Journal for Research in Mathematics Education
2013, Vol. 44, No. 4, 683–729

The Effects of Content Organization and Curriculum Implementation on Students' Mathematics Learning in Second-Year High School Courses

James E. Tarr and Douglas A. Grouws
University of Missouri

Óscar Chávez
University of Texas at San Antonio

Victor M. Soria
University of Missouri

We examined curricular effectiveness in high schools that offered parallel paths in which students were free to study mathematics using 1 of 2 content organizational structures, an integrated approach or a (traditional) subject-specific approach. The study involved 3,258 high school students, enrolled in either Course 2 or Geometry, in 11 schools in 5 geographically dispersed states. We constructed 3-level hierarchical linear models of scores on 3 end-of-year outcome measures: a test of common objectives, an assessment of problem solving and reasoning, and a standardized achievement test. Students in the integrated curriculum scored significantly higher than those in the subject-specific curriculum on the standardized achievement test. Significant student-level predictors included prior achievement, gender, and ethnicity. At the teacher level, in addition to Curriculum Type, the Opportunity to Learn and Classroom Learning Environment factors demonstrated significant power in predicting student scores, whereas Implementation Fidelity, Teacher Experience, and Professional Development were not significant predictors.

Key words: Curricular effectiveness; Curriculum; HLM; Integrated curriculum; Secondary mathematics

---

The research reported in this article was supported by the National Science Foundation under Grant No. (REC-0532214). The study was conducted as part of the Comparing Options in Secondary Mathematics: Investigating Curricula (COSMIC) project, http://cosmic.missouri.edu, supported by the National Science Foundation under 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 authors wish to thank Robert Ankenmann, Angela Bowzer, Dean Frerichs, Wei-Min Hsu, Melissa McNaught, Ira Papick, Greg Petroski, Robert Reys, Daniel Ross, Ruthmae Sears, and R. Didem Taylan for their insights into various stages of the research. We also wish to express our gratitude to the Editor and our five anonymous reviewers for their helpful comments and suggestions.

During the past half century, arguably, no discipline has been the focus of more intense public scrutiny than school mathematics. For decades, educational reform initiatives have spurred debate over what mathematics should be learned, by whom it should be learned, and when it should be learned (Jones, 1970; Romberg, 2010). Waves of reform in school mathematics include, but are not limited to, the "new math" programs of the late 1950s and 1960s, the "back-to-basics" movement of the 1970s, and the problem solving focus promoted by the *Agenda for Action* (National Council of Teachers of Mathematics [NCTM], 1980). These movements were followed by the Standards documents for school mathematics (NCTM, 1989, 1991, 1995, 2000), No Child Left Behind's research-based programs for teaching and learning (No Child Left Behind [NCLB], 2002), and the *Common Core State Standards for Mathematics* (National Governors Association Center for Best Practices [NGA Center] and the Council of Chief State School Officers [CCSSO], 2010a). Each of these initiatives has sought to increase the breadth or depth of mathematics to be taught in order to raise student achievement, better prepare students for college success, and increase our competitive status in an increasingly global economy.

Because mathematics textbooks are a primary determinant in teachers' selection of lesson content and pedagogical techniques (Grouws & Smith, 2000; Grouws, Smith, & Sztajn, 2004; Weiss, Pasley, Smith, Banilower, & Heck, 2003) and are the centerpiece of U.S. mathematics instruction (Hiebert et al., 2003), curricular materials have the potential to be instruments of reform. Accordingly, and reflecting national priorities in the 1990s, the National Science Foundation (NSF) invested millions of dollars in the development of reform curricula that differed from traditional mathematics textbooks in fundamental ways. Most notably, NSF-funded curricula dramatically restructured the organization of mathematics content by integrating several branches of mathematics within high school courses, focusing on the development of mathematical thinking and problem solving, and deemphasizing skills and symbol manipulation (Nathan, Long, & Alibali, 2002).

The Need for Rigorous, Scientific Curriculum Research

As NSF-funded curricula made inroads into the textbook market, the federal government began to play a more powerful role in education. Under federal educational policies, No Child Left Behind (NCLB, 2002) stipulated that public schools use scientific, research-based programs for improving mathematics achievement for all students. To that end, the What Works Clearinghouse (WWC) established standards for reviewing and synthesizing educational research in order to assess the rigor of research evidence on the effectiveness of interventions, including curricular programs developed with NSF funding (United States Department of Education Institute of Educational Sciences, 2011). It was argued that evidence standards were necessary because, historically, studies of curricular effectiveness have been plagued by numerous methodological limitations. Few have used an

experimental design, made provisions for the multistructured nature of educational data, included multiple measures of student learning outcomes, or been sensitive to treatment integrity (National Research Council [NRC], 2004). Moreover, many studies suffered from a perceived evaluator bias because most were conducted in the context of field tests in which evaluators were not independent of the program developers (NRC, 2004; Senk & Thompson, 2003). However, according to the WWC, "currently, only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence, while quasi-experimental designs (QEDs) with equating may only meet standards with reservations" (U.S. Department of Education Institute for Educational Sciences, 2011, p. 11). Applying these evidence standards to secondary mathematics curricular programs, the WWC declared that one NSF-funded textbook series, the Core-Plus Mathematics Project, demonstrated "potentially positive effects on mathematics achievement for high school students" (U.S. Department of Education Institute for Educational Sciences, 2010, p. 2). However, this conclusion of curricular effectiveness was based on a single study, Schoen and Hirsch (2003), which met the WWC evidence standards "with reservations." The remaining 16 studies of Core-Plus did not satisfy evidence standards, leaving school districts without a strong evidentiary base for scientific, research-based programs in secondary mathematics.

Parallel Curricular Paths

Despite the more recent efforts of researchers to design studies that satisfy the evidence standards of the WWC, debates continue to rage regarding the relative merits of NSF-funded curricular programs. Traditionalists argue 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, 2004, p. 253). As a compromise to assuage concerns held by stakeholders (i.e., administrators, teachers, parents, and students), some school districts recently began to offer parallel curricular paths in which students are free to study mathematics using one of two content organizational schemes, an integrated approach or a (traditional) subject-specific approach. Such parallel pathways are contrasted in *Designing High School Mathematics Courses Based on the Common Core State Standards* (NGA Center & CCSSO, 2010b) as:

> (1) An approach typically seen in the United States (Traditional) that consists of two algebra courses and a geometry course, with some data, probability and statistics included in each course; [and] (2) An approach typically seen internationally (Integrated) that consists of a sequence of three courses, each of which includes number, algebra, geometry, probability and statistics. (p. 3)

It is in the context of high schools offering parallel curricular paths that the Comparing Options in Secondary Mathematics: Investigating Curricula (COSMIC) project addresses the need for well-designed comparative studies of curricular effectiveness. The primary goal of the COSMIC project is to examine high school

students' mathematics learning from textbooks embodying two fundamentally distinct approaches to content organization: an integrated approach (Core-Plus Mathematics Project) and a subject-specific approach (in which students follow an Algebra I, Geometry, Algebra II sequence). Our research design enables the systematic study of complex relationships between curriculum organization, curriculum implementation, teacher and student characteristics, and high school students' mathematics learning using multiple achievement measures.

In the first in a series of project-related studies, Grouws et al. (2013) investigated the differential effects of mathematics content organization on students' learning in their first-year high school mathematics course. The Year-1 study involved 2,161 students in 10 schools in 5 states in which approximately one half of the students studied from an integrated curriculum (Core-Plus Course 1), and one half studied from a subject-specific curriculum (Algebra 1). Taking account of curriculum implementation and student prior achievement, hierarchical linear modeling with three levels showed that students who studied from the integrated curriculum were significantly advantaged over students who studied from a subject-specific curriculum on three end-of-year outcome measures: a test of common objectives, an assessment of problem solving and reasoning, and a standardized achievement test. These results underscore the significance of mathematics content organization in influencing student learning in first-year high school mathematics courses and how teachers are the ultimate arbiters of students' opportunity to learn.

Research Questions

Despite differential curriculum effects on students' learning in first-year high school mathematics (Grouws et al., 2013), little is known about the influence of mathematics content organization in second-year high school mathematics courses, Geometry and Course 2 of an integrated program. Specifically, in the second in a series of studies conducted by the COSMIC project, we examined the following research questions:

1. Is there a differential mathematics learning effect when high school students study from an integrated textbook (Course 2) and when students study from a subject-specific textbook (Geometry)?
2. What are the relationships between curriculum type, curriculum implementation, and student learning in second-year high school mathematics courses? In particular, what student characteristics and teacher practices and characteristics are associated with students' learning in second-year high school mathematics courses?

In the following sections, we describe the related literature and conceptual framework that informed our research. We provide a robust description of our research design and methodology and report models of student achievement. In a final section, we discuss our results and offer implications for research.

Review of Related Literature

Stein and Kaufman (2010) argued that teachers' use of curricular materials has a greater influence on student learning than teacher characteristics such as education level, experience, and knowledge of mathematics teaching. With respect to experience, there is evidence that some beginning teachers rely more heavily on mathematics textbooks than more experienced teachers, who draw on their own experiences when planning and implementing instruction (Remillard & Bryans, 2004; Tarr, Chávez, Reys, & Reys, 2006). In fact, there is evidence that more experienced teachers may be resistant to key features of recent reform curricula and instead revert to the use of conventional textbooks with which they are more accustomed to teaching (Remillard & Bryans, 2004; Superfine, 2009). Similarly, teachers are more likely to follow the curriculum when their beliefs align with the pedagogical orientation of the program (Remillard & Bryans, 2004; Stein & Kaufman, 2010; Stein, Remillard, & Smith, 2007); when teachers' beliefs are in conflict with underlying program theory, they may be resistant to using the curricular materials (Bowzer, 2008; Chávez, 2003).

Numerous studies have documented considerable variability in how teachers use curricular materials in teaching mathematics (e.g., Cai, Wang, Moyer, Wang, & Nie, 2011; Kilpatrick, 2003; Stein & Kaufman, 2010; Tarr, McNaught, & Grouws, 2012). With respect to content selection, teachers can emphasize some mathematics topics at the expense of others, teach textbook lessons in an alternative sequence, adapt textbook lessons in purposeful ways, use alternative curricular materials, or skip entire chapters of textbook content (Grouws et al., 2013; Tarr et al., 2006). With respect to pedagogical considerations, teachers may have students work in collaborative groups, use graphing calculators, or provide instruction on a continuum of didactic lecture to inquiry-based learning (Chávez, 2003; Grouws et al., 2013; Moyer, Cai, Wang, & Nie, 2011). Such variation underscores that despite heavy textbook "use," teachers are active developers of the enacted curriculum, which is shaped by their experiences, knowledge, beliefs, and classroom characteristics (Ball & Cohen, 1996; Remillard & Bryans, 2004; Stein & Kaufman, 2010; Stein et al., 2007; Tarr et al., 2006; Thompson & Senk, 2010).

In addition to recommending the careful documentation of curriculum use, the National Research Council (2004) suggested that future research also include studies of student learning in schools employing districtwide adoptions of mathematics curricula in which teachers are required to implement the programs. Consistent with this recommendation, Post et al. (2008) studied the effects of student- and classroom-level variables in modeling student achievement using two different NSF-funded middle grades mathematics textbook series, the Connected Mathematics Project and MATHThematics, compared to national norms. Hierarchical linear models controlled for student-level variables, such as prior mathematics achievement, gender, and classroom-level predictors including school location, hours of teacher professional development, and aggregate measures of ethnicity, socioeconomic status (SES), and special education services. At the student level, Post et al. found significantly lower performance for low-SES students, African American students,

nonnative English speakers, and special education students. At the classroom level, students of the NSF-funded curricula scored significantly higher on problem solving and open-ended items than on procedural (computational) questions on standardized achievement tests.

Employing a similar research design, Harwell et al. (2007) developed hierarchical linear models of mathematics achievement for students in five school districts studying from one of three integrated high school curricula, Interactive Mathematics Program, Mathematics: Modeling Our World, and Core-Plus Mathematics Project. Four of five participating school districts scored above average on the mathematics portion of the Minnesota Basic Skills Test, their baseline measure of students' prior knowledge. At the student level, models controlled for prior achievement, SES (using free or reduced lunch as a proxy), gender, and attendance, and at the classroom level accounted for ethnicity, aggregate measures of SES, English Language Learners, special education students, and female students as well as attendance, curriculum type, and school district affiliation. Results indicated that SES level and prior achievement were significant but modest predictors of student learning. Neither gender nor attendance accounted for a significant amount of variation in outcomes. Although there were no significant differences in performance across curriculum types, students in all three programs generally scored at or above the national mean on the standardized achievement subtests. However, it is worth noting that these students as a group also scored above national norms on the standardized test used as a measure of prior student achievement.

