NCEE TECHNICAL APPENDIX September 2013

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

AFTER TWO YEARS, THREE ELEMENTARY MATH CURRICULA OUTPERFORM A FOURTH

This appendix provides the details that underlie the analyses reported in the evaluation brief, “After Two Years, Three Elementary Math Curricula Outperform a Fourth.” The details are organized in six sections: Study Curricula and Design (Section A), Data Collection (Section B), Construction of the Analysis File (Section C), Curriculum Effects on Student Math Achievement (Section D), Curriculum Implementation (Section E), and Effects of Switching Curricula (Section F).[1]

A. Study Curricula and Design

This section provides details about the curricula included in the study, recruitment of study participants, random assignment of curricula to schools, and site participation and curriculum implementation during years one and two.

1. Curriculum Descriptions and Training

A competitive process was used to select the study curricula. Nine submissions were received. A panel of independent experts in math and math instruction reviewed the submissions and recommended to IES curricula suitable for the study. Following in-person meetings with publishers whose curricula were considered strong candidates for the study, IES selected the four curricula for the study. Curricula submitted for consideration but not selected are not disclosed, because the proposals are confidential. In June 2006, IES selected the following four curricula for the study:

- ***Investigations in Number, Data, and Space*** (Investigations) is published by Pearson Scott Foresman and uses a student-centered approach encouraging metacognitive reasoning and drawing on constructivist learning theory. The lessons focus on understanding, rather than on students answering problems correctly, and build on students’ knowledge and understanding. Students are engaged in thematic units of three to eight weeks in which they first investigate and then discuss and reason about problems and strategies. Students frequently create their own representations.

- ***Math Expressions*** is published by Houghton Mifflin Harcourt and blends student-centered and teacher-directed approaches to mathematics. Students question and discuss mathematics but are also explicitly taught effective procedures. The curriculum emphasizes using multiple specified objects, drawings, and language to represent concepts and learning through the use of real-world situations. Students are expected to explain and justify their solutions.

[1] Section G presents results from supplemental analyses that examine curriculum effects during a second year of implementation in the 1st grade. The purpose of these supplemental analyses is to provide information to district and school staff that is useful for understanding whether the curriculum used in a particular grade matters during subsequent years of usage in a grade.

- ***Saxon Math*** (Saxon) is published by Houghton Mifflin Harcourt and is a scripted curriculum that blends teacher-directed instruction of new material with daily distributed practice of previously learned concepts and procedures. The teacher introduces concepts or efficient strategies for solving problems. Students observe and then receive guided practice, followed by distributed practice. Students hear the correct answers and are explicitly taught procedures and strategies. Frequent monitoring of student achievement is built into the program. Daily routines are extensive and emphasize practice of number concepts and procedures and use of representations.

- ***Scott Foresman-Addison Wesley Mathematics*** (SFAW) is published by Pearson Scott Foresman and is a basal curriculum that combines teacher-directed instruction with a variety of differentiated materials and instructional strategies. Teachers select the materials that seem most appropriate for their students. The curriculum is based on a consistent daily lesson structure, which includes direct instruction, hands-on exploration, the use of questioning, and practice of new skills.

Throughout the brief and this appendix, we refer to SFAW/enVision because the publisher of SFAW revised the curriculum during the study and renamed the curriculum enVision Math (enVision).[2] This change did not affect three of the seven districts that participated in the study for two years because their participation preceded the change to enVision. However, the change affected the other four districts, because it occurred halfway through their participation. Their schools, which were initially assigned to SFAW, used it during the first year but used enVision during the second. Therefore, during the second year, about one-quarter of students attending a school initially assigned to SFAW experienced that curriculum; the other three-quarters experienced enVision.

As the brief mentions, this curriculum change is relevant for analyses designed to answer the research questions about two years of experience with the curricula and student achievement. Therefore, we examined the implications of the SFAW-enVision change by comparing results based on students from all seven districts and those from the three districts that were not affected by the SFAW-enVision change. The pattern of results is similar across the two sets districts and, therefore, we report the ones based on the larger sample of seven districts. The other curricula also were updated, but the changes were not substantial. Publishers described updates to Investigations, Math Expressions, and Saxon as “copyright refreshes.” The publishers of Investigations and Saxon advertised their updated materials as new editions, with Investigations issuing a second edition and Saxon issuing a third edition. The three districts mentioned above used these updated versions during their second year of participation; the other four districts used the updated versions during both years.

[2] enVision is a K–6 curriculum that develops math concepts through interactive activities and visual learning strategies. The curriculum is based on a consistent daily lesson structure that involves a review of previously learned materials, hands-on interactive activities to introduce new concepts, direct instruction of new skills or concepts using visual learning strategies, guided practice, differentiated independent practice, and a closure activity that allows teachers to assess student understanding and make clear to students the concept or skill covered that day.

a. Important Differences between the Curricula

Generally speaking, the curricula vary in the extent to which they emphasize student-centered or teacher-directed approaches, but there are many specific differences between the curricula that could be important for student achievement. To summarize important differences across the study's curricula, we drew on research on effective mathematics instruction and curriculum materials as resources for teachers, to identify features of curricula that are likely to have a strong influence student achievement. We used three features from the literature to frame a comparative analysis:[3]

1. **Mathematical emphasis.** This feature refers to the mathematics knowledge and practice that is valued and the quality and treatment of mathematics in each curriculum. Our particular interest was in the kinds of mathematics students had opportunities to learn and how those opportunities were structured.
2. **Instructional approach.** Students' opportunities to learn the mathematics offered in a curriculum are influenced by the instructional approach taken in the curriculum. The instructional approach refers to the roles the teacher is expected to play during instruction, the types of activities in which students are expected to engage, and the nature of the classroom interactions.
3. **Supports for teachers.** We examined differences in the types of guidance curriculum authors provided teachers to implement the curricular designs. This dimension of curriculum design is particularly relevant when a curriculum requires teachers to teach in unfamiliar ways.

This comparative analysis was based on a careful review of the curriculum materials, including a systematic review of one component of mathematical emphasis: cognitive demand. The following sections provide details about the analytical methods. As they explain, we found substantial variation in each feature across the four curricula.

Mathematical emphasis. There is general agreement among researchers of mathematics education that both conceptual understanding and procedural fluency are necessary components of high-quality mathematics instruction. Hiebert and Grouws (2007) identified two key features of instruction found to lead to high levels of conceptual understanding. First, instruction must attend explicitly to concepts, "to connections among mathematical facts, procedures, and ideas" (pp. 383). Second, students should struggle (or grapple) with important mathematics.[4]

To assess mathematical emphasis of the curricula, including the ways in which they attend to conceptual understanding and procedural fluency, we examined three components in the curricula: (1) the cognitive demand of the curricula, (2) regular routines that provide

[3] Our analysis does not include other categories of instruction that could be important for student achievement because they were outside of the scope the study. For example, research has shown formative assessment to be an important aspect of instruction.

[4] Hiebert and Grouws (2007) elaborate on the term "struggle" by stating that it is "the opposite of simply being presented information to be memorized or being asked only to practice what has been demonstrated" (pp. 387–388).

opportunities for engagement with concepts, facts, and procedures, and (3) repeated practice to develop procedural fluency.

