 .052	.178 .235	.225 .132	-.105 .489	.325 .027*	.283 .057	.047 .757	.227 .129	-.150 .321

*p < .05. **p < .01.

Similarly, because the interaction between Curriculum Type and %FRL was significantly correlated with student achievement, we examined its effect by partialling out this interaction from our initial bivariate correlations. When holding constant the interaction between Curriculum and %FRL, many correlations become mediated (Table 13). Nevertheless, despite the overall mediation in correlations, Curriculum Type is significantly correlated with each of the three dependent measures: .399 (Test A), .604 (Test B), and .320 (ITED-15). Moreover, Factor 3 (Technology & Collaboration) becomes significantly correlated with Test B and the ITED-15 while Knowledge of Standards hold significant relationships with the two project-developed measures.

Table 13
Correlations between Student Outcomes and Other Variables, Partialling Out the Interaction between Curriculum and %FRL

				Implementation Factors				Teacher Characteristics		
	Curr Type	**% FRL**	**Time LD**	**FAC1 (SBI)**	**FAC2 (Fidelity)**	**FAC3 (T&C)**	**FAC4 (OTL)**	**FAC5 (KoS)**	**FAC6 (Exp)**	**FAC7 (PD)**
Test A	.399 .006**	-.078 .608	.005 .973	.080 .595	-.112 .460	.201 .180	.207 .168	.329 .025*	.130 .388	.125 .409
Test B	.604 .000***	-.019 .899	.227 .130	.171 .255	-.012 .935	.341 .020*	.229 .126	.322 .029*	.085 .575	.164 .277
ITED-15	.320 .030*	-.151 .318	.256 .086	.182 .227	-.052 .729	.375 .010*	.152 .313	-.019 .899	.173 .251	-.127 .402

*$p < .05$. **$p < .01$. ***$p < .001$.

Finally, with each variable that was partialled out, we monitored changes in the magnitude of correlations between Curriculum Type and student outcomes, and several are worth reporting. For example, when %FRL is partialled out, the magnitude of correlations related to Curriculum Type increase in the range of .128 to .155, suggesting that %FRL should be included in multi-level modeling of student achievement. Partialling out Technology & Collaboration decreased the magnitude of correlations related to Curriculum Type by .108 to .211, suggesting that Factor 3 should be entered into the model. Additionally, partialling out OTL increased the correlation for Curriculum Type between .071 and .120. No interactions affected the magnitude of the correlation between Curriculum Type and student outcomes in substantial ways.

DISCUSSION

Until recently, there existed scant data regarding the effectiveness of "reform" curricula, either in their own right or in relation to more traditional approaches (Schoenfeld, 2006). The relatively few studies of curricular effectiveness have been wrought with limitations and the subject of much scrutiny. In particular, previous investigations of NSF-funded curricula were often field tests and, as such, questions were raised about the objectiveness of researchers who conducted the studies (NRC, 2004). Moreover, previous curriculum evaluation studies typically have not carefully accounted for *curriculum implementation* and, as a consequence, the degree to which curriculum explains variation in students' mathematics achievement cannot be ascertained. The COSMIC Project has sought to address these deficiencies by conducting a systematic examination of the many variables, including student- and teacher variables as well as curricular effects that might contribute to student learning in mathematics. While our analyses are ongoing and multi-level modeling of student achievement has yet to be completed, our preliminary analyses have provided insight into key issues in data collection, reduction, and coherence. Although we do not directly address the question "which curriculum is best?", we offer ideas for the identification of student- and teacher variables in studies of curricular effectiveness.

Not surprisingly, our analyses revealed a significant and positive relationship between students' prior achievement and their scores on a variety of outcome measures. In particular, prior achievement correlated with scores on Test A, Test B, and ITED-15 in the magnitude of .70, suggesting that much of student performance on year 1 assessment can be attributed to their initial referent of mathematics achievement. These strong correlations between baseline and outcome measures suggest that prior achievement must be carefully considered in searching for other determinants of student outcomes.

The importance of prior achievement cannot be discounted, but there were significant methodological challenges associated with specialized context of our research project that had to be overcome. More specifically, given that participating schools were located in numerous US states, it was necessary to place students on a common prior achievement scale. Without such, valid comparisons across schools would simply be impossible. Accordingly, in this study, we developed an algorithm (Figure 2) for generating the COSMIC Prior Achievement (CPA) scale. In doing so, the CPA enabled us to examine student outcomes after adjusting for prior achievement; that is, the CPA scale score served as a covariate in subsequent analyses. Both within and across schools, there was substantial variation in test scores (prior- and post-achievement); the existence of variation is necessary in order to identify correlations between Curriculum Type, student variables, and teacher factors and to enable the construction of statistical models of student achievement.