Despite the sophisticated analytic techniques employed, Post et al. (2008) and Harwell et al. (2007) did not gauge classroom instruction or characterize the implementation of curricula. By way of contrast, the Longitudinal Investigation of the effect of Curriculum on Algebra Learning (LieCal) project (Moyer et al., 2011) studied the effects of an NSF-funded curriculum (Connected Mathematics Project [CMP]) and publisher-developed curricula using a variety of data sources, including observations in middle grade classrooms. In a related analysis using a two-level HLM growth curve and repeated measures ANOVA, the researchers found significantly greater growth in problem-solving skills for students using the NSF-funded curriculum but detected no differences in symbolic manipulation skills across curriculum types. Moreover, in addition to curriculum type, they found the nature and quality of classroom instruction to be significant predictors of students' achievement gains over the 3 middle school years on at least some outcome measures. In a pivotal study of the Connected Mathematics Project, King et al. (2011) measured use of curricular materials using a curriculum coverage index and measured curriculum adaptation using a lesson modification index. Using HLM, they found both indices to be significant predictors of student achievement; student scores were significantly higher among teachers who covered more of the curriculum and were likewise higher when the teacher adapted the curriculum more than average. The results of Moyer, Cai, Wang, and Nie (2011) and King et al. (2011) warrant careful documentation of both opportunity to learn and fidelity of implementation of curricular materials.

Notwithstanding recent studies of curricular effectiveness (Cai et al., 2011;

Grouws et al., 2013; Harwell et al., 2007; King et al., 2011; Post et al., 2008), there are opposing perspectives on how teachers should use curricular materials (Huntley & Chval, 2010). One view is that teachers should carefully follow their textbooks because (a) curriculum developers are perceived "experts" in their field, and (b) they were written, piloted, refined, and field-tested over several years (Snyder, Bolin, & Zumwalt, 1992). A competing view is that teachers should adapt the curricular materials in response to the needs of their students and specialized teaching context (McClain, Zhao, Visnovska, & Bowen, 2009; Remillard, 2005). We, along with others, take the position that causal links between mathematics curriculum and student learning are not defensible without the careful documentation of how teachers enact the textbook curriculum (NRC, 2004; Thompson & Senk, 2010). Without such, even valid and reliable measures of student learning may not be sufficient to accurately gauge curricular effectiveness (George, Hall, & Uchiyama, 2000; NRC, 2004; Thompson & Senk, 2010).

Conceptual Framework for Curriculum Research

In concert with the related research, our research is informed by the Center for the Study of Mathematics Curriculum (CSMC) framework (Center for the Study of Mathematics Curriculum, n.d.), depicted in Figure 1. We accept the notion that

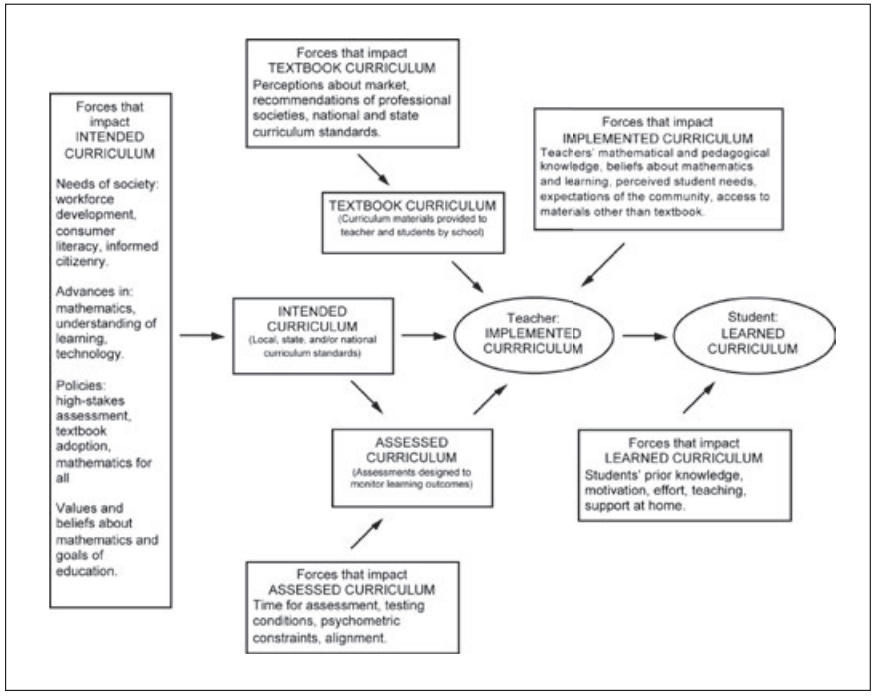

Figure 1. Curriculum Research Framework (Center for the Study of Mathematics Curriculum, n.d.).

teachers are active developers of the enacted curriculum, transforming the *textbook curriculum* into an *implemented curriculum* that may be markedly different in nature from the developers' intentions. For example, teachers may teach most of the textbook or not; they may supplement the textbook or not; they may assign homework directly from the textbook or not; they may encourage students to use technology or dissuade them from such use; they may emphasize some mathematics topics at the expense of others; and they may write assessments that emphasize procedural fluency over conceptual understanding, or vice versa. Because textbooks are not self-enacting, we invested heavily in the documentation of teachers' implementation of curricular materials in order to gauge fidelity of implementation; that is, we examined the extent to which the implemented curriculum is attributable to the textbook curriculum. As experienced by students, the implemented curriculum, in turn, influences the *learned curriculum*. Furthermore, consistent with the CSMC framework, we acknowledge that the implemented curriculum is not the sole determinant of the learned curriculum because student learning is also influenced by numerous *student characteristics* including (but not limited to) prior achievement, motivation, teaching, and home support. Therefore, in addition to carefully documenting the implemented curriculum that students experience in mathematics classrooms, we sought trusted measures of students' prior learning and collected information on gender, race, socioeconomic status, and qualifications for special education services.[1] These key forces that impact student learning (Figure 1) are described in a subsequent section. Although curriculum innovations are often considered a lever for educational reform, classroom teachers, who often are allowed to operate autonomously, ultimately implement them. It is therefore imperative that studies of curricular effectiveness carefully document the relationships between the textbook curriculum, implemented curriculum (including opportunity to learn and the classroom learning environment), and learned curriculum while accounting for student characteristics.

Method

Research Design

The research reported herein is the second in a series of studies conducted by the COSMIC project. It capitalizes on the strengths of the initial study of curricular effectiveness in first-year high school mathematics courses (Grouws et al., 2013) to examine relationships between mathematics content organization, curriculum implementation, and students' learning in second-year high school mathematics courses. Our conceptual framework for curriculum research (Figure 1) and the aforementioned literature informed the design of this study in several important ways. First, it is clear that studies of curricular effectiveness must account for student characteristics (e.g., prior achievement, gender, race/ethnicity, special

[1] We did not measure student motivation or account for support at home, but instead assumed these to be roughly comparable across the two groups of students.

education status, English Language Learners) that have demonstrative predictive power and act as forces on the learned curriculum. Secondly, curriculum research should account for teacher characteristics (e.g., experience, beliefs, professional development) that have the potential to offer explanatory power and influence the implemented curriculum. Third, it is essential to characterize teachers' classroom use of curricular materials—vis-à-vis fidelity of implementation, opportunity to learn and the classroom learning environment—to ascertain the effects of the textbook curriculum on student learning (learned curriculum). Consistent with recommendations of the National Research Council (2004), multiple measures of student outcomes were used. Finally, we attended to the WWC evidence standards by adopting a quasi-experimental design with equating. Randomized controlled trials were not possible because students were not randomly assigned to classes and, furthermore, classes were not randomly assigned to teachers. Accordingly, discretion should be exercised in drawing causal inferences between the textbook curriculum and learned curriculum.

Sample Selection: Schools, Teachers, and Students

As reported in Grouws et al. (2013), the COSMIC project conducted an extensive search for U.S. high schools that offered parallel curricular paths, integrated and subject-specific, and we further narrowed our search to schools using Core-Plus as the integrated curriculum. Furthermore, schools were ineligible to participate if students were tracked into one curriculum type; therefore, schools were excluded if a particular curriculum was designated for high (or low) achieving students. It was our position that schools offering parallel curricular paths inherently value each curriculum as a viable secondary mathematics program. Furthermore, by selecting schools that offer parallel curricular paths, we were able to provide some control for a variety of contextual factors, including the length of the academic year, duration of the school day, length of class periods, district-mandated professional development, homework and technology policies, assessment practices, student demographics, and school climate. Additionally, we sought diversity across schools with respect to geography, race/ethnicity, and socioeconomic levels. Eligible schools were invited to participate if they also agreed to provide student records (including scores on state tests, free or reduced lunch status, and race/ethnicity), permit multiple observations of teachers, and commit to the administration of three project-related assessments.

Our selection process yielded 11 high schools in six school districts that were located in five geographically dispersed states, none of which were in an initial adoption year for either of their mathematics textbook series. Our participant recruitment process was not independent of that employed in Grouws et al. (2013) because most students from the first-year study participated in the second year of the COSMIC project. In this study, those teachers responsible for teaching Geometry or Course 2 of Core-Plus were invited to participate, and all agreed. By virtue of their enrollment in Geometry or Course 2, all such students participated in the study.

Characterization of the Study Curricula

Because our study's primary aim was to examine whether there is a differential effect of content organization on student learning, we classified textbooks into two categories, *integrated* and *subject-specific*. All teachers of the integrated curriculum used Core-Plus Course 2 (Coxford et al., 2003), whereas teachers of the subject-specific curriculum used high school geometry textbooks from several different publishers, including Glencoe-McGraw Hill (Boyd, Cummins, Malloy, Carter, & Flores, 2005), Prentice Hall (Bass, Charles, Jonson, & Kennedy, 2004), Holt (Burger et al., 2007), and McDougal Littell (Larson, Boswell, & Stiff, 2001). Although there are modest differences between the geometry textbooks, all of them covered the same core topics and were organized similarly. For example, lessons in subject-specific textbooks include a lesson opener, an explanation of key ideas, definitions and worked-out examples, and opportunities for guided practice and conclude with practice problems that are similar to the worked examples. Teachers are expected to bring closure to individual lessons in a single class period. In contrast, the Core-Plus integrated textbook includes multiple mathematical strands including geometry, algebra, discrete mathematics, statistics, and modeling. Spanning several days, lessons in this textbook are generally situated in a real-world context and begin with a launch. Students typically work in collaborative groups (often using graphing calculators) during the explore lesson component, then share and summarize the mathematical ideas before solving problems in an apply homework section. Formalism such as definitions, symbolic representations, and mathematical theorems are developed when content is revisited later in the same course or in subsequent courses.

Not all integrated textbooks are identical because they are written using different design principles, and likewise some subject-specific textbooks may have a pedagogical orientation that is similar to some integrated textbooks. For example, the Center for Mathematics Education project (2009) includes topics beyond a singular content focus and is organized around mathematical habits of mind (see Cuoco, Goldenberg, & Mark, 2010), and some integrated textbooks series, such as the Interactive Mathematics Program, do not share the same pedagogical orientation of Core-Plus. Because of these subtle yet possibly important distinctions, we caution against drawing inferences regarding curricular effectiveness of textbooks not included in our sample.

Teacher Characteristics and Measures of Curriculum Implementation

Using a variety of data sources, we gathered information regarding teacher characteristics, teaching practices, and the implementation of curricular materials in mathematics classrooms.

Teacher surveys. On an Initial Teacher Survey (ITS) administered prior to the school year, teachers provided background and demographic data such as years of teaching experience and quantity of professional development and its perceived

effect on instructional practices, and on a Likert scale indicated their beliefs about teaching and learning. On a Mid-Course Teacher Survey (MTS) administered in the late fall, they reported the number of years teaching the textbook curriculum, and indicated on a Likert scale their preparedness to teach the textbook curriculum as well as their overall satisfaction with the textbook (Appendix A).

Table of Contents (TOC) records. As a measurement of fidelity of implementation, on a daily basis, teachers indicated for each lesson (referred to as an *investigation* in the integrated curriculum) one of the following codes: (a) content taught primarily from textbook, (b) content taught from the textbook with some supplementation, (c) content taught primarily from an alternative source, and (d) content not taught. We collected TOC records quarterly. Based on codes for every lesson in the table of contents, we developed three indices to capture the nature and extent of textbook use: Opportunity to Learn (OTL), Extent of Textbook Implementation (ETI), and Textbook Content Taught (TCT). Essentially, the OTL Index represents the percentage of textbook lessons taught without considering teachers' use of supplemental or alternative curricular materials. In contrast, the ETI Index was determined by assigning weights according to the strength of the relationship between the textbook curriculum and implemented curriculum. Lessons taught directly from the textbook were weighted 1, those taught with some supplementation were weighted 2/3, and textbook lessons taught from alternative sources were weighted 1/3; textbook content not taught was weighted 0. The ETI Index was calculated by summing the weights across the textbook lessons, dividing by the number of lessons contained in the textbook, and multiplying by 100. The TCT Index differs from the ETI Index by considering only those lessons where content was taught in some manner, thereby ignoring content that was not taught. Lessons were weighted in the same manner as in the ETI, but the index was calculated by dividing by the number of lessons reported as being taught in any manner and again multiplied by 100. For a full description of the development of these indices, see Tarr, McNaught, and Grouws (2012). Individually, these indices measure subtle but important aspects of curriculum implementation. Collectively, they offer utility in characterizing the extent to which teachers rely on the textbook to select the content.