First, we assessed ***cognitive demand*** (Stein et al. 1996), to identify the extent to which the mathematics tasks attended to connections among concepts, procedures, and facts and the potential opportunities to grapple with mathematics. To perform the systematic comparison of cognitive demand, we selected 10 lessons from each curriculum, 5 from each of the two grades examined in this report (1st and 2nd). Because a few of the schools that were initially assigned to SFAW switched to enVision for 2nd grade, we also examined 5 of the 2nd-grade enVision lessons. We identified comparable lessons across curricula that contain two main tasks (main tasks were defined as those requiring the majority of time), following the procedure used by Stein et al. (1996). To ensure comparability across curricula, we selected lessons that addressed the same topics within the Number and Operations strand from each program. We selected this strand because developing number concepts and fluency with operations is the central focus of early elementary curriculum; moreover, the strand is given even greater prominence in grades 1 and 2 in the Common Core State Standards. Each task was then coded using the cognitive demand criteria provided by Stein et al. (2000). Two authors of the current report—Harris and Remillard—initially coded several tasks in each curriculum together, discussing how the criteria are operationally defined for the purposes of coding, and resolved differences when they occurred. Harris then coded the remaining tasks, and Remillard reviewed these codings, making revisions when appropriate.

In Stein et al.'s framework (2000), high-demand tasks are intellectually and conceptually challenging and place emphasis on underlying concepts, patterns, and properties. These tasks are classified as *Doing Mathematics* (DM), which involves nonroutine thinking, or *Procedures with Connections* (PWC), which emphasizes underlying meaning within procedural routines or practices. They differ in that DM tasks tend to allow for multiple solution paths and require students to make connections and develop strategies, often drawing on their informal knowledge; PWC tasks provide students with solution paths to follow but in a way that connects them to underlying concepts. As such, DM tasks are more likely to provide opportunities for students to struggle with important mathematics, whereas PWC tasks attend explicitly to connections among concepts, procedures, and facts (Hiebert and Grouws 2007). Low-demand tasks are classified as *Procedures without Connections* (PWOC) or *Memorization* (M), both of which focus on routine and procedural elements of mathematical tasks, often in isolation, and without connections to mathematical sense-making. (See Stein et al. 2000 for more details.)[5]

Second, we examined the ***regular routines*** built into the curriculum to assess the extent to which the curriculum provided opportunities for regular engagement with concepts and their relationships with facts and procedures. Routines can occur at the beginning of the lesson, or at another time during the school day. By examining the ways that routines are incorporated into the selected lessons, and through reading other portions of the teacher's guide that provide

[5] Mathematics tasks in curriculum materials infrequently call for a single type of thinking; when classifying tasks, therefore, we attended to their primary emphasis determined by how the child would be expected to spend most of his or her time. In a number of Saxon and SFAW/enVision tasks, for example, underlying concepts are represented initially, but the majority of student activity during the task does not make reference to these ideas. Such tasks were coded as PWOC.

guidance to teachers on how to implement routines, we characterized the frequency, extent, and mathematical emphasis of the daily routines.

Third, we looked at how ***repeated practice*** is treated to compare differences in how procedural fluency is developed within each curriculum. In particular, we examined the extent to which students are expected to engage in regular practice of skills and procedures within and beyond the specific content of the daily lessons.

The analysis of ***regular routines*** and ***repeated practice*** drew on the systematic coding of cognitive demand, and on a careful review of the entire package of curriculum materials provided by each publisher. Specifically, each curriculum includes a teacher's guide and supplementary materials that describe how to use the curriculum. Each teacher's guide includes introductory pages about the design of the curriculum and key features; some include detailed explanations of the mathematics. We reviewed the introductory information in each teacher's guide, the supplementary materials from each curriculum, and the daily lesson guides for each task that was coded for cognitive demand.

In coding the regular routines and repeated practice, we drew on all reviewed materials to describe how routines and practice were characterized in the curriculum materials. Two study authors (Harris and Remillard) independently described each curriculum by examining when the routines or practice occur, their frequency and duration, and the approach or types of activities that should occur. After summarizing each curriculum, the authors discussed and resolved any differences. The authors then examined whether the curricula differ along the dimensions examined.

Findings about mathematical emphasis. Compiling findings across the three components of the mathematical emphasis offers insight into the extent to which the curricula attend to building conceptual understanding, developing procedural fluency, and encouraging mathematics thinking, reasoning, and strategic competence (NRC 2001).

All four curricula include at least a majority of high-demand tasks (Table A.1). This indicates that all four curricula pay substantial attention to developing conceptual understanding. However, Investigations and Math Expressions differ from Saxon and SFAW in that nearly all (95 percent) of Investigations' and Math Expressions' tasks are high demand, whereas 57 to 65 percent of Saxon's and SFAW/enVision's task are low demand.

Table A.1. Cognitive Demand of the Primary Tasks in Each Curriculum

	Low-Demand Tasks		High-Demand Tasks	
Curriculum and Number of Tasks Examined[a]	Memorization	Procedures Without Connections	Procedures With Connections	Doing Math
Investigations n=20	-	1 (5%)	11 (55%)	8 (40%)
Math Expressions n=20	1 (5%)	-	13 (65%)	6 (30%)
Saxon n=20	-	7 (35%)	13 (65%)	-
SFAW/enVision n=30	1 (3%)	12 (40%)	17 (57%)	-

[a]For each curriculum, we coded the 2 tasks that required the majority of time in each of 10 lessons (5 from grade 1 and 5 from grade 2), for a total of 20 tasks for each curriculum. Because some schools using SFAW in 1st grade transitioned to enVision for 2nd grade, 2 tasks in each of 5 2nd-grade enVision lessons were analyzed along with the SFAW lessons.

Looking at the other two components of mathematical emphasis (daily routines and use of practice), we conclude that Saxon and Math Expressions both emphasize procedural fluency, although Saxon does so in greater measure (Tables A.2 and A.3). The routines in Saxon stand out in the extent to which they are integrated into the curriculum and the amount of time devoted to them. On a daily basis, Saxon's routine activities expose students to a number of skills, with a focus on mastery and fluency of foundational skills, more than making conceptual connections. The routines in Math Expressions and Investigations are similar in design. Each program includes a set of conceptually oriented routines that the teacher draws from. Mathematically, the Math Expressions routines place extensive focus on place value and quantitative concepts and take a PWC approach. The Investigations routines also focus on number concepts but tend to be structured like many DM activities, providing students with opportunities to develop strategies and make mathematical connections. SFAW/enVision does not include a specified regular routine, although it provides a problem of the day and a repeated practice worksheet that can be used in a routine, if the teacher chooses.

Table A.2. Characteristics of the Routines in Each Curriculum

Curriculum	Frequency	Length	When (Level of Specificity)	Types of Activity
Investigations	Daily	10 minutes	Outside of lesson (use clearly specified)	Single activity identified from a set of four conceptually oriented activities focused on number relationships and time
Math Expressions	Daily	5–10 minutes	At beginning of lesson or outside of lesson (use underspecified)	Set of conceptually oriented activities focused on place value, number relationships, and time
Saxon	Daily	20 minutes	At beginning of lesson or outside of lesson (use scripted)	Set of six to nine daily fluency activities that include practice reading calendars and clocks, practice computation facts, problem solving, counting, representing quantities, and graphing the weather and attendance
SFAW/enVision	Optional	Unspecified	At beginning of lesson or outside of lesson (use unspecified)	No routine is identified; a problem of the day and review worksheet could be used

All four programs employ repeated practice in the lesson, but the amount and focus vary. Investigations and Math Expressions both place minimal emphasis on repeated practice within the lesson, although each provides follow-up worksheets to reinforce the content of the lesson. In Investigations, these sheets generally maintain emphasis on underlying concepts, whereas in Math Expressions, they tend to emphasize fluency. In Saxon and SFAW/enVision, repeated practice of the skill taught is incorporated into each daily lesson and the focus of their practice tends to be fluency.