In the COSMIC Project, it was important to determine the equivalence (or comparability) of student groups prior to their experience in one of two curricular treatments, *integrated* or *subject-specific*. Our analysis of CPA scale scores confirmed our sample selection criterion that students not be *tracked* into one curriculum path or another. This finding is important for at least two reasons. First, the comparability of prior achievement is consistent with our stipulation that participating schools provide a *free choice* between parallel curricular options. Although students were not randomly assigned to treatment groups, the two sets of students nonetheless appear to have begun with similar prior knowledge of mathematics. It follows that differential performance across curriculum types cannot be attributable to differences in student characteristics, namely prior achievement.

Another key finding is the relationship between opportunity-to-learn (OTL) and student learning. In particular, we found that OTL was the only variable to significantly correlate with *all three* dependent measures. This finding is consistent with a growing body of evidence suggesting "that students do not learn content to which they are not exposed" (Stein, Remillard, & Smith, 2007, p. 327). Moreover, this suggests that studies of curricular effectiveness should take into account whether students are afforded equitable opportunities to study the content on which they are likely to be assessed. As reported in another paper (McNaught et al., 2010), teachers of integrated curricula covered significantly less textbook content than teachers of subject-specific curricula. This finding may have moderated the effect of Curriculum Type. Without consideration of OTL, Curriculum Type was significantly correlated with two of three dependent measures. However, when controlling for OTL, correlations between Curriculum Type all three dependent measures were strengthened. A similar pattern emerged with %FRL, suggesting that %FRL and OTL may be acting as suppressor variables in the relationship between Curriculum Type and mean gains on the tests, and therefore will be important variables to

include in our modeling processes.

Of further importance is the relationship between OTL and other variables. In particular, initial bivariate correlations indicated a significant negative relationship between FRL status and student scores on Test A and ITED-15. However, when controlling for OTL, the correlation between %FRL and dependent measures decrease in magnitude, and they are no longer significantly different than 0. Likewise, when %FRL is partialled out, the relationship between OTL and student outcomes is diminished to a level of insignificance. Stated differently, by holding %FRL constant, we learn that the effect of OTL is reduced substantially. This finding suggests that OTL and %FRL are likely closely affiliated; this is, there is likely overlap in what each variable is measuring. On some level, the relationship between OTL and %FRL may be attributable to the differential (slower) pace of content coverage in classes with more FRL students. By controlling for OTL and %FRL, we can more carefully measure the impact of curriculum on student learning.

The NRC (2004) stipulates that examinations between curriculum and student learning must carefully document teachers' use of curricular materials. In this study, we measured *fidelity of implementation* in ways that were previously unprecedented. As depicted in Table 4, we measured more than 30 teacher variables that potentially held predictive power in explaining student outcomes. While we retain our position that multiple measures of curriculum implementation are important, we concede that it is possible to attend to too many variables. In theory, each of the 31 teacher variables could be utilized in the construction of models of student achievement. However, in doing so, it is conceivable that *all* (or nearly all) variation is collectively accounted for by the 31 variables. Such models would inherently offer precision that is unwarranted and therefore inappropriate. Consequently, it was necessary to explore the relationship between teacher variables—*independent of student outcomes*—to achieve coherence in the data and ultimately position us to construct relatively parsimonious models.

We successfully reduced the number of teacher variables to "only" seven, and these factors clustered around two coherent themes, curriculum implementation and teacher characteristics. With regard to the former, it was somewhat surprising that neither Standards-Based Instruction (Factor 1) nor Implementation Fidelity (Factor 2) were significantly correlated with any of the dependent measures. These findings might be considered counterintuitive and certainly warrant further research. Similarly, with regard to teacher characteristics, some might be surprised to learn of the lack of significance of Experience (Factor 6) and especially Professional Development (Factor 7) in relation to student outcomes. This result is consistent with a recent report of the American Institutes for Research (AIR) that found an intensive teacher professional development program for middle school mathematics teachers had no statistically significant impact on student achievement (Garet et al., 2010). With respect to our findings, one might ask, "Was the PD long enough?", "Was the PD sustained?", "Was the PD targeted (or focused) enough?", "Were teachers receptive to the PD?" It is plausible that the *amount* of professional development may not represent an ideal measure of its impact on instructional practices.

Finally, although teacher variables clustered around two coherent themes, it seems reasonable that a single measure *within each factor* might serve as a proxy in analyses of student achievement data. Consider the case of Technology & Collaboration (Factor 3). For this factor, scale scores for each teacher were generated using data from Classroom Visit Protocols and the Initial Teacher Survey. Rather than collect data on the extent to which

students are using technology (*Tech_Students*), the teacher is using technology (*Tech_Teacher*), whether students are seated in groups (*Seating*) and working collaboratively (*Collaboration*), and teachers' beliefs regarding reform practices (*Belief 1*), is it possible that only one of these variables essentially "tells the story"? If one variable were to effectively serve as a proxy, this would considerably reduce the cost of data collection, render more parsimonious models, and provide greater ease in the interpretation of coefficients used in hierarchical models of student achievement. Through ongoing analyses, we continue our long journey toward the attainment of such models.