Classroom Visit Protocols (CVP). We interviewed textbook authors and asked them to describe specific observable teacher behaviors they would expect to see in a faithful implementation of their curriculum materials. Our questions focused on particular structural components of integrated textbook lessons (i.e., launch, explore, share and summarize, and apply) and subject-specific textbook lessons (i.e., lesson preview, teach, and practice and apply). From our interviews, we developed curriculum-specific observation protocols and accompanying user's guides. During three classroom visits (fall, winter, and spring), trained observers documented classroom activities and use of curricular materials in each class. Essentially, observers either made dichotomous judgments—"Did I observe this action?" or "Did I not observe this action?"—or assigned scores on a 5-point Likert

scale for *content fidelity* (content taught) and *presentation fidelity* (pedagogical considerations). Observers also characterized selected elements of the classroom learning environment by rendering scores for 10 items classified into three subscales: *reasoning about mathematics, students' thinking in instruction*, and *focus on sense making*.

Through field-testing and extensive training of classroom observers, we achieved a high degree of coding reliability. More specifically, for a sample of 15 double-coded lessons, overall agreement was attained on approximately 94% of codes. Agreement was achieved on 14 of 15 (or 93%) rating pairs for content fidelity and 10 of 15 (or 67%) rating pairs for presentation fidelity, with the remaining 5 pairs all within one unit of each other. With respect to the classroom learning environment, agreement was realized in approximately two thirds of codes, but more than 92% of codes were within one unit of each other on a 5-point scale. For a more detailed description of the development of classroom visit protocols, see Tarr et al. (2012).

Student Assessments

Prior achievement measure. We used existing scores on state-mandated Grade 8 tests, typically administered during the 2004–2005 school year, as a viable measure of students' mathematics achievement prior to participation in the study. Our multistate sample necessitated the transformation of prior achievement scores to a common scale that would take into account differences in student achievement across states. To that end, we mapped individual scores on state assessments to the 2005 National Assessment of Educational Progress (NAEP) scale. More specifically, we converted students' scores in each state to z-scores before mapping these onto the NAEP scale (see National Center for Education Statistics [NCES], 2007). We called the resulting score the COSMIC Prior Achievement Score (CPA Score). For additional illustrative examples and a full description of the scaling process, see Chávez, Papick, Ross, and Grouws (2011) and Tarr et al. (2010).

Project-developed tests. To inform the development of student assessments, a research mathematician, mathematics teacher educator, and Ph.D. student in mathematics education (with a graduate degree in mathematics) worked collaboratively to analyze the content of the textbooks used by the teachers in our sample. We examined Core-Plus Course 2 as the integrated textbook and Glencoe's *Geometry* as the representative subject-specific textbook because it was most commonly used across participating schools. Informed by the content analysis, we developed two student assessments. A detailed description of the content analysis process and test development is presented in Chávez et al. (2011).

***Test of Common Objectives* (*Test C*).**[2] We developed an assessment that included only objectives common to both curriculum types, including geometry of the

[2] We named tests sequentially throughout the COSMIC project. Test A and Test B were used in the Year 1 study (Grouws et al., 2013).

Cartesian coordinate system with concepts related to lines, distance, midpoints, and slope; transformations in the coordinate plane; perimeter, area, volume, and surface area; proportionality and similarity; and trigonometric ratios. Given the cumulative nature of school mathematics, we also included some common content found in the textbook of the previous year in each series (Algebra 1 and Core-Plus Course 1), including systems of linear equations, graphs of linear equations, quadratic equations, radicals and rational exponents, the study of some basic polynomial functions, three-dimensional figures and nets, symmetry and transformations, Pythagorean theorem, and quadrilaterals. The Test of Common Objectives (Test C) included nine constructed-response rubric-scored items and one multiple-choice item. Two of the 10 items were subdivided, for a total of 14 subitems. Our scoring rubrics included separate scores for the work shown in each subitem and for the solution given, yielding 28 scorable units. Inter-rater reliability of the scoring of the Test of Common Objectives was 97.3%.

Test of Problem Solving and Reasoning (Test D). The Test of Problem Solving and Reasoning (Test D) included five constructed-response rubric-scored items, four focused on geometry and one on algebra. These items required mathematical reasoning and were based on topics appropriate to the grade level. Several items were subdivided, for a total of 11 subitems. Our scoring rubrics considered separate scores for work and answer; thus, the Test of Problem Solving and Reasoning had 18 scorable units. Inter-rater reliability of the scoring of the Test of Problem Solving and Reasoning was 94%.

With respect to the use of technology, we developed test items such that student access to calculators would not have a direct influence on the quality of an item response. Although no items *required* the use of calculators for successful attainment of an answer, calculator use was *permitted* on all items.

From raw scores on each of the project-developed tests, Test of Common Objectives and Test of Problem Solving and Reasoning, scale scores for every student were generated using two-parameter Item Response Theory (IRT); item difficulty and item discrimination indices were used to generate theta scores for each student on each test.[3] Ancillary analyses of student responses were undertaken in order to verify and ensure local item independence, a fundamental assumption of traditional IRT modeling. Because of the large size of the student sample, sufficient number of test items, and sound psychometric properties of the assessments, our models converged, and we conducted reliability checks to ensure proper model fitting.

Iowa Test of Educational Development. We attended to the NRC (2004) recommendations that studies of curricular effectiveness use multiple measures of student learning. In addition to the two project-developed tests, we selected the Iowa Test

[3] Despite an abundance of statistical power, the generation of IRT subscales (e.g., geometry subscale) was not possible.

of Educational Development: Mathematics: Concepts and Problem Solving Form B (hereafter, ITED-16; Feldt, Forsyth, Ansley, & Alnot, 2003) as a standardized achievement test. According to the publisher's website,

> the primary intent of the Mathematics: Concepts and Problem Solving test is to measure students' ability to solve quantitative problems. The questions in this test present practical problems that require basic arithmetic and measurement, estimation, data interpretation, and logical thinking. Since this is a test of the students' ability to use appropriate mathematical reasoning, the number of items requiring computation is minimal. (Riverside Publishing, n.d.)

Although the ITED-16 measures students' problem-solving skills, it differs from the Test of Problem Solving and Reasoning in important ways. The ITED-16 contains many more items (40) than the Test of Problem Solving and Reasoning (5) and uses multiple-choice questions, whereas the latter uses only constructed-response items. Notwithstanding these differences, the correlation between the scores on these two measures was approximately .70. Classroom teachers administered our three measures of student outcomes during the regular class day near the close of the school year.

Sample Description and Equivalence of Treatment Groups

Student sample. The 3,258 students in the sample were spread across the six school districts, ranging from a low of 176 students in District C to a high of 1,176 students in District W, with slightly more females (50.4%) participating than males (49.6%; see Table 1). White students comprised three fourths of the total sample but more than 90% in District B, District C, and District T. Hispanic students comprised about one ninth of the total sample but ranged from a low of 3.2% in District T to a high of 41.2% in District R. African American students accounted for 9% of the total sample, with none in District C and 19.1% in District W. The prevalence of student classifications such as Individualized Education Program (IEP), Limited English Proficiency (LEP), and Free or Reduced Lunch (FRL) varied across districts. District I had the greatest percentage of students qualifying for FRL, 50.8%, and the lowest number of students classified as LEP, 0%. In contrast, District R had the lowest percentage of FRL, 17.4%, and the highest percentage of students classified as LEP, 5.2%. Students with an IEP accounted for 6.6% of the total sample and ranged from 2.7% in District W to 12.8% in District I.

It is important to note that we aggregated the percentage of FRL students for each class primarily because one district restricted access to FRL status for individual students but was willing to share the percentage of FRL students at the class level. This class-level data was further aggregated across sections taught by each teacher and was used as a proxy for SES in multilevel modeling of student outcomes. Moreover, class-level FRL percentages were further aggregated and used as a school-level variable in our multilevel modeling of student outcomes.

Although we examined students' mathematics learning in second-year high school courses, we acknowledge that not every student enrolled in Geometry or Course 2 was in his or her second year of high school. In the subject-specific

Table 1
Characteristics of Student Sample, by School District

		Gender		Race/Ethnicity				Qualifications		
District	Students	Male	Female	AfrAm	Hispanic	White	Other[a]	IEP	LEP	FRL
B	250	47.6	52.4	2.0	6.8	90.0	2.4	8.0	2.4	26.8
C	176	46.0	54.0	0.0	5.1	94.9	0.0	5.1	3.4	43.2
I	390	50.0	50.0	9.2	6.9	80.5	3.3	12.8	0.0	50.8
R	507	46.7	53.1	4.1	41.2	47.5	7.1	7.3	5.2	17.4
T	759	50.9	49.1	1.2	3.2	92.1	3.6	9.0	1.3	25.0
W	1,176	50.8	49.2	19.1	7.1	67.9	5.9	2.7	1.8	19.0
Totals	**3,258**	**49.6**	**50.4**	**9.0**	**11.4**	**75.0**	**4.6**	**6.6**	**2.1**	**25.8**

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

curriculum, Grade 9 students comprised 21.4% of the sample, Grade 10 students accounted for 71.3%, and high school juniors and seniors comprised 6.1% and 1.2%, respectively. For the integrated curriculum, a similar distribution of grade levels was evident, with 24.2%, 68.7%, 6.2%, and 0.6% of students in Grades 9, 10, 11, and 12, respectively. A one-way analysis of variance (ANOVA) detected no significant differences in the distribution of grade levels across curriculum types.[4]

COSMIC Prior Achievement (CPA) scale scores appear approximately normally distributed (Figure 2) with a mean (289.88) about one fourth of a standard deviation greater than the national NAEP mean (278.82). An ANOVA showed the mean CPA for students in the subject-specific pathway (291.68) was significantly higher ($F = 19.68$, $p < .001$) than the mean CPA for students in the integrated pathway (286.75). Given the large sample size, the detection of a statistically significant difference is not surprising. However, the 4.93-point difference in group means has little practical significance because it represents only one sixth of a standard deviation. Nevertheless, in developing models of student outcomes, we controlled for CPA as well as numerous other student characteristics that have potential predictive power.

As shown in Table 2, more students were in the subject-specific (64%) than the integrated curriculum (36%) pathway. Students enrolled the most often into the integrated curriculum in District I, which is in contrast with other districts. The difference in student enrollment between the two curriculums was greatest in

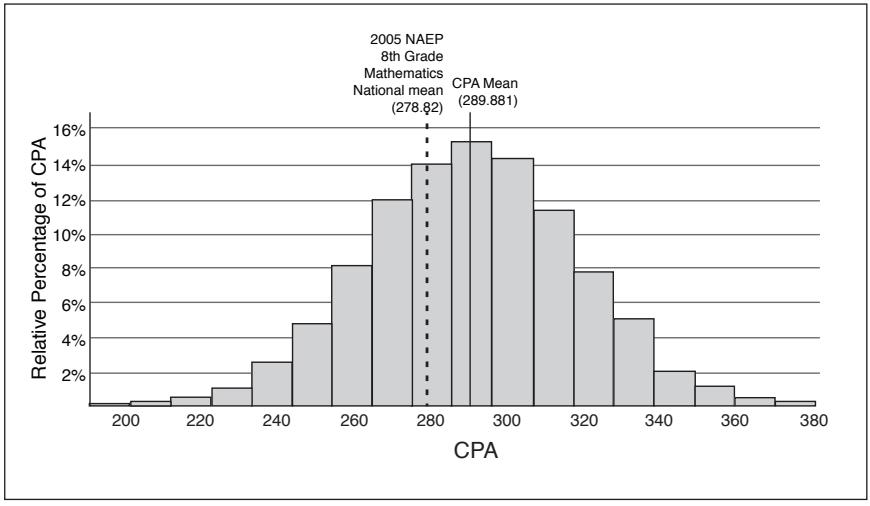

Figure 2. Distribution of COSMIC Prior Achievement (CPA) scores.

[4] We detected a significant negative correlation ($r = –.352$, $p < .001$) between grade level and prior achievement; students from lower grade levels were associated with higher COSMIC Prior Achievement scores. However, rather than using grade level as a student variable, we instead use the CPA scale score in analyses of student outcomes; the former is an ordinal categorical variable, whereas the latter is a numerical variable on an interval scale that offers more precision.

District W, where there were 660 more students enrolled in subject-specific than in integrated curriculum classes.

Table 2
Class and Student Sample, by School District and Curriculum Type

	Classes		Students	
District	Integrated	Subject-specific	Integrated	Subject-specific
B	2	2	109	141
C	1	2	82	94
I	7	2	331	59
R	2	5	61	446
T	8	4	330	429
W	7	8	258	918
Total	**27**	**37**	**1,171**	**2,087**

Because students had the opportunity to choose between curricular paths, it was conceivable that specific attributes might be more common among students who selected one path over another. To investigate this possibility, we used logistic regression to generate a *propensity score* for each student that represents the probability that he or she was assigned to a particular curriculum type based on prior achievement (CPA), gender, ethnicity, IEP status, and LEP status.[5] We found African American and Hispanic students were more likely to be in the subject-specific curriculum group, as were IEP students. Our analysis further revealed that in practice CPA was a significant predictor of curriculum type, with higher performing students more likely in the subject-specific path. However, ancillary analysis identified that differences in CPA across curriculum types occurred in only two schools, District C, a small rural school, and District I, an urban school. In both schools, students with higher prior achievement more frequently selected the subject-specific curricular option. Significant differences in mean CPA scores across curriculum types differ from the Year 1 study (Grouws et al., 2013), which examined relationships between content organization, curriculum implementation, and student learning in Algebra 1 and integrated Course 1 classrooms.