In addition to repeated practice within the lesson, as shown in Table A.3, three of the four curricula employ repeated practice outside of the main lesson. Math Expressions and Saxon provide a daily activity focused on fluency of basic math facts or number concepts; these activities can occur prior to the math lesson, or at another time of the day (and they are separate from the routines). Investigations provides some repeated practice within the routine, although the extent of practice depends on the specific routine being implemented (teachers use four routines in regular rotation) and the emphasis is not necessarily fluency. SFAW/enVision imbeds repeated practice in its optional "Spiral Review" activity, which some teachers choose to use as part of their regular routine.

Table A.3. Approaches Taken to Repeated Practice of Skills Beyond the Lesson

Curriculum	Frequency	Length	Approach
Investigations	Regularly, but not daily	10 minutes	Some Routines are focused on number concepts
Math Expressions	Daily	5-10 minutes	"Quick Practice" activity focused on fluency or number concepts
Saxon	Daily	10-15 minutes	"Fact Practice" activity focused on fluency of facts
SFAW/enVision	Optional with each lesson	Unspecified	"Spiral Review" worksheet provides two to four practice problems

Instructional approach. In each of lessons analyzed for cognitive demand, we assessed the instructional approach by examining the teacher's role during instruction, the types of activities in which students are expected to engage, and the nature of the classroom interactions. Specifically, in each lesson coded for cognitive demand, we looked at instructions to the teacher about how to present the content to students, how to interact with the students, and the types of activities in which students should engage. We also reviewed supplementary documents of each curriculum (such as implementation guides and introductory materials for teachers) to see how the curriculum materials described the desired teacher and student roles in the overall curricular approach. In addition, we examined these materials to identify the most desired classroom interactions and participant structures.[6]

The analytical approach followed the same approach as the careful review of ***regular routines*** and ***repeated practice***. Two study authors (Harris and Remillard) independently reviewed the curriculum materials to identify the teacher's role during instruction, the types of activities in which students are expected to engage, the nature of the classroom interactions, and participant structures. After summarizing each curriculum along these dimensions, the authors discussed and resolved differences, if they occurred. The authors then examined whether the curricula differ along the dimensions examined.

Findings about instructional approach. The instructional approaches offered in the four curricula differ with respect to the roles played by the teacher, students, and text in shaping classroom interactions and student learning (Table A.4). Investigations, Math Expressions, and Saxon have designed classroom exchanges where students and teachers interact with one another around the intended mathematics activities and concepts. In addition, Investigations and Math Expressions emphasize student-to-student interactions and provide opportunities for students to work together and communicate their mathematics knowledge. In SFAW/enVision, the predominant classroom interaction is between the student and the text (worksheet, workbook, or

[6] "Participant structure" refers to the different ways that teachers arrange interactions with students, such as whole-class instruction, teacher-led small groups, independent seatwork, and group projects (Philips 1972). We use the phrase to describe who can do and say what, and when.

other curriculum material). Munter et al. (2013) use the terms dialogic and direct to differentiate instructional models that place emphasis on student generation and exchange of ideas from models in which skills and knowledge are passed from teacher to student. Investigations uses a dialogic instructional model. The primary pathway for learning is between students with the guidance of the teacher; the predominant teacher role involves interacting with students and facilitating student production of ideas. In contrast, Saxon uses a direct instructional model; the knowledge moves from the teacher to the student, and the associated teacher's role involves explaining concepts, demonstrating procedures, and guiding students while they work. The Math Expressions curriculum takes an approach that blends these two models, providing opportunities for student production of ideas as well as direct teacher explanations. In contrast to these three programs, SFAW/enVision promotes classroom interactions focused on the workbook pages, which seem to be the primary pathway for learning. Like Saxon, SFAW/enVision adopts a direct instructional model. The teacher's role involves explaining, demonstrating, and guiding; the aim of this guidance, however, is to support students as they complete the pages.

Table A.4. Key Aspects of the Instructional Approach in each Curriculum

Curriculum	Classroom Interactions	Teacher's Role	Pathway for Learning (Instructional Model)
Investigations	Teacher-Student (Student-Student)	Facilitate student production of ideas	Between students and teacher (Dialogic)
Math Expressions	Teacher-Student (Student-Student)	Explain, model, facilitate production of ideas	From teacher to students Between students (Blended Dialogic and Direct)
Saxon	Teacher-Student	Explain, demonstrate, guide	From teacher to students (Direct)
SFAW/envision	Student-Text	Explain, demonstrate, guide	From text to students (Direct)

Supports for teachers. Drawing on research on curriculum design and use (Davis and Krajcik 2005; Remillard and Bryans 2004; Stein and Kaufman 2010; Stein and Kim 2009), we analyzed both how each program provides guidance to teachers and the topics of guidance. Once again, the analytical approach followed the same approach as the careful review of ***regular routines***, ***repeated practice***, and ***instructional practice***. That is, two study authors (Harris and Remillard) independently reviewed the curriculum materials, summarized each curriculum along the dimensions identified in the literature, and discussed and resolved differences, if they occurred. The authors then examined whether the curricula differ along the dimensions examined.

Conventional teacher's guides tend to provide guidance that direct teachers' instructional actions, by providing teachers with a collection of tasks to present to students and questions to ask. Remillard (2000) refers to this approach as speaking through the teacher. Some curriculum developers have begun to design teacher's guides that also speak to teachers about the design of the lessons, the mathematical and pedagogical ideas underlying them, and how students might respond. This latter type of guidance may be especially important for curriculum programs that adopt instructional models and mathematical emphases that are likely to be challenging for teachers to implement, since Stein and Kim (2009) argue that high cognitive demand tasks and dialogic instructional approaches place substantial demands on teachers and require more support in the teacher's guide.

How texts guide teachers. There are generally two different approaches used to guide teachers: (1) explicit scripts and (2) descriptive scripts (Remillard and Reinke 2012). Explicit scripts include a high level of detail that specifies exact sentences for teachers to deliver verbally, exact words to write on the board, or specific visual models to demonstrate. Descriptive scripts guide teachers' and students' actions or dialogue by describing what should be said, written, visually demonstrated, or done.

Based on our review of the curricula, all four programs blend these two approaches, but in different ways (Table A.5). Investigations blends descriptions of teacher actions with selective explicit scripts of questions the teacher should ask or ways to respond to students. Math Expressions is even more detailed in its descriptive script, and less frequently scripts the teacher's words. Unlike the other three, Saxon provides a fully scripted lesson containing everything the teacher should say, and a detailed descriptive script. SFAW/enVision's guidance is less extensive than the other curricula, providing general guidance for the teacher actions.

Topics of Guidance. Another difference in guidance is in what the curriculum communicates to the teacher. In addition to directing teachers' actions, some curriculum designs offer support by (a) helping teachers attend to student thinking, (b) providing subject-specific content support, and (c) clarifying curriculum designers' rationale or intent (Ball et al. 2005; Remillard 2000).

Based on our review of the curricula, both Investigations and Math Expressions provide guidance on a variety of teaching components, including mathematical concepts, student thinking, and ways to adapt a lesson for specific students. SFAW/enVision provides guidance on few topics, as does Saxon, which primarily focuses on classroom organization and management.