CONCLUSION

The historical sub-standard performance of US mathematics students in international comparisons has resulted in a reform movement, including the development, adoption, and enactment of NSF-funded integrated curricula in secondary classrooms. In some US schools, dual curricular options represent a compromise between seemingly warring factions that each profess to know which curriculum is *best*. The COSMIC Project has embarked on a rigorous study to examine the impact of curriculum and numerous other variables that have the potential to explain student achievement in mathematics. However, the results reported herein represent only one piece of the large, complex puzzle of developing comprehensive profiles of what students learn when mathematics content is organized in fundamentally different ways. In short, correlational studies are insufficient to determine whether traditional or integrated approaches to curriculum organization yield different profiles of student learning. The application of more sophisticated analytic techniques such as HLM have not been completed and therefore readers are urged to exercise great caution when interpreting results reported in this manuscript. Data represent the findings for year 1 only; longitudinal findings as well as result of year 2 and year 3 analyses will provide important additions to our understanding of the impact of mathematics curriculum on student learning. This paper provides insight into the design of studies of curricular effectiveness and offer recommendations related to key issues in data collection, reduction, and coherence that are prerequisite to the productive modeling of student achievement data.

REFERENCES

Cai, J., & Moyer, J. C. (2006). *A conceptual framework for studying curricular effects on students' learning: Conceptualization and design in the LieCal Project*. Newark, DE: The University of Delaware.

Chavez, O., Papick, I., Ross, D. J., & Grouws, D. A. (2010). *The essential role of curricular analyses in comparative studies of mathematics achievement: Developing "fair" tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO, May 2010.

Elis, M. E., Malloy, C. E., Meece, J. L. & Sylvester, P. R. (2007). Convergence of observer ratings and student perceptions of reform teaching practices in sixth-grade mathematics classrooms. *Learning Environments Research*, *10*(1), 1-15.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., and Doolittle, F. (2010). *Middle School Mathematics Professional Development Impact Study: Findings After the First Year of Implementation* (NCEE 2010-4009). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Harwell, M. R., Post, T. R., Maeda, Y., Davis, J. D., Cutler, A. L., & Kahan, J. A. (2007). *Standards*-based mathematics curricula and secondary students' performance on standardized achievement tests. *Journal for Research in Mathematics Education*, *38*(1), 71-101.

Harwell, M. R. & LeBeau B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher, 39*(2), 120-131.

Hoffman, B. (in press). "I think I can, but I'm afraid to try": The role of self-efficacy beliefs and mathematics anxiety in mathematics problem-solving efficiency. *Learning and Individual Differences*.

McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P., Bugliari, D., & Robyn, A. (2001). Interactions among instructional practices, curriculum and student achievement: The case of *Standards*-based high school mathematics. *Journal for Research in Mathematics Education*, *32*(5), 493-517.

McNaught, M.D., Tarr, J. E. & Sears, R. (2010). *Conceptualizing and measuring fidelity of implementation of secondary mathematics textbooks: Results of a three-year study.* Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO, May 2010.

Nathan, M. J., Long, S. D., & Alibali, M. W. (2002). Symbol precedence in mathematics textbooks: A corpus analysis. *Discourse Processes, 33*, 1-21.

National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: The Council.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics.* Reston, VA: The Council.

National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academies Press.

National Center for Education Statistics (2007). *Mapping 2005 State Proficiency Standards Onto the NAEP Scales* (NCES 2007-482). U.S. Department of Education. Washington, DC: Author.

Nie, Y. & Lau, S. (in press). Differential relations of constructivist and didactic instruction to students' cognition, motivation, and achievement. *Learning and Instruction*.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, 7(1).

Post, T. R., Harwell, M. R., Davis, J. D., Maeda, Y., Cutler, A. & Andersen, E. (2008). *Standards*-based mathematics curricula and middle-grades students' performance on

standardized achievement tests. *Journal for Research in Mathematics Education, 39*(2), 184-212.

Schafer, W. D. (2005). Review [of the Kaufman Iowa Tests of Educational Development, Forms A and B]. In R. A. Spies & B. S. Plake (Eds.), *The Sixteenth Mental Measurements Yearbook* (pp. 488-491). Lincoln, NE: Buros Institute of Mental Measurements.

Schoen, H. L., Cebulla, K. J., Finn, K. F., & Fi, C. (2003). Teacher variables that relate to student achievement when using a *Standards*-based curriculum. *Journal for Research in Mathematics Education, 34*(3), 228-259.

Schoenfeld, A.H. (2004). The math wars. *Educational Policy*, 18, 253–286.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the works clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher*, 35(2), 13–21.

Senk, S. L., & Thompson, D. R. (2003). (Eds.) *Standards-based school mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Silver, E. A., Alacaci, C., & Stylianou, D.A. (2000). Students' performance on extended constructed-response tasks. In Silver & Kenney (Eds.), *Results from the Seventh Mathematics Assessment of the National Assessment of Educational Progress* (pp. 301-341). Reston, VA: NCTM.