The student sample was restricted to those with complete data. Student mobility, incomplete school records, and additional factors resulted in missing data for 593 students (18.2%). For the vast majority of these students, a prior achievement score was simply not available. Because prior learning is arguably the best predictor of

[5] Our use of propensity scores controls for *measurable* student characteristics only. It is possible, although unlikely, that there were unmeasured differences between the two groups of students that might have influenced their choice of curriculum type and therefore could affect outcomes in unknown ways, and we acknowledge this as a potential limitation of our study.

student achievement, we decided to exclude such students from analyses of outcome measures; for many reasons, imputation of prior achievement scores is problematic in our research design. Moreover, because we developed separate models for each of our three dependent measures, the sample sizes for each model varied slightly. Although missing data was prevalent in the subject-specific (22.3%) and the integrated (11.3%) sample, excluded student cases from each group had similar demographics. Students with complete background data participated in summative assessments at high rates, ranging from 89.0% to 96.2% on the Test of Common Objectives (Test C), 87.8% to 97.2% on the Test of Problem Solving and Reasoning (Test D), and 90.8% to 97.6% on the ITED-16. Overall, approximately 84% of students took all three tests, whereas only about 2% took none.

Class sample. We collected self-report data from all 64 study teachers using an Initial Teacher Survey (ITS) and Mid-Course Teacher Survey (MTS). When the same teacher taught several sections of a course, these sections were combined to form a class. The sample included 37 subject-specific teachers and 27 integrated teachers with four teachers of both curricula. We conducted an ANOVA comparing teachers of the two curricula and found no significant differences on almost all measured characteristics, beliefs, and practices. Both groups of teachers possessed approximately the same number of years of experience ($M = 10.77$, $SD = 8.70$), familiarity with the textbook, and amount of professional development. With respect to beliefs, there were no significant differences with respect to didactic approaches to teaching mathematics and students' self-efficacy. Furthermore, both groups shared similar knowledge of NCTM Standards and their perceived implementation of Standards. Although the groups were largely equivalent on most characteristics, a few differences were detected. Teachers of the integrated curriculum held beliefs about reform-oriented practices that differed markedly from subject-specific teachers ($F = 11.91$, $p = .001$), and their declared agreement with the vision of the NCTM Standards was greater than that of teachers of subject-specific curricula ($F = 7.08$, $p = .010$). We found no other significant differences.

Reduction of Teacher Characteristics and Curriculum Implementation Data

We utilized Principal Components Analysis (PCA) to reduce the large number of variables in our teacher database. For conceptual reasons, a few class-level variables were excluded from the PCA, such as Curriculum, Free or Reduced Lunch (FRL) percentage, and the percentage of instructional time devoted to lesson development (Time_LD).[6] The initial extraction was conducted with 27 variables and 151 cases in the entire COSMIC project; each case represented one class; that is, one teacher of a given curriculum type in a given year.[7] Three variables were subsequently

[6] Although not included in the Principal Components Analysis, all of these variables were used in constructing models of student outcomes.

[7] An integrated class and a subject-specific class taught by the same teacher are considered to be two separate cases.

excluded because of low communalities. Extracted factors were rotated using the Varimax method, converging in seven iterations and explaining about 70% of the variance. Using the Anderson-Rubin technique, we generated factor scores, each with a mean of approximately 0 and standard deviation of about 1. The seven factors essentially clustered around two themes, curriculum implementation and teacher characteristics (Figure 3). Related to curriculum implementation, we named the four factors Classroom Learning Environment (CLE), Implementation Fidelity (FID), Technology (TECHNOLOGY), and Opportunity to Learn (OTL).[8] Related to teacher characteristics, we named the three factors NCTM Standards (STANDARDS), Experience (EXPERIENCE),[9] and Professional Development (PD). Each factor is distinctive because the individual variables that load most heavily are clearly related conceptually. Our factor analysis reduced the expansive class-level data (e.g., seating, collaboration) into a coherent data set by addressing the inherent interdependencies related to curriculum implementation as well as detecting variables that were tenuous or deficient of key measurement properties.

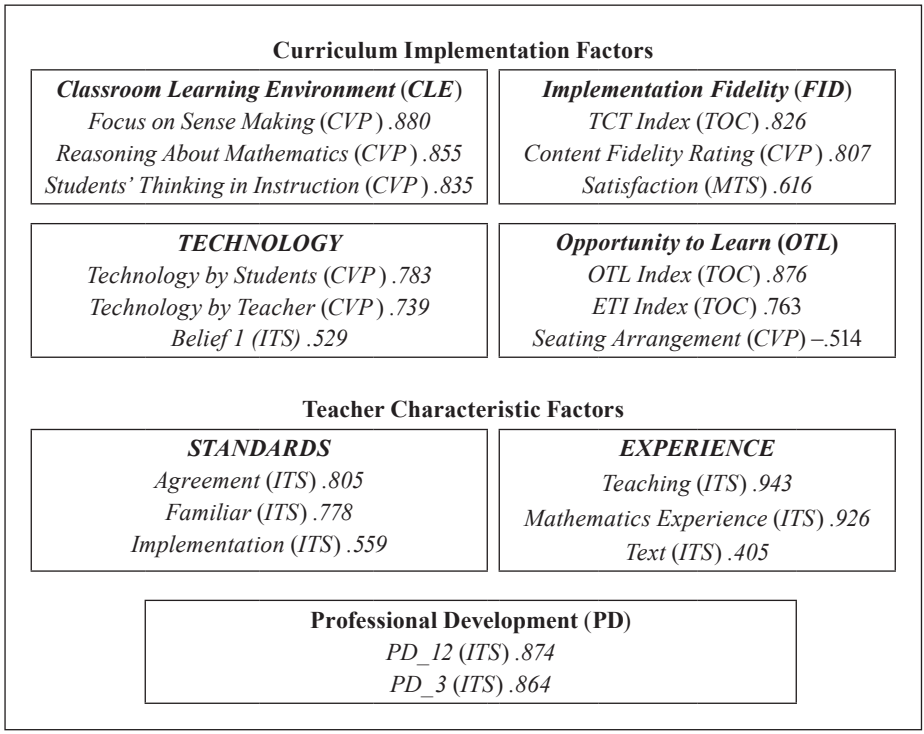

Figure 3. Factors related to curriculum implementation and teacher characteristics, data source, and primary loadings.

[8] The OTL factor described here is different from the OTL Index. The OTL Index is one of several class-level variables that loaded heavily on the OTL factor, as depicted in Figure 3.

Results

Curriculum Implementation in Mathematics Classrooms

Across curriculum types, there were several notable differences in implementation as measured by three implementation indices (Table 3).[10] With respect to the Opportunity to Learn (OTL) Index, on average about three fourths of textbook lessons were taught. Teachers of the subject-specific curriculum taught approximately 86% of textbook lessons, whereas teachers of the integrated curriculum taught only 58%, on average. In an extreme case, one teacher[11] reported teaching none of the content in Course 2 but instead spent the academic year teaching Course 1 content she did not cover the previous year. Taking into account the degree to which teachers used study curricula in teaching mathematics, the Extent of Textbook Implementation (ETI) Index suggested that teachers occasionally supplemented their textbooks but seldom used alternative materials. The mean OTL Index for teachers of the subject-specific curriculum was significantly greater ($F = 52.47$, $p < .001$) than that for teachers of the integrated curriculum, and the same is true of the mean ETI Index ($F = 15.24$, $p < .001$). Notwithstanding these differences, when teachers taught textbook content, they primarily taught directly from their textbooks, as indicated by the large Textbook Content Taught (TCT) Index values.

Table 3
Table of Content Records: Means and Standard Deviations of Implementation Variables by Curriculum Type

			Teachers					
			Subject-specific[a]		Integrated[b]		All[c]	
Index	Min	Max	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)
OTL	0.0	130.0[d]	37	86.3 (16.0)	27	57.6 (15.2)	64	74.2 (21.1)
ETI	0.0	122.5[d]	37	61.1 (16.8)	27	45.4 (14.6)	64	54.5 (17.6)
TCT	34.6	100.0	37	71.5 (15.5)	27	79.8 (14.8)	64	75.0 (15.6)

[a]$n = 37$. [b]$n = 27$. [c]$n = 64$. [d]Indices are measured on a 0–100 scale. Indices exceeding 100 represent accelerated courses in which more than the entire book was taught in an academic year.

[9] Rather than using factor scale scores, we partitioned the teacher sample into two categories, thereby rendering *teaching experience* a dichotomous variable. Experienced teachers ($n = 52$) had taught for at least 3 years prior to the study, whereas inexperienced teachers ($n = 12$) had taught for fewer than 3 years.

[10] It is important to note that variables in Tables 3 and 4 were not examined individually in relation to student outcome measures but instead were ultimately reduced to cluster factors using principal components analysis. Factor scores were related to student outcomes in the subsequent development of multilevel models.

[11] For each dependent variable, we constructed models that *included* and *excluded* this teacher-case. Because both models yielded the same set of significant predictors, our final models are based on inclusion of this teacher-case.

In fact, teachers exhibited significantly greater reliance on the integrated textbook ($F = 4.71$, $p = .034$) than the subject-specific textbook. Therefore, although teachers of the integrated curriculum generally provided fewer opportunities to learn textbook content, they nevertheless closely followed their textbook when teaching mathematics. For each index, there was substantial variation across teachers regardless of curriculum type.

Based on 186 classroom observations, we found relatively high mean content fidelity ratings for both sets of teachers, suggesting that teachers largely implemented the content of textbook lessons with considerable integrity (Table 4). However, mean presentation fidelity ratings were somewhat lower, indicating that teachers were less faithful to the pedagogical recommendations espoused by textbook authors, and this was more evident among teachers of the integrated program. Teachers of the integrated curriculum were significantly more likely to use graphing calculators during instruction, although they did so in only 19% of lessons on average. Similarly, students in the integrated pathway were much more likely to use graphing calculators than those in classes using the subject-specific curriculum; they also sat in groups and worked collaboratively more often. Across curriculum types, students were engaged (i.e., on task) in comparable proportions of lessons and spent relatively similar amounts of class time on various aspects of textbook lessons, with the exception of lesson closure (Table 4).

With respect to the Classroom Learning Environment subscales, relatively low means were observed across the teacher sample, ranging from 2.12 (reasoning about mathematics) to 2.50 (students' thinking in instruction). Indeed, none of the 64 teachers achieved a rating of 5 on the 5-point Likert scale used for the three subscales. Across curriculum types, we found significant differences between curriculum types, as teachers of the integrated curriculum scored higher ($F = 7.305$, $p = .009$) on the Reasoning subscale than did teachers of the subject-specific curriculum. No differences were detected on the students' thinking in instruction or focus on sense making subscales. Collectively, the data in Tables 3 and 4 convey substantial variability in all measures of curriculum implementation, both within groups and between groups. These measurement properties are essential to the examination of relationships between curriculum, implementation fidelity, and student outcomes.

Student Learning

Hierarchical linear modeling. We examined relationships between curriculum type, implementation measures, and student characteristics vis-à-vis three end-of-course measures of student learning. For each student outcome measure, we fit a three-level hierarchical linear model (HLM), taking into account the nested data structure of students within classes and classes within schools.

Level 1: Student-level model. At the student level (Level 1), our models included prior achievement (CPA) and numerous demographic data (e.g., gender, race/ethnicity) that have demonstrated predictive power in related studies of curricular effectiveness (e.g., Harwell et al., 2007; Klibanoff, Levine, Huttenlocher, Vasilyeva, & Hedges,

Table 4
Classroom Visit Protocol (CVP): Means of Implementation Variables by Curriculum Type

	Classroom teachers				
			Subject-specific[a]	Integrated[b]	All[c]
	Min	Max	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Lesson Fidelity Rating[d]					
Presentation	1	5	3.26 (0.88)	2.90 (1.18)	3.11 (1.03)
Content	1	5	3.57 (1.02)	3.77 (1.13)	3.65 (1.07)
Technology Use[e]					
By Teacher	0	100	5 (28)	19 (32)	11 (26)
By Students	0	100	24 (31)	66 (30)	42 (37)
Classroom Learning Environment[d]					
Reasoning About Mathematics	1	4	1.86 (0.75)	2.48 (1.11)	2.12 (0.96)
Students' Thinking in Instruction	1	4	2.39 (0.80)	2.66 (0.92)	2.50 (0.86)
Focus on Sense Making	1	4	2.30 (0.73)	2.60 (0.84)	2.42 (0.79)
Lesson Characteristics					
Closure[e]	0	100	14 (19)	28 (34)	20 (27)
Engage[f]	0	2	1.62 (0.55)	1.63 (0.63)	1.63 (0.58)
Seating[f]	0	2	0.38 (0.72)	1.44 (0.89)	0.83 (0.95)
Collaboration[f]	0	2	0.16 (0.37)	0.93 (0.62)	0.48 (0.62)
Time Allocation[e]					
Lesson Development	10	97.2	42 (18)	48 (25)	45 (21)
Non-instructional	0	29.5	8 (6)	8 (8)	8 (7)
Practice and Apply	0	100	19 (16)	14 (17)	16 (16)

Notes. [a]$n = 37$. [b]$n = 27$. [c]$n = 64$. [d]On a 5-point Likert scale. [e]Reported as a percentage. [f]Interval scale, 0–2.