Table A.5. How Each Program Provides Guidance to Teachers and the Topics of Guidance

Curriculum	How the Text Guides Teachers	Topics of Guidance
Investigations	BLEND: Descriptive scripts guide teacher actions, with selective explicit scripts containing exact words to use.	Rationale behind design decisions Mathematics concepts Student thinking and responses Ways to adapt the content for specific students
Math Expressions	BLEND: Detailed descriptive scripts and explicit guidance of teacher actions. [Rarely scripts teacher's words.]	Mathematics concepts Student thinking and responses Ways to adapt the content for specific students
Saxon	EXPLICIT SCRIPT: Fully scripted lesson; detailed description of teacher actions and room arrangements.	Classroom organization and management
SFAW/enVision	DESCRIPTIVE SCRIPT: Minimal Description of teacher actions.	Anticipate student errors Ways to adapt the content for specific students

Why the differences are important. The potential for curricula to influence teaching practices, which in turn, impact student achievement, is determined by two key factors: (1) the content of the curriculum and (2) the way the curriculum is implemented in the classroom. To examine the potential for each of the study curriculum to influence student achievement, we examined the differences in mathematical emphasis of each curriculum (as a measure of quality), the instructional approach of each curriculum (including the difficulty of implementing the approach), and the supports provided to the teachers to enact the curriculum as intended. Hypothesizing the potential for the curricula to affect student achievement is not an easy endeavor given the differences described above—each curriculum in the study has strengths and weaknesses.

Taking the mathematical emphasis and instructional approach together, Math Expressions has the greatest potential to provide students with opportunities to learn a wide range of mathematical skills and concepts. Math Expressions appears to emphasize conceptual understanding, mathematical thinking and reasoning, and procedural fluency. Saxon and Investigations have less potential than Math Expressions to address the necessary range of skills and concepts. Saxon's design is more likely to support procedural fluency, and Investigations is more likely to support mathematical thinking and reasoning. SFAW/enVision appears to have the least potential; it is limited in its attention to conceptual understanding and mathematical thinking. It places primary emphasis on procedural fluency but lacks routines and instructional interactions that might support fluency development.

Considering both the instructional approach of each curriculum and the support for teachers to implement the curriculum as intended, we find that Math Expressions offers a challenging curriculum to implement, but the widest range of support, including detailed descriptions and explicit guidance for the teacher along with information about mathematics, student thinking, and curriculum adaptations. The instructional approach in Math Expressions is not as challenging as it is in Investigations, where the mathematical emphasis and dialogic instructional approach are particularly challenging for teachers to enact. Investigations provides a wide range of information for the teacher but fewer explicit details than Math Expressions about how to use these resources. SFAW/enVision's instructional approach is easier to implement than that of Math Expressions and Investigations, but SFAW/enVision provides minimal support to teachers. Saxon is somewhat comparable to SFAW/enVision in terms of difficulty of implementing the instructional approach, but the script might make it the easiest curriculum to use. Although the range of support provided by Saxon is much narrower than the others, the extensive detail and script provided increase the likelihood that the curriculum will be enacted closer to the authors' intent in coverage and time.

b. Curriculum Training

Publishers provided study teachers with training on their assigned curriculum. Training was provided in the summer before each school year, and follow-up training was provided during the school year. Summer training consisted of group sessions held in each district, with separate training for each curriculum. Typically, summer training occurred two to four weeks before the first day of school. Prior to the first year each school participated, these sessions were initial training sessions. Prior to the second year each school participated, publishers provided initial training for new study teachers and refresher training to returning teachers. In addition, with the transition from Scott Foresman-Addison Wesley Math (SFAW) to enVision between the 2007–2008 and 2008–2009 school years, the publishers provided initial training to all teachers (including returning teachers) whose schools implemented enVision in the 2008–2009 school year. For teachers unable to attend the summer training sessions, publishers often scheduled make-up sessions just before or after the first day of school.

Investigations, Saxon, and SFAW/enVision each offered one day of initial training; Math Expressions offered two days. Each publisher provided follow-up training and support to teachers during both school years. Unlike the initial and refresher trainings, follow-up training was often provided one school or one teacher at a time; the structure of the training differed across and within curricula, and could have been provided through group sessions, classroom observations by trainers followed by brief feedback sessions with teachers, or demonstration lessons. Most trainers attempted to provide the first round of follow-up support within the first six weeks of each school year and then provided additional support at different intervals for each curriculum; in some cases, the additional training varied by school.

2. Recruiting Study Participants

The study team identified and recruited geographically dispersed districts and schools to participate in the study. Suitable districts had to include a minimum of four elementary schools (to support the study's design) and Title I schools (consistent with the policy interest that underlies Title I and its focus on effective approaches to help low-income children meet state standards for academic achievement).

Districts and schools volunteered to participate, as did all teachers and other relevant staff (such as math coordinators, math coaches, and supplemental teachers) at the relevant grade levels. A school was considered a participant when the study team received consent forms from all teachers in the target grade levels in the school. All teachers at the target grade levels in each school enrolled in the study. A teacher's initial enrollment in the study covered all potential years of participation. Throughout each study year, updated teacher lists were obtained to track teacher turnover and identify new math teachers at the target grade levels. The study team sent informational packets to any new teachers who had not yet enrolled in the study. All replacement teachers enrolled in the study.

A total of 111 schools from 12 districts enrolled in the study. Among these, 40 schools from four districts enrolled during the 2006–2007 school year (Group 1), and the remaining 71 schools from eight districts enrolled during the 2007–2008 school year (Group 2).

3. Random Assignment of Curricula to Schools

A blocked random assignment procedure was implemented in each district involving all four study curricula, which allocated similar numbers and types of schools, teachers, and students to each curriculum. Schools in each district were divided into blocks, where each block contained four to seven schools with similar baseline characteristics. For example, if a district contained eight schools, two blocks with four schools each were constructed. Random assignment of curricula to schools then took place within each block. The procedure helped minimize chance differences in school characteristics and sample sizes across curriculum groups, thus helping increase the design's statistical power and face validity. Agodini et al. (2008) provide more detail about the blocked random assignment procedure.

4. Site Participation and Curriculum Implementation during Years One and Two

Although a total of 111 schools from the 12 districts agreed to participate in the study, not all districts and schools participated for more than one year. As Table A.6 shows, three of the four Group 1 districts participated for a second year, and four of the eight Group 2 districts participated for a second year. As the table also shows, one Group 1 district participated in the study for three years and contains two cohorts of students who experienced the curricula in 1st and 2nd grades. The analyses include both of those student cohorts.

Table A.6. Number of Districts That Participated During Each Year of the Study

	Number of Districts by School Year		
	2006–2007	2007–2008	2008–2009
Group 1 districts	4	3	1
Group 2 districts	—	8	4

The analyses are based on the 58 schools from seven districts that participated in the study for two school years. The seven districts originally contained 66 schools that enrolled in the study. One of the 66 schools dropped out during the first year; 5 schools dropped out after the first year; and 2 schools refused to implement their assigned curriculum in the 2nd grade. Table A.7 illustrates curriculum implementation that occurred during each year of the study, separately for Group 1 and Group 2 districts.

Table A.7. Curriculum Implementation during Each Year of the Study

	Curriculum Implementation		
	2006–2007	2007–2008	2008–2009
Group 1 districts	1st grade	1st and 2nd grades	2nd grade[a]
Group 2 districts	—	1st and 2nd grades	1st and 2nd grades

[a] The Group 1 district that participated for three years also implemented the curricula in 3rd grade. This report does not present those data because several issues (such as attrition, small sample sizes, and crossover across treatment conditions) compromise the results based on those data.

Table A.8 presents information that is useful for understanding the types of schools that participated in the study. Compared to the average U.S. school, the study schools have a higher fraction of schoolwide Title I eligibility, students eligible for free or reduced-price meals, and minority students (Table A.8).[7]

[7] The Title I program provides financial assistance to schools with a high number or percentage of children from low-income families to help all students meet state academic standards. Title I–eligible schools have at least 35 percent of students from low-income families. Schools in which children from low-income families make up at least 40 percent of enrollment are eligible to use Title I funds for schoolwide programs that serve all children in the school.

Table A.8. Characteristics of U.S. Elementary Schools and Participating Schools

	U.S. Elementary Schools	Study Schools
Title I–eligible (percentage)[a]	71.4	82.8
Schoolwide Title I–eligible (percentage)	43.5	58.6
Urbanicity (percentage)		
City (large, midsize, and small)	28.3	36.2
Suburb or urban fringe	34.2	44.8
Town	8.2	5.2
Rural	29.4	13.8
Region[a] (percentage)		
Appalachia	6.2	8.6
Central	8.3	29.3
Midwest	19.7	12.1
Northeast and Mid-Atlantic	20.1	19.0
Pacific, Northwest, Southwest, and West	39.4	10.3
Southeast	12.1	20.7
Student Enrollment (average)		
1st grade	71	70
2nd grade	69	73
Students eligible for free or reduced-price meals (percentage)	46.9	63.3
Student gender (percentage)		
Male	51.8	52.4
Female	48.2	47.6
Student Race/Ethnicity (percentage)		
White	58.2	27.5
Non-Hispanic black	16.4	48.8
Hispanic	19.3	19.4
Asian	4.0	2.0
American Indian or Alaskan Native	2.2	2.4
Sample Size	**54,960**	**58**

Source: Author calculations using the 2005–2006 and 2006–2007 CCD. The “U.S. Elementary Schools” calculations include elementary schools with at least one 1st- or one 2nd-grade student. “Study Schools in Report” calculations include all schools in the Group 1 and Group 2 districts that participated for more than one year, except six schools within those districts. One school participated during part of the first school year and then stopped using its assigned curriculum (Math Expressions) and did not allow the study to collect follow-up data. The other five schools (three Investigations, one Math Expressions, and one Saxon) participated in the first year but not in the second.

[a] Regions are defined by IES’ Regional Educational Laboratory Program.

B. Data Collection

Below we provide information about the study's data collection activities. The data collection instruments are included in the study's design report (Agodini et al. 2008).

1. Student Testing

The study team administered to students the math assessment developed for the Early Childhood Longitudinal Study-Kindergarten Class of 1998–1999 (ECLS-K). The ECLS-K assessment is an adaptive test in that it is tailored to a student's achievement level. The test begins with a short, first-stage routing test used to broadly measure each student's achievement level. Depending on the score on the routing test, the student is then assigned to one of three longer second-stage tests: (1) an easy test, (2) a middle-difficulty test, or (3) a difficult test. Some items on the second-stage tests are identical across the second-stage tests, and this overlap across tests is used by IRT techniques (Lord 1980) that analyze patterns of correct and incorrect answers, to place scores from the different forms on the same scale to allow comparisons. The assessment includes both open-ended and multiple-choice questions designed to measure conceptual understanding, procedural knowledge, and problem solving in five content areas: (1) number sense, properties, and operations; (2) measurement; (3) geometry and spatial sense; (4) data analysis, statistics, and probability; and (5) patterns, algebra, and functions.

The ECLS-K K–1 math assessment was administered to students in the 1st grade. An ECLS-K math assessment for the 2nd grade did not exist; therefore, Mathematica worked with the Educational Testing Service (ETS), the developer of the ECLS-K, to create a 2nd-grade assessment by selecting appropriate items from existing ECLS-K math assessments (including the K–1, 3rd, and 5th grade instruments). ETS used information from the ECLS-K bridge study,[8] which included a small sample of second graders, combined with information about the current study's sample to ensure that the administered items appropriately targeted the estimated range of second graders' ability levels.[9] The study also used a Spanish version of the assessment for classes in which math instruction was conducted entirely in Spanish.

During 1st grade, students were tested in both the fall and spring; during 2nd grade, students were tested only in the spring. The test was administered during the school day by the study's field testers, who were trained and certified to administer the assessment to students.

Timing of the tests. Fall tests were administered within four weeks of the first day of classes; spring tests were administered one to six weeks before the end of the academic year. The test schedule aimed to administer the fall test as close as possible to the beginning of the school year and the spring test as close as possible to the end of the school year while keeping the average number of days between tests comparable across curriculum groups.

[8] The ECLS-K bridge study was conducted to ensure that item overlap between the ECLS-K, K-1, and ECLS-K 3rd-grade items was adequate to place student achievement in a longitudinal scale (Pollack et al. 2005).

[9] The present study has a relatively high proportion of children of low socioeconomic status, and test results for the study's fall 2006 1st-grade sample showed mean math ability slightly below that of national ECLS-K fall first graders, by about 1/8 of a standard deviation. The selection of items included in the 2nd-grade test accounted for these factors.

Student sampling. Before initiating student testing, trained field staff collected class rosters from each participating classroom. Field staff used a protocol to randomly select eligible students from the roster in each classroom for testing. The number of students sampled in each class was a function of the number of classrooms in the target grade levels and class size, so that both a fall and spring test was administered to an average of 30 eligible students at the target grade levels in each study school.[10]

Obtaining consent. In fall, parents of all students in study classrooms (regardless of whether they were sampled) received consent packets that included a letter and brochure describing the study and a consent form requesting permission to test the child and collect demographic data. Of the 12 study districts, 10 required passive consent, meaning that parents had to return signed forms only if they refused to permit testing of their child. The other 2 districts required active consent, meaning that parents had to return signed permission forms indicating their consent or refusal. Parents had at least one week to return the forms before testing began. In the spring, consent packets were distributed again to any parent who had not returned the form in the fall.

Parental consent was not a factor in determining whether a student could be sampled for testing. Therefore, some students selected for testing were not tested because of parental refusal.

Test scoring and reliability. Student answers on the assessment were sent to ETS for scoring.[11] A three-parameter IRT model was used to place the scores from the different second-stage tests on the same scale. The scores were calibrated using item parameters determined in the ECLS-K national analyses; specifically, the pooled K-1 and 3rd-grade ECLS-K item parameters were used.[12] This calibration supports comparisons between this study's sample and the ECLS-K national sample. Reliabilities were equal to 0.89 for the fall score and 0.92 for the spring score among 1st-grade students; these reliabilities are consistent with the national ECLS-K sample (Rock and Pollack 2002, pp. 5–7 through 5–9).[13] Reliabilities for the 2nd-grade students equal 0.88 for the fall score and 0.91 for the spring score; these reliabilities cannot be compared to the ECLS-K national sample, because a 2nd-grade assessment was not administered to the national sample. No floor or ceiling effects were observed in either the fall or spring scores for any of the analysis samples.

2. Classroom Observations

Members of the study team were trained to use a classroom observation protocol that captures elements of teacher instruction, student behavior, student-teacher interactions, and

[10] The study team collected class rosters before each subsequent round of testing to track student movement across classrooms and to identify any students who had moved out of the school since the baseline administration. Students who moved to another school participating in the study were administered the follow-up assessment at their new school. The study team did not collect follow-up data for students who moved to schools outside the study.

[11] ETS was a developer of the ECLS-K mathematics assessment.

[12] The 3rd-grade ECLS-K item parameters were used because many of the items on the study's 2nd-grade form were drawn from the ECLS-K 3rd-grade item pool, which includes all of the 3rd-grade items as well as the K–1 items.

[13] Reliabilities are based on the internal consistency (alpha) coefficients.

classroom activities related to math instruction. The protocol includes nearly 100 items that were either thought to be useful by the study team for discriminating the instructional approaches of the study's four curricula, or they were practices with prior evidence suggesting they are related to student achievement. The items were coded during one day's worth of math instruction in each study classroom that was observed in real time; this instruction included the math lesson and the morning meeting or calendar time, which was typically about 70 minutes per day, on average. About two-thirds of the items on the protocol were coded during the observation; the remaining items were coded immediately after the observation.