2006; Post et al., 2008, Staub & Stern, 2002). For each dependent measure, the conditional model equation for the student level (Level 1) was as follows:

$$\begin{aligned}\text{TEST}_{ijk} = {} & \pi_{0jk} + \pi_{1jk} * (CPA)_{ijk} + \pi_{2jk} * (FEMALE)_{ijk} \\ & + \pi_{3jk} * (AFRICAN\ AMERICAN)_{ijk} \\ & + \pi_{4jk} * (HISPANIC)_{ijk} + \pi_{5jk} * (OTHER)_{ijk} \\ & + \pi_{6jk} * (IEP)_{ijk} + e_{ijk},\end{aligned}$$

where TEST_{ijk} is the IRT scaled score of student *i* of teacher *j* in school *k*; CPA is a prior achievement score; and AFRICAN AMERICAN, HISPANIC, and OTHER are dichotomous 0/1 indicators of ethnicity with White ethnicity serving as the referent. Similarly, FEMALE is a 0/1 indicator for gender, and IEP is a 0/1 indicator for a student having an Individualized Education Program. The π_{ijk} are Level-1 regression parameters and e_{ijk} the Level-1 random error component. With the variable coding used here, π_{0jk} is the expected test score of teacher *j* in school *k* for a White, male student without an IEP and with prior achievement score equal to the grand mean CPA. It follows that the coefficients π_{1jk} through π_{6jk} reflect differences from this "reference student" due to student characteristics.

Level 2: Class-level model. At the class level (Level 2), our models included a dichotomous indicator for curriculum type (Curriculum)—0 for subject-specific, 1 for integrated—and five additional class-level variables with potential capacity to explain variation in student outcomes: socioeconomic status (FRL used as a proxy), Classroom Learning Environment factor (CLE), Implementation Fidelity factor (FID), Opportunity to Learn factor (OTL), Technology factor (TECHNOLOGY), NCTM Standards factor (STANDARDS), time devoted to lesson development variable (TIME_LD), teaching experience variable (Experience), and Professional Development factor (PD).[12] In addition to these main effects, our models include the two-way interactions of curriculum type with the CLE, FID, and OTL factors, and the FRL variable. The conditional class-level model is given by the following equation:

$$\begin{aligned}\pi_{0jk} = {} & \beta_{00k} + \beta_{01k} * (Curriculum)_{jk} + \beta_{02k} * (FRL)_{jk} + \beta_{03k} * (CLE)_{jk} \\ & + \beta_{04k} * (FID)_{jk} + \beta_{05k} * (OTL)_{jk} + \beta_{06k} * TECHNOLOGY_{jk} \\ & + \beta_{07k} * STANDARDS_{jk} + \beta_{08k} * TIME\_LD_{jk} + \beta_{09k} * Experience_{jk} \\ & + \beta_{010k} * PD_{jk} + \beta_{011k} * (Curriculum \times CLE)_{jk} \\ & + \beta_{012k} * (Curriculum \times FID)_{jk} + \beta_{013k} * (Curriculum \times OTL)_{jk} \\ & + \beta_{014k} * (Curriculum \times FRL)_{jk} + r_{0jk}, \\ \pi_{1jk} = {} & \beta_{10k} + \beta_{11k} * (Curriculum)_{jk}\end{aligned}$$

[12] Variables and indices that loaded on each of these factors are summarized in Figure 3.

where r_{0jk} is a random effect varying over teachers. The intercept term is the mean test score for teacher k in school j prior to adjustment for the other Level-2 factors reflected in the coefficients β_{01k} through β_{010k}. The inclusion of π_{1jk} results in a cross-level interaction between curriculum type (Curriculum) and prior learning (CPA).

Level 3: School-level model. At the school level (Level 3), our model follows and includes a measure of SES formed as the average of the Level-2 FRL variables: $\beta_{00k} = \gamma_{000} + \gamma_{001}(SCHFRL)_k + u_{00k}$, where β_{00k} is the mean test score in school k, γ_{000} is the grand mean over all schools, γ_{001} is the school-k deviation from the mean, and u_{00k} is the Level-3 random effect. Combining the individual models gives the following mixed model expression for the full regression model used for each of the three test scores:

$$
\begin{aligned}
TEST_{ijk} = \gamma_{000} &+ \gamma_{001} * SCHFRL_k + \gamma_{010} * Curriculum_{jk} + \gamma_{020} * FRL_{jk} \\
&+ \gamma_{030} * CLE_{jk} + \gamma_{040} * FID_{jk} + \gamma_{050} * OTL_{jk} \\
&+ \gamma_{060} * (Curriculum \times CLE)_{jk} + \gamma_{070} * (Curriculum \times FID)_{jk} \\
&+ \gamma_{080} * (Curriculum \times OTL)_{jk} + \gamma_{090} * (Curriculum \times FRL)_{jk} \\
&+ \gamma_{010} * TECHNOLOGY_{jk} + \gamma_{011} * STANDARDS_{jk} \\
&+ \gamma_{012} * TIME\_LD_{jk} + \gamma_{013} * Experience_{jk} + \gamma_{014} * PD_{jk} \\
&+ \gamma_{100} * CPA_{ijk} + \gamma_{110} * CPA_{ijk} \times Curriculum_{jk} + \gamma_{200} * FEMALE_{ijk} \\
&+ \gamma_{300} * AFRICAN\ AMERICAN_{ijk} + \gamma_{400} * HISPANIC_{ijk} \\
&+ \gamma_{500} * OTHER_{ijk} + \gamma_{600} * IEP_{ijk} + r_{0jk} + u_{00k} + e_{ijk}
\end{aligned}
$$

All variables, except dichotomous variables and interactions, were centered on the grand mean. All interactions were computed on mean-centered transformations of the variables involved. Effect sizes (g) were computed for significant main effects by dividing the individual predictor beta coefficient by the pooled post-test standard deviation. Significant effects were taken to be those with p-values less than or equal to .05.

For each measure, we provide three models: (a) the unconditional model, (b) a model fitted to research and control variables, and (c) a parsimonious model containing only significant variables and curriculum type. For each model, we report the distribution of variance across each of the three levels.

Modeling Test of Common Objectives (Test C) student scores. In Table 5, we report our model of student scores on Test C. Regarding the research question related to content organization, we detected no significant main effect of mathematics curriculum type (p = .990). Students from the two curriculum pathways performed nearly identically on Test C. Similarly, FID was not associated with student performance in any meaningful way (p = .861).

Several Level-1 student covariates were significantly associated with scores on Test C, most notably prior achievement. Although the magnitude of the slope

(0.195) appears to be small, CPA was the best predictor of performance ($p = .001$) with an associated effect size (g) equal to 0.548. Female students scored 0.606 points higher than males, and this coefficient is significant ($p = .030$). Neither African American nor special needs (IEP) exhibited significant predictive power, and although Hispanic students scored about 1 point lower, this result is of marginal significance ($p = .054$). Compared to White students, those classified as other races scored lower ($p = .034$), although its significance was lost in the reduced model.

At Level 2, there was one significant main effect, OTL, and a significant interaction effect, Curriculum × OTL. In particular, as OTL scores increased, there was a corresponding significant increase ($p < .001$) in student performance with an effect size (g) of 0.264. Interestingly, there was a differential effect of OTL across curriculum types. For classes in the subject-specific curriculum, each incremental increase in OTL scores resulted in a 2.656-point increase in student scores. However, the positive effect of OTL on student performance was greatly diminished for classes in the integrated curriculum. An increase of one standard deviation in OTL resulted in a 0.316-point increase in student performance, markedly lower than the slope of 2.656 for the entire sample of classes. No additional significant interactions with curriculum were determined at the class level. Other non-significant effects include Experience, FRL, Time_LD, and class factors of TECHNOLOGY, STANDARDS, and PD. Although the CLE failed to demonstrate predictive power in the full model, it emerged as a significant predictor when non-significant interaction variables were excluded and maintained significance as the model was reduced further.

At the school level (Level 3), the effect of FRL was not significant ($p = .631$), indicating that this proxy for school-level SES was essentially unrelated to student scores on Test C. We determined a significant cross-level interaction ($p = .002$, $g = 0.106$) between prior achievement (CPA) and Curriculum. This finding suggests that students with higher prior mathematics achievement scores benefited more from the integrated curriculum than the subject-specific curriculum as measured by performance on Test C (Table 5). Variance estimates reveal that our model accounted for a substantial portion of the variance at each level, with student variables capturing much of the variation.

Modeling Test of Problem Solving and Reasoning (Test D) student scores. In Table 6, we report our model of student scores on Test D using the same general equations previously presented. Regarding content organization, no significant main effect of mathematics curriculum type was found ($p = .902$), indicating similar performance across the two groups of students. Likewise, we determined no significant effect of implementation fidelity ($p = .882$) on Test D scores. Three of six student covariates were significantly associated with performance, including prior achievement ($p < .001$) that had an effect size (g) of 0.586. Although Hispanic students' performance was not significantly different than that of White students, African American students and those of other races scored lower than White students by about 3 and 2 points, respectively. There was no

Table 5
HLM Fixed Effects Model Outcomes and Variance Components for Test of Common Objectives (Test C) Scores

	Unconditional model			
	Coeff	*SE*	*df*	*p*
Intercept	64.431***	0.839	10	0.000
	Random effect	*SD*	*df*	*p*
Level 1	72.707	8.527		
Level 2	27.544	5.248	53	0.000
Level 3	2.090	1.446	10	0.155

	Conditional models							
	Full model				Final model			
Fixed effect	Coeff	*SE*	*df*	*p*	Coeff	*SE*	*df*	*p*
Intercept	63.638***	1.136	9	0.000	63.528***	0.580	10	0.000
Level 1 (Student)								
CPA	0.195***	0.007	2431	0.000	0.199***	0.007	2446	0.000
Female	0.606*	0.280	2431	0.030	0.638*	0.278	2446	0.022
African American	–0.802	0.563	2431	0.154				
Hispanic	–0.976	0.507	2431	0.054				
Other	–1.483*	0.701	2431	0.034				
IEP	–0.304	0.609	2431	0.618				
Level 2 (Class)								
Curriculum	–0.019	1.413	49	0.990	0.160	0.929	59	0.864
FRL	–0.053	0.047	49	0.271				
CLE	0.288	0.597	49	0.632	0.959**	0.331	59	0.006
FID	–0.092	0.520	49	0.861				
OTL	2.656**	0.701	49	0.001	2.904***	0.619	59	0.000

TECHNOLOGY	0.243	0.499	49	0.628				
STANDARDS	−0.342	0.432	49	0.432				
Time_LD	0.009	0.019	49	0.633				
Experience	0.152	0.941	49	0.873				
PD	0.439	0.359	49	0.228				
Curriculum × FRL	0.011	0.055	49	0.847				
Curriculum × CLE	0.836	0.752	49	0.272				
Curriculum × FID	0.544	0.948	49	0.569				
Curriculum × OTL	–2.340*	1.150	49	0.047	–2.314*	0.881	59	0.011
Level 3 (School)								
FRL	0.0247	0.050	9	0.631				
Interaction (Cross-level)								
Curriculum × CPA	0.038**	0.012	2431	0.002	0.037**	0.011	2446	0.002

Random effects

	Variance component	*SD*	*df*	*p*	χ^2	Variance component	*SD*	*df*	*p*	χ^2
Level 1	46.259	6.801				46.413	6.813			
Level 2	4.747	2.179	39	0.000	256.9	5.623	2.371	49	0.000	279.6
Level 3	0.015	0.122	9	>0.500	8.4	0.001	0.037	10	0.412	10.3

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

significant effect of gender or special needs (IEP) status.

At Level 2, only OTL was a main effect, with increased opportunities significantly associated with higher scores on Test D (p = .012), an effect size (g) of 0.163. The FRL percentage was not significant (p = .087), but this result was elevated to the level of significance (p = .042) in the reduced model. As was the case with Test C, CLE emerged as a significant predictor (p = .008) when non-significant interaction variables were dropped and preserved its significance in further reductions of the model. No other class variables or interactions with curriculum type (including cross-level interactions) were found to be significant. Furthermore, the Level-3 variable FRL failed to yield predictive power. Variance estimates for Test D scores appear at the bottom of Table 6 (see on pp. 712–713).