Observers were trained to use the protocol by watching multiple classroom videos and coding these behaviors, interactions, and activities. After coding each video, a master coder led a group discussion of the results to bring observers to a consensus on how to code each item. Observers were required to pass a certification test on the entire protocol prior to conducting observations in the field. To become certified, an observer had to code within one category of the master observer on 85 percent of the items in the protocol.[14]

Observers worked with teachers to schedule observations in advance and asked teachers to identify all points in time during the observation day when students were involved in math instruction. The observers then entered and exited the class as needed so they could be present at all times math instruction took place (such as the morning meeting or calendar time, the math lesson, and any subsequent math instruction, such as drills or activity at math centers). In some classrooms, observers were in the classroom for a single block of time; in others, they were in and out of the classroom numerous times throughout the day.

For 1st-grade classrooms, all observations took place in the spring (March–April). In Group 1 schools, attempts were made to observe all classrooms. In Group 2 schools, attempts were made to observe all English-speaking classrooms in schools with four or fewer classrooms. In Group 2 schools with more than four English-speaking classrooms, four were randomly sampled for observation. This sampling was conducted to keep the average number of observations per school consistent between groups 1 and 2. In Group 2 schools, attempts were also made to observe all Spanish-speaking classrooms.

For 2nd-grade classrooms, the observations were evenly distributed within each curriculum group across three points in the school year: fall (October–November), winter (January–February), and spring (March–April). Attempts were made to observe all classrooms in schools with seven or fewer classrooms. In schools with more than seven classrooms, seven were randomly sampled for observation.

About 10 percent of the classroom observations were simultaneously coded by two observers to assess item reliability. During these reliability observations, a master coder and classroom observer sat in the same classroom and independently observed all math instruction during the day of the observation. They completed and submitted the classroom observation

[14] Continuous (tallied) items in Sections A, B, C, and F were converted to the following seven categories: 0 (0 tallies), 1 (1–2 tallies), 2 (3–5 tallies), 3 (6–10 tallies), 4 (11–15 tallies), 5 (16–20 tallies), and 6 (21 or more tallies) for reliability assessments. Percentage agreement was calculated within one of these constructed categories for continuous items.

protocol separately and did not change any responses regardless of any similarities or differences in coding. These paired observations were assessed for reliability using the same methods used to certify observers during the observation training effort. Percentage agreement was calculated within 1 for all categorical and continuous items on the protocol. Exact agreement was required for dichotomous items. Items with inter-rater reliability below 75 percent were considered unreliable. All items, except one, met the study's criteria for inter-rater reliability (see Agodini et al. 2010).

The items presented in the table are a subset of the items on the observation protocol. The items selected for the table were considered by the study team to be those that are most closely aligned with the aspects of the curricula examined through the comparative curriculum analysis.

3. Other Data Collection

In addition to the classroom observation data, to help interpret measured effects, the study team conducted several other data collection efforts that are reported later in this appendix to set a context for the student achievement results:

- ***Assessment of Teacher Knowledge of Math Content and Pedagogy.*** Teacher math content knowledge and pedagogical knowledge were assessed at the initial teacher training sessions before the curricula were introduced, using an assessment developed by researchers at the University of Michigan.[15]

- ***Curriculum Training Received by Teachers.*** The study team took attendance at the initial teacher training sessions the publishers conducted before the start of the school year. Attendance at the followup sessions that occurred during the school year was recorded and provided by the publishers and was collected from teachers through the teacher surveys.

- ***Teacher Surveys.*** Two surveys were administered to teachers each year of the study. The fall survey focused on background information about the teacher, classroom characteristics, curriculum training provided by the publishers up to that point, and math instruction approaches used before joining the study. The spring survey gathered information on follow-up training provided by the publishers; use of the assigned curriculum and any other math curricula; and math instructional practices used during the year, including details about adherence to the teacher's assigned curriculum.

[15] The teacher assessment includes items about teacher pedagogical content knowledge in two major domains: (1) knowledge of mathematics for teaching and (2) knowledge of students and mathematics. Hill et al. (2004) provides details about the assessment's development process. The teacher assessment was scored using item response theory (IRT) techniques. An overall scale score and separate scores for content and pedagogical knowledge were calculated. Only the overall score was used in the analysis, because the reliability of one of the separate measures (the pedagogical score) was below an acceptable level. The reliability of the overall teacher assessment score for the study's sample equals 0.80.

- ***Student Characteristics from Class Rosters.*** The study team collected rosters for each classroom in the study to select the student sample. Student demographic information was also collected, including gender, date of birth, race/ethnicity, eligibility for free or reduced-price meals, whether the student had limited English proficiency or was an English language learner, and whether the student had an individualized education plan (IEP) or received special services (for students with a disability).

C. Construction of the Analysis File

Below, we describe students included in the analysis file, measures included in the file, and response rates to the measures.

1. Students Included in the Analysis File

The analysis sample includes students who were tested in the fall of their first year of study participation and the spring of their second year of participation. In nearly all cases, the analysis file includes students who were tested in the fall of 1st grade and spring of 2nd grade.[16] Students were linked to their teachers and schools during the fall 1st-grade assessment; the characteristics of these teachers and schools were measured as of the start of the school year.

The analysis sample contains 2,045 students sampled from 222 classrooms in 58 schools and seven districts. Below, we describe the selection and attrition that led to this sample.

Among districts that participated in the study for two years, Table A.9 shows the number of schools that enrolled in the first year, along with the number that participated for a second year. Figure A.1 shows the flow of districts and schools through the two years of the study, and Figure A.2 shows the flow of students through the two years of the study. Table A.10 shows the number of schools, classrooms, and students in the analysis sample (after accounting for all attrition).

Table A.9. School Attrition in Districts That Participated for Two Years, by Curriculum

		Schools by Curriculum			
	All Schools	Investigations	Math Expressions	Saxon	SFAW/ envision
Schools in first year (number)	66	17	16	15	18
Schools in second year (number)	58	14	14	12	18
Attrition rate (percentage)	12.1	17.6	12.5	20.0	0.0

[16] A small number of students repeated 1st grade, so the second-year spring test is a 1st-grade assessment.