Modeling ITED-16 student scores. Using the same general equations previously presented, we report our model of student scores on the ITED-16, a nationally normed test of problem solving and reasoning (Table 7) (see on pp. 714–715). Regarding content organization, we initially detected no significant main effect for mathematics curriculum type (p = .346). However, curriculum type was significant (p = .001) when we excluded non-significant interaction terms and thereafter as we further reduced the model. In the parsimonious model, students in the integrated curriculum outperformed those in the subject-specific curriculum by more than 10 points, with an effect size (g) of 0.294. Consistent with models of student scores on Tests C and D, there was no significant effect of FID (p = .528) on ITED-16 scores. Of the student covariates, only CPA (p < .001, g = .531) was positively associated with performance, in contrast to the remaining five student covariates: African American (p < .001, g = –0.355), other (p = .008, g = –0.188), Hispanic (p = .001, g = –0.170), IEP (p = .024, g = –0.135), and female (p = .001, g = –0.093). These student-level control variables enhanced our precision in the estimation of class-level effects, including curriculum.

Beyond curriculum type, we identified only one main effect at Level 2, namely OTL. As was the case in modeling scores on Tests C and D, a one-standard-deviation increase in OTL scores translated to about a 9-point increase in scores on the ITED-16, an effect size (g) of 0.254. Similar to the Test C model, there was a differential effect of OTL across curriculum. For classes in the integrated curriculum, the magnitude of the slope decreases dramatically from 8.887 to 0.650, essentially nullifying the effect of OTL. Finally, FRL (the Level-3 variable) was not a significant predictor of performance on the ITED-16. Variance estimates for ITED-16 are included in Table 7 and indicate most of the variation is captured by student-level variables.

Summary of Significant Effects

There were patterns of significant effects in our models of student scores. As shown in Table 8 (see on pp. 716–717), all student variables held significant explanatory power on one or more dependent measures. Most notably, our metric of prior achievement (CPA) was clearly the strongest predictor of performance, and this was true for all three dependent measures with effect sizes consistently

exceeding 0.5. Gender (being female) and race (being African American or Hispanic) or having an IEP held predictive power on some but not all dependent measures, whereas students coded as "other" (races) scored significantly lower on all three assessments. Such results are compelling in their own right but more importantly enabled us to more precisely estimate the effects of teacher-level variables that are at the center of our research program.

At the class level, with respect to our primary research question, we found a significant effect in favor of the integrated curriculum on the standardized measure (ITED-16) when non-significant interaction terms were excluded; the associated effect size of 0.294 is noteworthy. Curriculum type, however, had little bearing on student outcomes on our two project-developed tests.

The OTL factor was a significant predictor of student achievement on all three assessments with effect sizes ranging from 0.163 to 0.264. The interaction Curriculum × OTL significantly moderated student performance on two of three tests with substantial effect sizes. Similarly, the cross-level interaction CPA × Curriculum significantly moderated performance on Test C, with higher performing students benefitting from the integrated curriculum. This result is particularly notable given that students with higher prior achievement were more likely to have studied from the subject-specific curriculum.[13] Although our class-level aggregate measure of socioeconomic status (FRL) generally failed to demonstrate predictive power, a miniscule effect was detected in a reduced model of Test D scores. Interestingly, the classroom factors of FID, TECHNOLOGY, STANDARDS, and PD did not yield predictive value, as was also the case for the teaching experience dichotomous variable. There were no significant interactions of curriculum type with CLE, FID, or FRL in any of the full models. Nevertheless, when these interactions were removed, the CLE factor emerged as a significant predictor of scores on both Test C and Test D. With larger sample sizes, some class variables may have strengthened their associations with student outcomes.

Discussion

The purpose of our study was to examine the complex relationships between the textbook, implemented, and learned curriculum in second-year high school mathematics courses. More specifically, we sought to identify the effect of content organization on student learning and to identify significant curriculum implementation factors, as well as student and teacher characteristics, associated with student outcomes.

[13] The reasons why students selected one curriculum type over another were beyond the scope of our study. However, the propensity of higher performing students in two schools to have selected the subject-specific pathway provides some evidence that students, parents, counselors, and/or administrators might have perceived that the best and brightest mathematics students would be better served by a subject-specific curriculum. However, the significant CPA × Curriculum effect provides empirical evidence that higher performing students are instead better served by an integrated program.

Table 6
HLM Fixed Effects Model Outcomes and Variance Components for Test of Problem Solving and Reasoning (Test D) Scores

	Unconditional model			
	Coeff	*SE*	*df*	*p*
Intercept	63.763***	0.860	10	0.000
	Random effect	*SD*	*df*	*p*
Level 1	66.140	8.133		
Level 2	22.921	4.788	53	0.000
Level 3	3.152	1.776	10	0.056

	Conditional models							
	Full model				Final model			
Fixed effect	Coeff	*SE*	*df*	*p*	Coeff	*SE*	*df*	*p*
Intercept	64.142***	1.128	9	0.000	63.765***	0.753	10	0.000
Level 1 (Student)								
CPA	0.201***	0.007	2399	0.000	0.200***	0.005	2414	0.000
Female	–0.141	0.268	2399	0.598				
African American	–2.776***	0.545	2399	0.000	–2.612***	0.538	2414	0.000
Hispanic	–0.907	0.493	2399	0.066				
Other	–1.986**	0.686	2399	0.004	–1.836**	0.681	2414	0.007
IEP	–0.095	0.589	2399	0.872				
Level 2 (Class)								
Curriculum	–0.149	1.195	49	0.902	0.798	0.743	59	0.288
FRL	–0.070	0.040	49	0.087	–0.061*	0.030	59	0.042
CLE	0.606	0.537	49	0.265	0.880**	0.316	59	0.008
FID	–0.065	0.431	49	0.882				

OTL	1.563*	0.597	49	0.012	1.280**	0.461	59	0.008
TECHNOLOGY	0.459	0.451	49	0.314				
STANDARDS	0.104	0.376	49	0.784				
Time_LD	0.000	0.016	49	0.989				
Experience	−0.253	0.768	49	0.743				
PD	0.143	0.301	49	0.637				
Curriculum × FRL	−0.041	0.048	49	0.395				
Curriculum × CLE	0.448	0.620	49	0.473				
Curriculum × FID	−0.089	0.831	49	0.916				
Curriculum × OTL	−1.474	0.930	49	0.119				
Level 3 (School)								
FRL	0.057	0.058	9	0.354				
Interaction (Cross-level)								
Curriculum × CPA	−0.004	0.011	2399	0.739				

Random effects

	Variance component	*SD*	*df*	*p*	χ^2	Variance component	*SD*	*df*	*p*	χ^2
Level 1	42.255	6.500				42.311	6.505			
Level 2	2.422	1.556	39	0.000	157.4	2.725	1.651	49	0.000	169.8
Level 3	3.411	1.847	9	0.000	65.7	4.114	2.028	10	0.000	74.9

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

Table 7
HLM Fixed Effects Model Outcomes and Variance Components for ITED-16

	Unconditional model			
	Coeff	*SE*	*df*	*p*
Intercept	279.116***	3.189	10	0.000
	Random effect	*SD*	*df*	*p*
Level 1	887.958	29.799		
Level 2	266.702	16.331	53	0.000
Level 3	51.558	7.180	10	0.018

	Conditional models							
	Full model				Final model			
Fixed effect	Coeff	*SE*	*df*	*p*	Coeff	*SE*	*df*	*p*
Intercept	280.654***	3.528	9	0.000	278.801***	2.065	10	0.000
Level 1 (Student)								
CPA	0.661***	0.026	2482	0.000	0.692***	0.020	2496	0.000
Female	–3.266**	0.978	2482	0.001	3.247**	0.978	2496	0.001
African American	–12.451***	1.971	2482	0.000	–12.242***	1.967	2496	0.000
Hispanic	–5.963**	1.759	2482	0.001	–5.958**	1.763	2496	0.001
Other	–6.598**	2.477	2482	0.008	–6.538**	2.479	2496	0.009
IEP	–4.733*	2.099	2482	0.024	–5.266*	2.079	2496	0.012
Level 2 (Class)								
Curriculum	4.043	4.247	49	0.346	10.311**	2.837	61	0.001
FRL	–0.156	0.139	49	0.269				

CLE	1.469	1.836	49	0.427				
FID	–0.970	1.526	49	0.528				
OTL	8.887***	2.072	49	0.000	9.184***	1.510	61	0.000
TECHNOLOGY	2.350	1.530	49	0.131				
STANDARDS	0.300	1.313	49	0.820				
Time_LD	0.035	0.055	49	0.526				
Experience	–1.132	2.724	49	0.679				
PD	1.331	1.051	49	0.212				
Curriculum × FRL	–0.270	0.165	49	0.108				
Curriculum × CLE	0.656	2.206	49	0.767				
Curriculum × FID	–1.473	2.874	49	0.610				
Curriculum × OTL	–8.237*	3.340	49	0.017				
Level 3 (School)								
FRL	0.128	0.158	9	0.441				
Interaction (Cross-level)								
Curriculum × CPA	0.062	0.041	2482	0.126				

	Random effects									
	Variance component	*SD*	*df*	*p*	χ^2	Variance component	*SD*	*df*	*p*	χ^2
Level 1	579.105	24.065				579.544	24.074			
Level 2	32.081	5.664	39	0.000	163.4	45.634	6.755	51	0.000	217.9
Level 3	9.343	3.057	9	0.008	22.5	14.368	3.791	10	0.005	25.2

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

Table 8
Summary of HLM Findings Across Dependent Measures

	Significant on all outcomes
Student level	
CPA	Strong positive slopes with effect sizes ranging from 0.531 (ITED-16) to 0.549 (Test C) to 0.586 (Test D)
Other	Strong negative slopes with effect sizes ranging from –0.147 (Test C) to –0.188 (ITED-16) to –0.207 (Test D)
Class level	
OTL	Strong positive slopes with effect sizes ranging from 0.163 (Test D) to 0.254 (ITED-16) to 0.264 (Test C)

	Significant on some outcomes
Student level	
Female	Female students scored significantly higher than males on Test C with effect sizes of about 0.060 but scored significantly lower on ITED-16 with effect sizes of about –0.093
African American	African American students scored significantly lower than White students on Test D with effect sizes ranging from –0.289 to –0.272 in full and reduced models, respectively, and on ITED-16 with effect sizes of about –0.35
Hispanic	Hispanic students scored significantly lower than White students on ITED-16 only with effect sizes of about –0.17
IEP	Students with an IEP scored significantly lower on ITED-16 only with effect sizes ranging from –0.135 to –0.15 in full and reduced models, respectively
Class level	
Curriculum	Students of teachers using an integrated curriculum scored significantly higher on the ITED-16 with an effect size of 0.294 in the reduced model
Curriculum × OTL	The positive effect of OTL was moderated by curriculum type. For teachers of an integrated curriculum, the benefit of increased OTL was essentially nullified on Test C with effect sizes around –0.23, and on ITED-16 with an effect size of –0.235

CLE	Students of teachers with higher levels of CLE scored significantly higher on Test C and Test D with an effect size of about 0.09 in the reduced model
FRL	Small yet significant effect on Test D scores with an effect size of −0.006 in the reduced model only
Cross level	
CPA × Curriculum	Differential effect of prior achievement across curriculum types; students with higher prior achievement benefitted more from the integrated curriculum on Test C only, with an effect size around 0.10
Nonsignificant predictors	
Class level	
FID	Negligible negative slopes on all three outcomes
TECHNOLOGY	Negligible positive slopes on all three outcomes
STANDARDS	Negligible negative (Test C) or negligible positive (Test D and ITED-16) slopes on all three outcomes
Time_LD	Slopes near 0 on all three outcomes
Experience	Slopes near 0 (Test C) or negligible negative (Test D and ITED-16)
PD	Negligible positive slopes on all three outcomes
Curriculum × CLE	Negligible positive slopes on all three outcomes
Curriculum × FID	Negligible positive (Test C) or modest negative (Test D and ITED-16) slopes
Curriculum × FRL	Negligible positive slopes on all three outcomes

Discussion of Class-Level Findings

At the class level, only the OTL factor consistently offered predictive power, ranging from effect sizes of 0.181 to 0.294. These effect sizes were comparable in magnitude to the OTL effect sizes determined in the Grouws et al. (2013) study of first-year mathematics courses. Although some consider these to be small effect sizes (e.g., Cohen, 1988), such findings are compelling for several reasons. First, after controlling for prior achievement, this finding is consistent with the assertion that "opportunity to learn is widely considered the single most important predictor of student achievement" (Kilpatrick, Swafford, & Findell, 2001, p. 334). Second, as Floden (2002) argued,

> If OTL is not taken into account, its effects may be mistakenly attributed to some other attribute of the education system. A general rule in developing and testing models of schooling is that misspecification of the model, such as omitting an important variable like OTL, can lead to mistaken estimates of the effects of other factors. (p. 239)

Therefore, by essentially controlling for OTL and student attributes, the demonstrable effect of content organization in mathematics classrooms can be ascertained.

Our primary interest is in the relationship between content organization and student performance. Developed with substantial funding from the National Science Foundation, integrated curricula continue to be the subject of intense public scrutiny. Despite outspoken criticism of integrated curricula, results of this study do not support the notion that students learn more when studying mathematics that is organized in a subject-specific (traditional) manner. Indeed, after controlling for a large number of student and classroom attributes, students in the two curricular pathways performed comparably on a test of common objectives as well as an assessment of mathematical reasoning and problem solving. Moreover, students in the integrated curriculum outperformed others on a nationally normed test of problem solving and concepts, one with sound psychometric properties.