Figure A.1. Flow of Districts and Schools Through the Two Years of the Study

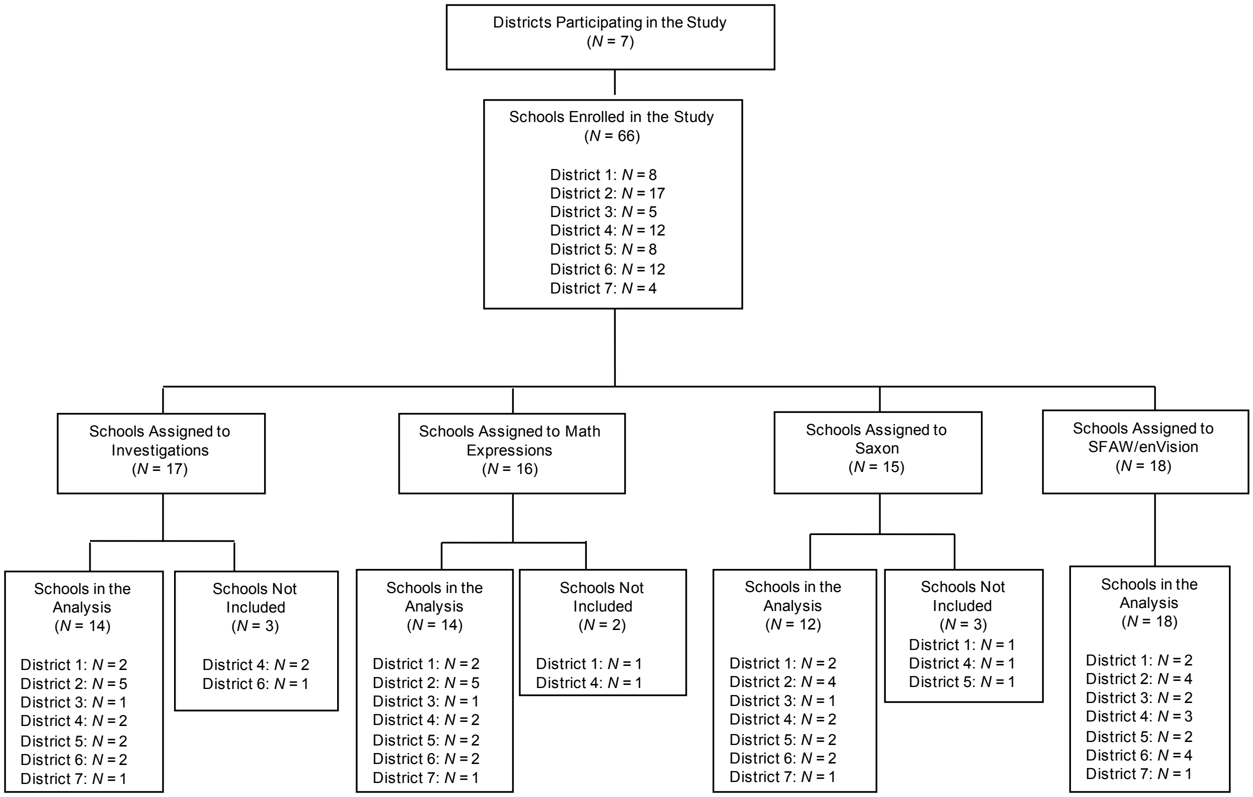

Figure A.2. Flow of Students Through the Two Years of the Study

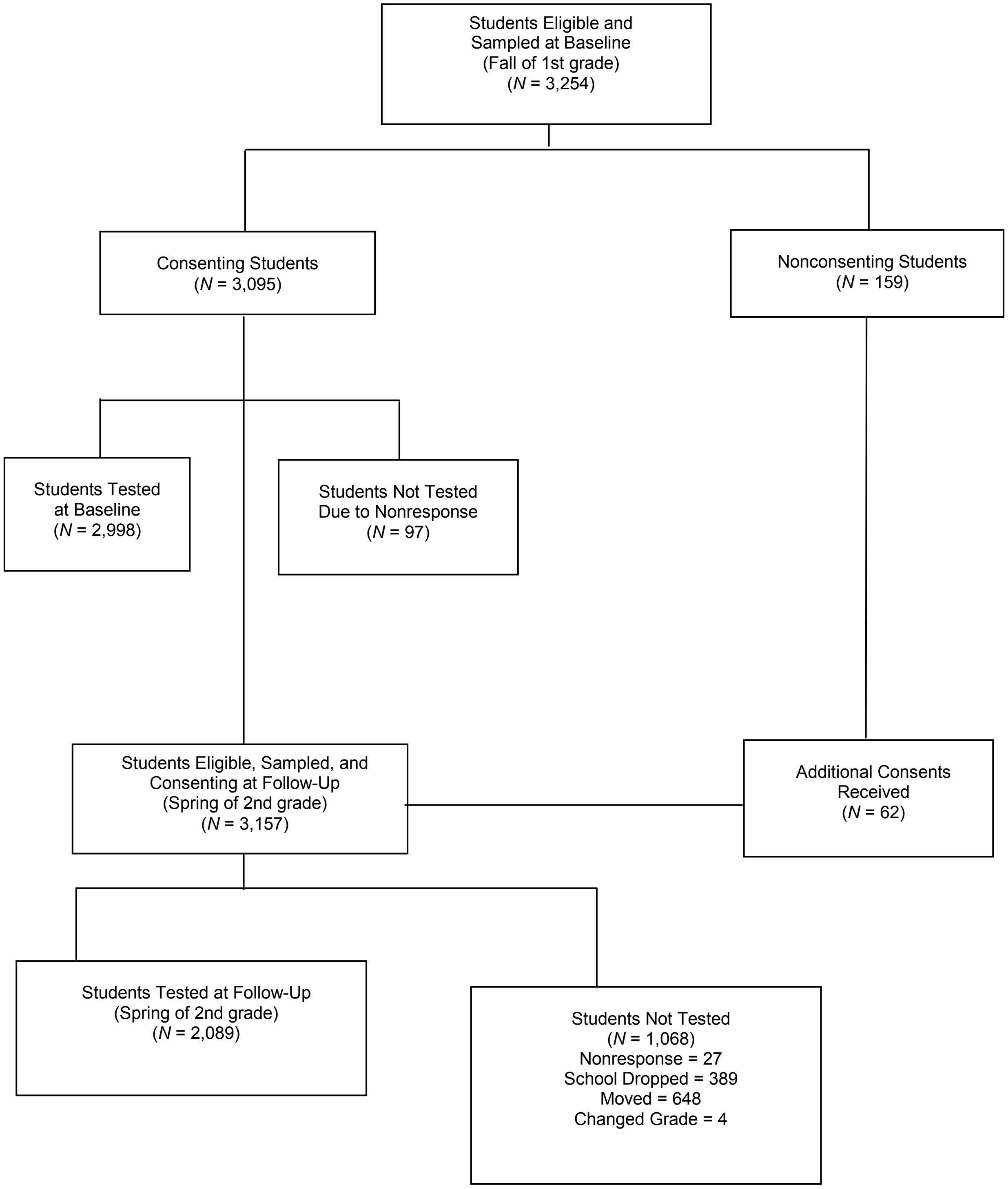
Students Eligible and Sampled at Baseline (Fall of 1st grade) (*N* = 3,254)
Consenting Students (*N* = 3,095)
Nonconsenting Students (*N* = 159)
Students Tested at Baseline (*N* = 2,998)
Students Not Tested Due to Nonresponse (*N* = 97)
Students Eligible, Sampled, and Consenting at Follow-Up (Spring of 2nd grade) (*N* = 3,157)
Additional Consents Received (*N* = 62)
Students Tested at Follow-Up (Spring of 2nd grade) (*N* = 2,089)
Students Not Tested (*N* = 1,068)
Nonresponse = 27
School Dropped = 389
Moved = 648
Changed Grade = 4

Table A.10. Number of Schools, Classrooms, and Students in the Analysis, by Curriculum

		Samples by Curriculum			
	All	Investigations	Math Expressions	Saxon	SFAW/ envision
Schools	58	14	14	12	18
Classrooms	222	45	54	54	69
Students	2,045	430	479	452	684

2. Measures Included in the Analysis File

The analysis files contain student-, teacher-, and school-level measures. Students in each analysis file were linked to their teachers and schools during the fall assessment; the characteristics of these teachers and schools were measured as of the start of the school year. Student-level math test scores were obtained from ETS, which scored the assessments and created a scale score based on IRT techniques. School records were used to construct other student-level measures included in the analysis files, including student demographics (age, gender, and race/ethnicity), whether the student is LEP or an ELL, and whether that student had an IEP or received special services. In addition, the analysis file includes the number of days between the beginning of school and the fall assessment and the number of days between the fall and spring assessments.