However, the curriculum effect, in favor of those studying from an integrated program, was detected only after constructing more parsimonious models in which non-significant predictors (e.g., interaction terms) were removed in iterative stages. Given that teachers of the integrated curriculum were significantly more likely to espouse agreement with the vision of the NCTM Standards, use graphing calculators, and organize students into collaborative learning groups, it follows that the curriculum type and the classroom factors TECHNOLOGY and STANDARDS are not unrelated but indeed entangled. Therefore, the interrelationship of these variables with curriculum type may have suppressed the curriculum effect in the initial (full) model as these three simultaneously competed for variation. Excluding non-significant classroom factors (TECHNOLOGY and STANDARDS) as well as non-significant interaction terms, the effect of the integrated curriculum was significant. Results of this study are consistent with those in Year 1 of the COSMIC project, in which we detected a significant curriculum effect in favor of the integrated program on three measures of high school students' learning in Algebra 1 and Core-Plus Course 1 (Grouws et al., 2013).

Somewhat surprising is the interaction between curriculum type and the OTL factor, in which the magnitude of the slope diminished by about 90% for classes in the integrated curriculum. Research strongly suggests that more opportunity to learn results in greater student learning, but in the case of integrated mathematics, the benefits of OTL were modest at best. On average, teachers in classrooms using the integrated curriculum reported teaching approximately 58% of lessons, while teachers using the subject-specific curriculum taught 86% of lessons. Despite such differences in OTL, teachers of both curriculum types taught the target content (i.e., mathematics content that comprised Test C). Regardless of whether teachers of the integrated curriculum covered less than half of or nearly the entire textbook, their students might perform comparably, as long as the target (assessed) content was taught.

Although this may explain the somewhat peculiar interaction of curriculum and OTL on Test C, it is more difficult to explain its relationship to ITED-16 scores. Perhaps teachers of the integrated curriculum went into greater depth in teaching those lessons that were covered, and such depth resulted in greater student learning on the ITED-16. More specifically, by providing markedly less coverage of Core-Plus, teachers of the integrated program may have traded breadth for depth. By teaching fewer lessons over the school year, teachers could have fostered student learning of the relationships between mathematical topics that comprise the integrated program, and the benefits of this less-is-more approach were then manifested on the ITED-16. Arguments for fewer topics with greater depth are certainly not new to school mathematics. Indeed, the U.S. school mathematics curriculum has long been criticized as "a mile wide and an inch deep" (Schmidt, 2001, p. 301).

Of particular interest are findings related to curriculum implementation. Arguably, we documented fidelity of implementation in unprecedented ways and found wide variations in teachers' use of curricular materials (Tables 3 and 4). These findings are consistent with Kilpatrick's (2003) assertion that "two classrooms in which the same curriculum is supposedly being 'implemented' may look very different" (p. 473). Moreover, our findings support the notion that teachers are active developers of the mathematics curriculum (Remillard, 2005; Remillard & Bryans, 2004) as they transform written curricular materials to meet the perceived needs of individual and classes of students. Ultimately, the Implementation Fidelity (FID) factor was not determined to have significant predictive power. However, it is worth noting that the slopes were consistently negative across all three dependent measures, and these findings parallel those of our related study (Grouws et al., 2013) of curricular effectiveness in first-year high school mathematics courses. Furthermore, our findings suggest that carefully thought out adaptations and deviations from textbook content and pedagogical recommendations might be warranted, although we have no data to support wholesale changes to the textbook curriculum.

Other findings warrant discussion, including why professional development was essentially unrelated to student performance. First, we measured PD only in terms of quantity, not quality, and therefore we cannot attest to whether the nature of the

PD mitigated its potential effect. However, from teacher surveys we determined that although many teachers participated in PD, its perceived impact was marginal at best because it merely confirmed what they were already doing. Consequently, student scores were likely unaffected by PD that did little to change teaching practices. Moreover, research indicates that consistent effects are found when teachers have received more than 100 hours of professional development (Banilower, Boyd, Pasley, & Weiss, 2006), an amount unattained by teachers in our sample.

The Classroom Learning Environment factor yielded predictive power when non-significant interactions (e.g., Curriculum × FRL) were not included in the full model. The positive effect of CLE is consistent with findings of earlier studies (e.g., Romberg & Shafer, 2008; Tarr et al., 2008) of curricular effectiveness. However, despite the positive contribution of CLE, we remind readers that generally low scores were determined on all three subscales of CLE (Table 4) that loaded greatest on the CLE factor. Such low scores for CLE subscales are consistent with results of the TIMSS Video Study (Jacobs et al., 2006), which found that typical mathematics teaching in the United States in both 1995 and 1999 "reflects the kind of traditional teaching that has been documented during most of the past century (Cuban, 1993; Fey, 1979; Hoetker & Ahlbrand, 1969; Welch, 1978), more so than the kind of teaching recommended in *Principles and Standards*" (pp. 28–29). Notwithstanding such low scores, the predictive value of CLE in parsimonious models of two outcome measures lends credence to the value of NCTM's recent initiatives (2000, 2010) to provide greater emphasis on reasoning and sense making in school mathematics.

Discussion of Student-level Findings

Given that "students are taking many more tests as a result of NCLB" (Jennings & Rentner, 2006, p. 111), officials in participating school districts were understandably unwilling to administer a fourth project-related assessment, namely a baseline test to gauge students' prior knowledge. Despite this, by using existing scores on state tests and mapping them to a common scale (CPA), we addressed the methodological difficulties inherent to our multistate research design. It is compelling to note that CPA consistently yielded the greatest student-level effect sizes, accounting for 0.53 to 0.59 standard deviations in student performance, depending on the outcome measure. Interestingly, the effect sizes documented herein are essentially identical to those reported in our study of students' learning in first-year high school mathematics courses (Grouws et al., 2013) and are well above those observed in recent studies of curricular effectiveness (e.g., Harwell et al., 2007; Post et al. 2008). Moreover, CPA was the only student-level variable with significant predictive power on all three outcome measures. For CPA, the effect sizes were at least double the magnitude of effect sizes determined for other student-level variables such as gender, race, and special education services (IEP status). Furthermore, effect sizes of CPA were approximately twice as large as those observed for class-level variables including curriculum type and the OTL factor. By accounting for prior knowledge and various student characteristics, our models offered increased

precision in the detection of key classroom factors—including content organization and implementation—that were the focus of this study. Consistent with our conceptual framework (Figure 1), we found students' prior knowledge to be a significant force on the learned curriculum. In our study, CPA was the greatest force influencing student achievement, although its effect was likely magnified because it was modeled at the student level, not at the class level as in prior studies (e.g., Post et al., 2008). Thus, by accounting for prior knowledge and various student characteristics, our models offered increased precision in the detection of significant classroom forces that influence student learning, including content organization and implementation, which were the primary focus of our study.

With respect to other student characteristics, our analyses yielded inconsistent findings. Regarding gender, females outperformed males on the Test of Common Objectives (Test C), scored comparably on the Test of Problem Solving and Reasoning (Test D), and were outperformed by males on the ITED-16. Similarly, although students with an IEP scored significantly lower on the ITED-16 than those without an IEP, no such disadvantages were observed in scores on the two project-developed tests. Further inconsistencies were evident with respect to performance data between races. Hispanic and African American students scored significantly lower than White students, but this achievement gap was statistically significant only on two dependent measures. Interestingly, the differential effects of numerous student characteristics were consistently manifested on the ITED-16 only. Why is it that this standardized measure of student learning is sensitive to these many student attributes? Clearly, the ITED is a trusted measure of student learning and, through extensive field-testing and refinement, its items are purportedly free of gender bias, for example.

With regard to achievement gaps between races, we offer a plausible explanation. Because Free or Reduced Lunch data were unavailable at the student level, it is possible that SES, not race, might explain these performance gaps, because non-White (especially Hispanic, African American, and American Indian/Native Alaskan) students are significantly more likely to qualify for the FRL program (NCES, 2010). Lending further credence to this notion is the fact that FRL was not found to have predictive power when aggregated at the class level and school level. If race were confounded with SES at the student level, then the predictive value of FRL at the class and school levels would likely be further diminished. Stated differently, if we had been able to link individual students to FRL status, then FRL might have yielded more predictive power while decreasing the effect of race at the student level.

Implications for Future Research

Debates regarding what mathematics should be learned, by whom it should be learned, when it should be learned, and how it should be taught are likely to persist at least for the next decade as the Common Core State Standards for Mathematics (CCSS-M) and corresponding accountability measures begin to be enacted in much of the United States. As suggested by NCTM (Rasmussen et al., 2011), the impact

of new common learning goals is worthy of extensive research, with the development of new curricular materials and as teachers transform them into daily mathematics lessons. Given the neutral position taken in the CCSS-M regarding whether to organize the high school mathematics curriculum into traditional or integrated pathways, more research is needed to examine what students learn over the course of three (or four) years of secondary school mathematics. Given our findings, the findings from previous research, and the fact that mathematics content is organized in an integrated manner in many countries, including the highest performing nations in international assessments (e.g., TIMMS, PISA), additional research should examine the effect of content organization as implemented in other integrated curriculum programs on student learning.

With respect to existing curricular options in the United States, subject-specific versus integrated, results of this study do not unequivocally answer the question "Which curriculum is best?" or even "Which organization of the school mathematics content is better?" However, given that students in the integrated program scored significantly higher on the ITED-16 and performed comparably on two other outcome measures, results of this study do not support the impassioned belief held by some that mathematics is best learned when the content is organized in a traditional manner. Our results, when coupled with those of related studies (Cai et al., 2011; Grouws et al., 2013), not only suggest that mathematics content organization is an important dimension of mathematics curriculum, they also indicate that the Core-Plus integrated program can yield greater student learning than approaches embodied by the subject-specific textbooks comprising our sample.

Nevertheless, the positive effect of the integrated curriculum on ITED-16 warrants further investigation. Why are performance differences manifested on the standardized measure but not on project-developed tests that were developed to be more sensitive to the study curricula? Furthermore, additional studies are needed to more closely examine student learning at the individual item level. Such finer grained analyses might identify important differences in student performance across curriculum types that were not detected because our use of IRT scale scores was based on the complete set of items.

Although we determined a curriculum effect on the standardized measure of student learning, the fact remains that school mathematics curricula are ever evolving. In the wake of the Common Core State Standards for Mathematics, the content, focus, and organization of high school mathematics textbooks are likely to undergo significant revisions in the near future. It follows that ongoing studies are needed to ascertain the curricular effectiveness of new programs, integrated and subject-specific, including representatives other than Core-Plus that embody substantially different program theory and embedded pedagogy. Future research should be conducted scientifically (Clements, 2007) and seek to avoid investigator bias that has historically plagued studies of curricular effectiveness (NRC, 2004).

Finally, comparative studies are needed to explore particular features of integrated curricula that may have contributed to the significant differences we found in student learning. For example, essential elements of program theory—such as

the role of technology and cooperative learning groups—are worthy of closer examination because they are facets of curriculum implementation that have potential explanatory power in developing a deep understanding of how curriculum impacts student learning.

Conclusion

The historical underperformance of U.S. students on international tests of mathematics achievement will likely continue to heighten public awareness and public scrutiny of school mathematics. As pressure increases to improve student achievement, curriculum will continue to be considered the primary leverage for change, as has been the case for decades. By partnering with schools that offered dual curricular paths, the COSMIC project was uniquely positioned to study the effectiveness of integrated and subject-specific mathematics programs over time. Our research design enabled us to ascertain complex relationships between curriculum organization, curriculum implementation, and student learning. Results of our initial study of students' learning in first-year high school mathematics courses (Grouws et al., 2013) showed that students in the integrated curriculum significantly outperformed those studying from a subject-specific curriculum on multiple outcome measures. Furthermore, the results reported herein continue to underscore the value of an integrated curricular program in second-year high school mathematics courses. Although numerous student characteristics and teacher practices and characteristics are associated with the learned curriculum, our series of studies has collectively determined that curriculum organization and its implementation are indeed key elements in efforts to increase students' learning in first- and second-year high school mathematics.

References

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6–8.

Banilower, E. R., Boyd, S. E., Pasley, J. D., & Weiss, I. R. (2006). *Lessons from a decade of mathematics and science reform: A capstone report for the Local Systemic Change through Teacher Enhancement Initiative*. Chapel Hill, NC: Horizon Research.

Bass, L., Charles, R. I., Jonson, A., Kennedy, D. (2004). *Geometry*. Upper Saddle River, NJ: Prentice Hall.

Bowzer, A. (2008). *Professional identity and curricular construction: A study of teacher interaction with mathematics curricula of two types* (Doctoral dissertation, University of Missouri, Columbia). Retrieved from https://mospace.umsystem.edu/xmlui/handle/10355/5610

Boyd, C. J., Cummins, J., Malloy, C., Carter, J., & Flores, A. (2005). *Geometry*. Columbus, OH: Glencoe-McGraw Hill.

Burger, E. B., Chard, D. J., Hall, E. J., Kennedy, P. A., Leinwand, S. J., Renfro, F. L., . . . Waits, B. K. (2007). *Geometry*. Austin, TX: Holt.

Cai, J., Wang, N., Moyer, J. C., Wang, C., & Nie, B. (2011). Longitudinal investigation of the curricular effect: An analysis of student learning outcomes from the LieCal Project in the United States. *International Journal of Educational Research, 50*(2), 117–136. doi:10.1016/j.ijer.2011.06.006

Center for Mathematics Education (CME). (2009). *CME Project: Geometry—Student edition for grades 8–12*. Boston, MA: Pearson.

Center for the Study of Mathematics Curriculum. (n.d.). *Curriculum Research Framework*. Retrieved from http://www.mathcurriculumcenter.org/research_framework.php

Chávez, Ó. (2003). *From the textbook to the enacted curriculum: Textbook use in the middle school mathematics classroom* (Doctoral dissertation, University of Missouri, Columbia). Retrieved from https://mospace.umsystem.edu/xmlui/handle/10355/10363

Chávez, Ó., Papick, I., Ross, D. J., & Grouws, D. A. (2011). Developing fair tests for mathematics curriculum comparison studies: The role of content analyses. *Mathematics Education Research Journal, 23*(4), 397–416. doi:10.1007/s13394-011-0023-2

Clements, D. H. (2007). Curriculum research: Toward a framework for "research-based curricula." *Journal for Research in Mathematics Education, 38*(1), 35–70.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Coxford, A. F., Fey, J. T., Hirsch, C. R., Schoen, H. L., Burrill, G., Hart, E. W., . . . Ritsema, B. (2003). *Contemporary mathematics in context: A unified approach* (*course 2*). Chicago, IL: Everyday Learning.

Cuoco, A., Goldenberg, E. P., & Mark, J. (2010). Organizing a curriculum around mathematical habits of mind. *Mathematics Teacher, 103*(9), 682–688.

Feldt, L. S., Forsyth, R. A., Ansley, T. N., & Alnot, S. D. (2003). Iowa Tests of Educational Development (Form B, Level 16) [Achievement test]. Chicago, IL: Riverside Publishing Company.

Floden, R. E. (2002). Measuring opportunity to learn. In A.C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national education surveys of educational achievement* (pp. 231–266). Washington, DC: National Academy Press.

George, A. A., Hall, G. E., & Uchiyama, K. (2000). Extent of implementation of a Standards-based approach to teaching mathematics and student outcomes. *Journal of Classroom Interaction, 35*(1), 8–25.

Grouws, D. A., & Smith, M. (2000). NAEP findings on the preparation and practices of mathematics teachers. In E. A. Silver & P. A. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 107–139). Reston, VA: National Council of Teachers of Mathematics.

Grouws, D. A., Smith, M. S., & Sztajn, P. (2004). The preparation and teaching practices of United States mathematics teachers: Grades 4 and 8. In P. Kloosterman, F. Lester Jr., & C. Brown (Eds.), *Results and interpretations of the 1990 through 2000 Mathematics Assessments of the National Assessment of Educational Progress* (pp. 221–267). Reston, VA: National Council of Teachers of Mathematics.

Grouws, D. A., Tarr, J. E., Chávez, Ó., Sears, R., Soria, V., & Taylan, R. D. (2013). Curriculum and implementation effects on high school students' mathematics learning from curricula representing subject-specific and integrated content organizations. *Journal for Research in Mathematics Education, 44*(2), 416–463.

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.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., . . . Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Huntley, M., & Chval, K. (2010). Teachers' perspectives on fidelity of implementation to textbooks. In B. J. Reys, R. E. Reys, & R. Rubenstein (Eds.), *Mathematics curriculum: Issues, trends, and future directions,* 2010 Yearbook of the National Council of Teachers of Mathematics (NCTM) (pp. 289–304). Reston, VA: NCTM.

Jacobs, J. K., Hiebert, J., Givvin, K. B., Hollingsworth, H., Garnier, H., & Wearne, D. (2006). Does eighth-grade mathematics teaching in the United States align with the NCTM Standards? Results from the TIMSS 1995 and 1999 video studies. *Journal for Research in Mathematics Education*, *37*(1), 5–32.

Jennings, J., & Rentner, D. S. (2006). Ten big effects of the No Child Left Behind Act on public schools. *Phi Delta Kappan, 88*(2), 110–113.

Jones, P. S. (Ed.). (1970). *A history of mathematics education in the United States and Canada*. (1970 Yearbook of the National Council of Teachers of Mathematics). Reston, VA: National Council of Teachers of Mathematics.

Kilpatrick, J. (2003). What works? In S. L. Senk & D. R. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 471–493). Mahwah, NJ: Lawrence Erlbaum.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

King, K. D., Mitchell, M. B., Tybursky, J., Simic, O., Tobias, R., Barriteau Phaire, C., & Torres, M. (2011, April). *Impact of teachers' use of Standards-based instructional materials on students' achievement in an urban district: A multilevel analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Klibanoff, R., Levine, S. C., Huttenlocher, J., Vasilyeva, M., & Hedges, L. (2006). Preschool children's mathematical knowledge: The effect of teacher "math talk." *Developmental Psychology, 42*(1), 59–69. doi:10.1037/0012-1649.42.1.59

Larson, R., Boswell, L., & Stiff, L. (2001). *Geometry*. Evanston, IL: McDougal Littell.

McClain, K., Zhao, Q., Visnovska, J., & Bowen, E. (2009). Understanding the role of institutional context in the relationship between teachers and text. In J. T. Remillard, B. A. Herbel-Eisenmann, & G. M. Lloyd. (Eds.), *Mathematics teachers at work* (pp. 56–69). New York, NY: Routledge.

Moyer, J. C., Cai, J., Wang, N., & Nie, B. (2011). Impact of curriculum reform: Evidence of change in classroom practice in the United States. *International Journal of Educational Research, 50*(2), 87–99. doi:10.1016/j.ijer.2011.06.004

Nathan, M. J., Long, S. D., & Alibali, M. W. (2002). The symbol precedence view of mathematical development: A corpus analysis of the rhetorical structure of textbooks. *Discourse Processes, 33*(1), 1–21. doi:10.1207/S15326950DP3301_01

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.

National Center for Education Statistics. (2010). *Status and trends in the education of racial and ethnic minorities* (NCES 2010-015). U.S. Department of Education. Washington, DC: Author.

National Council of Teachers of Mathematics. (1980). *An agenda for action*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

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

National Council of Teachers of Mathematics. (2010). *Focus in high school mathematics: Reasoning and sense making*. Reston, VA: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010a). *Common core state standards for mathematics*. Retrieved from http://www.corestandards.org

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010b). *Common core state standards for mathematics, appendix A: Designing high school mathematics courses based on the Common core state standards*. Retrieved from http://www.corestandards.org/assets/CCSSI_Mathematics_Appendix_A.pdf

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

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002). Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

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.

Rasmussen, C. L., Heck, D. J., Tarr, J. E., Knuth, E., White, D. Y., Lambdin, D. V., . . . Barnes, D. (2011). Trends and issues in high school mathematics: Research insights and needs. *Journal for Research in Mathematics Education, 42*(3), 204–219.

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research, 75*(2), 211–246. doi:10.3102/00346543075002211

Remillard, J. T., & Bryans, M. B. (2004). Teachers' orientations toward mathematics curriculum materials: Implications for teacher learning. *Journal for Research in Mathematics Education, 35*(5), 352–388. doi:10.2307/30034820

Riverside Publishing (n. d.). Iowa tests of educational development. Retrieved from http://www.riversidepublishing.com/products/ited/details.html

Romberg, T. A. (2010). Classic publications on the mathematics curriculum. In B. Reys & R. Reys (Eds.), *Mathematics curriculum: Issues, trends, and future directions* (pp. 1–21). Reston, VA: National Council of Teachers of Mathematics.

Romberg, T. A., & Shafer, M. C. (2008). *The impact of reform instruction on student mathematics achievement: An example of a summative evaluation of a standards-based curriculum*. New York, NY: Routledge.

Schmidt, W. H. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass.

Schoen, H. L., & Hirsch, C. R. (2003). The Core-Plus Mathematics Project: Perspectives and student achievement. In S. Senk & D. Thompson (Eds.), *Standards-based school mathematics curricula: What are they? What do students learn?* (pp. 311–343). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A. H. (2004). The math wars. *Educational Policy, 18*, 253–286. doi:10.1177/0895904803260042

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.

Snyder, J., Bolin, E., & Zumwalt, K. (1992). Curriculum implementation. In P.W. Jackson (Ed.), *Handbook of research on curriculum* (pp. 402–435). New York, NY: Macmillan.

Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology, 94*(2), 344–355. doi:10.1037/0022-0663.94.2.344

Stein, M. K., & Kaufman, J. H. (2010). Selecting and supporting the use of mathematics curricula at scale. *American Educational Research Journal, 47*(3), 663–693. doi:10.3102/0002831209361210

Stein, M. K., Remillard, J., & Smith, M. S. (2007). How curriculum influences student learning. In F. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 319–370). Charlotte, NC: Information Age Publishing.

Superfine, B. M. (2009). Deciding who decides questions at the intersection of school finance litigation and standards-based accountability policies. *Educational Policy, 23*(3), 480–514. doi:10.1177/0895904808314712

Tarr, J. E., Chávez, Ó., Reys, R. E., & Reys, B. J. (2006). From the written to the enacted curricula: The intermediary role of middle school mathematics teachers in shaping students' opportunity to learn. *School Science and Mathematics, 106*(4), 191–201. doi:10.1111/j.1949-8594.2006.tb18075.x

Tarr, J. E., McNaught, M. D., & Grouws, D. A. (2012). The development of multiple measures of curriculum implementation in secondary mathematics classrooms: Insights from a three-year curriculum evaluation study. In I. Weiss, D. Heck, K. Chval, & S. Zeibarth (Eds.), *Approaches to studying the enacted curriculum* (pp. 89–115). Greenwich, CT: Information Age Publishing, Inc.

Tarr, J. E., Reys, R. E., Reys, B. J., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middle grades mathematics curricula on student achievement and the classroom learning environment. *Journal for Research in Mathematics Education, 39*(3), 247–280.

Tarr, J. E., Ross, D. J., McNaught, M. D., Chávez, Ó., Grouws, D. A., Reys, R. E., . . . Taylan, R. D. (2010, April 29–May 5). *Identification of student- and teacher-level variables in modeling variation of mathematics achievement data*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Thompson, D. R., & Senk, S. L. (2010). Myths about curriculum implementation. In B. Reys, R. Reys, & R. N. Rubenstein (Eds.), *K–12 curriculum issues* (pp. 249–263). Reston, VA: National Council of Teachers of Mathematics.

United States Department of Education Institute of Educational Sciences. (2010). *What Works Clearinghouse intervention report: Core-Plus Mathematics*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_coreplus_092110.pdf

United States Department of Education Institute of Educational Sciences. (2011). *What Works Clearinghouse procedures and standards handbook*. Retrieved from http://ies.ed.gov/ncee/wwc/references/idocviewer/Doc.aspx?docId=19&tocId=4

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: A study of K–12 mathematics and science education in the United States*. Chapel Hill, NC: Horizon Research.

Authors

James E. Tarr, Department of Learning, Teaching, and Curriculum, University of Missouri, 303 Townsend Hall, Columbia, MO 65211-2400; tarrj@missouri.edu

Douglas A. Grouws, Department of Learning, Teaching, and Curriculum, University of Missouri, 303 Townsend Hall, Columbia, MO 65211-2400; grouwsd@missouri.edu

Óscar Chávez, Department of Mathematics, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249-0624; oscar.chavez@utsa.edu

Victor M. Soria, Department of Learning, Teaching, and Curriculum, University of Missouri, 303 Townsend Hall, Columbia, MO 65211-2400; soriav@missouri.edu

Accepted March 16, 2012

APPENDIX A

Teacher Variables by Data Source

Table of Contents Records

Variable	Description
OTL Index	Represents the percentage of textbook lessons taught
ETI Index	Represents the extent to which teachers followed their textbook using weighted averages
TCT Index	Represents the extent to which teachers, *when teaching textbook content*, followed their textbook, supplemented their textbook lessons, or used alternative curricular materials

Classroom Visit Protocols

Variable	Description
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	Percent of observed lessons that teacher utilized graphing calculators during instruction
Tech_Students	Percent of observed lessons that most students utilized graphing calculators during instruction
Reasoning	Classroom Learning Environment Inventory Subscale: Reasoning About Mathematics
Students' Thinking	Classroom Learning Environment Inventory Subscale: Students' Thinking in Instruction
Sense Making	Classroom Learning Environment Inventory Subscale: Sense Making About Mathematics
Closure	Percent of observed lessons that teacher brought closure
Engage	Extent to which most students were engaged (on task) during observed lesson
Seating	Percent of observed lessons that students were seated in groups
Collaboration	Percent of observed lessons that students worked collaboratively

APPENDIX A (*continued*)

Teacher Variables by Data Source

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 (seatwork)

Initial Teacher Survey

Variable	Description
Teaching_Exp	Number of years of teaching experience
Math_Exp	Number of years of 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 months
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)

Mid-Course Teacher Survey

Variable	Description
Text	Number of years teaching from the district-adopted textbook
Preparation	Rating of preparedness to teach the district-adopted textbook
Satisfaction	Rating of satisfaction with the district-adopted textbook