Teacher-level measures were obtained from the assessment of math content and pedagogical knowledge and the fall teacher survey. Teacher experience, education, race/ethnicity, and prior use of the assigned curriculum at the K–3 level were obtained from the fall teacher survey. Classroom size was obtained from class rosters. To measure the heterogeneity of the students in the classroom, the classroom variance and skewness of the fall student math score were computed.

School-level measures were obtained from the CCD and study records. Two school-level measures were extracted from the CCD: (1) the percentage of students eligible for free or reduced-price meals, and (2) whether the school was Title I eligible. In addition, the analysis files included the block into which the school was placed during the random assignment process, the curriculum assigned to the school, and the school district.

Imputing missing data. Complete data were available for the school-level measures and the fall and spring student math test scores. However, a small fraction of data was missing for some of the other student-level measures and for each of the teacher-level measures. Model-based imputations were used to replace missing data; the process is described in Agodini et al. (2010). Table A.11 provides the number of missing observations for each measure in the analysis, along with the pre- and post-imputation means.

Table A.11. Model-Based Imputation of Missing Data

Variable Name	N	Number Missing	Mean (Pre-Imputation)	Mean (Post-Imputation)
Student Level				
Fall math scale score	2,045	0	35.48	35.48
Age at fall test	2,037	8	6.55	6.55
Female	2,036	9	0.49	0.49
Race/ethnicity				
Hispanic	2,042	3	0.25	0.25
Non-Hispanic black	2,042	3	0.41	0.41
LEP/ELL	1,893	152	0.13	0.14
IEP/special services	1,882	163	0.10	0.09
Days between start of school and fall assessment	2,045	0	19.87	19.87
Days between assessments	2,045	0	604.63	604.63
Teacher Level				
Master’s degree	202	15	0.55	0.56
Experience	206	11	12.71	12.82
Prior use of the assigned curriculum	180	37	0.13	0.17
Race				
Black	199	18	0.16	0.16
Hispanic	199	18	0.13	0.13
Assessment				
Overall IRT score	210	7	-0.06	-0.07
Classroom Level				
Class size	222	0	20.25	20.25
Variance of fall score	222	0	118.44	118.44
Skewness of fall score	222	0	0.44	0.44

Source: Author calculations using the study-administered student assessment, student records, fall teacher survey, and the study-administered teacher assessment.

Weights. A sampling weight was developed for each student in each of the analysis files. In particular, students in each file were weighted up to the number of students who were eligible to be tested in the fall of 1st grade, separately for each classroom. For example, if 20 students in a classroom were eligible to be tested in the fall but only 12 could be tested in the fall and spring of their relevant follow-up period, each of the 12 students was assigned a weight of 1.67 (20/12). The weight was not adjusted for testing nonresponse because, with the extent of missing test data observed in the current study, other research shows simply analyzing students who were pre- and post-tested is equivalent to using weighting and imputation techniques that adjust for nonresponse (Puma et al. 2009).

3. Response Rates to the Data Collection

Teacher assessment and surveys. Table A.12 shows the response rates to the teacher knowledge assessment and the fall and spring surveys. The number of 2nd-grade classrooms differs from fall to spring, because some classrooms were added during the year (when overenrolled classes were divided).

Table A.12. Teachers Who Completed the Teacher Assessment and Surveys, by Grade and Curriculum

		Teacher Knowledge Assessment		Fall Teacher Survey		Spring Teacher Survey		
Curriculum	Classrooms	Number	Percentage	Number	Percentage	Classrooms	Number	Percentage
1st-Grade Teachers								
All Curricula	**222**	**215**	**97**	**213**	**96**	**222**	**202**	**91**
Investigations	45	44	98	45	100	45	43	96
Math Expressions	54	51	94	49	91	54	44	81
Saxon	54	52	96	51	96	54	51	94
SFAW	69	68	99	68	99	69	64	93
2nd-Grade Teachers								
All Curricula	**224**	**204**	**91**	**217**	**97**	**226**	**209**	**92**
Investigations	49	46	94	48	98	50	48	96
Math Expressions	52	43	83	47	90	52	44	85
Saxon	49	46	92	49	100	49	46	94
SFAW/enVision	74	69	95	73	99	75	71	95

Student assessment. Table A.13 reports the number of students who were sampled for testing in the fall of 1st grade, and the percentages that were tested in both the fall of 1st grade and spring of 2nd grade, by curriculum and overall.[17] Parent refusals accounted for approximately two-thirds of student nonresponse in the fall of 1st grade (derived from Figure A.2). At 2nd-grade spring testing, nearly one-third of the students were no longer in a study school or enrolled in a school that withdrew from the study (derived from Figure A.2).

School-level data. Two school-level measures were collected from the Common Core of Data (CCD): (1) the percentage of students eligible for free or reduced-price meals, and (2) whether the school was Title I eligible. In addition, the analysis included the block into which the school was placed during the random assignment process and the curriculum assigned to the school. All these measures are available for each school included in the analysis.

[17] As mentioned above, a small number of students repeated 1st grade, so the second-year spring test is a 1st-grade assessment.

Table A.13. Number of Students Sampled for Testing and Response Rates

Curriculum	Students Sampled for Testing in Fall	Number Tested in Fall	Number Tested in Both Fall and Spring	Percentage Tested in Both Fall and Spring
All Curricula	**3,254**	**2,998**	**2,045**	**63**
Investigations	784	715	430	55
Math Expressions	743	681	479	64
Saxon	782	724	452	58
SFAW/envision	945	878	684	72

Note: Spring response rates are based on follow-up data collected 1.5 years after baseline.

D. Curriculum Effects on Student Math Achievement

As described earlier, an experimental design was used to determine the effects of the study's four curricula on student math achievement. The design involved randomly assigning participating schools in each district to the study's four curricula.

1. Baseline Equivalence

Tables A.14A, A.15, and A.16A show the comparability of the curriculum groups along baseline school, teacher, and student characteristics. Tables A.14B and A.16B provide the pre-attrition baseline characteristics of the school and student samples.

2. Two-Year Effects on Student Math Achievement

Table A.17 presents average fall and spring math achievement of students in each curriculum group and the average gain (spring minus fall) score for each group. The fall test corresponds to fall of 1st grade. The spring test corresponds to spring of 2nd grade.

Model for estimating curriculum effects and statistical significance. To assess whether the differences in achievement between the curriculum groups are statistically significant, we used a statistical model that accounts for the nested structure of the data (students clustered in classrooms and classrooms clustered in schools). To help increase the precision of the estimates, we also included baseline values of measures that explain variation in spring achievement.

Table A.14A. Baseline School Characteristics for the Analysis Sample

		Schools by Curriculum				
	All Schools	Investigations	Math Expressions	Saxon	SFAW	*p*-Value
Title I–Eligible (Percentage)	82.8	78.6	92.9	83.3	77.8	0.68
Schoolwide Title I–Eligible (Percentage)	58.6	57.1	57.1	58.3	61.1	0.99
Students Eligible for Free or Reduced-price Meals (Percentage)	63.3	68.4	60.6	56.7	66.4	0.54
Student Enrollment (Average)						
1st grade	70	64	69	79	71	0.72
2nd grade	73	68	65	87	72	0.40
Student Gender (Percentage)						
Male	52.4	51.6	52.3	52.2	53.2	0.61
Female	47.6	48.4	47.7	47.8	46.8	0.61
Student Race/Ethnicity (Percentage)						
White	27.5	28.0	30.4	25.0	26.4	0.97
Non-Hispanic black	48.8	53.1	49.1	52.1	42.9	0.84
Hispanic	19.4	15.2	17.1	20.7	23.5	0.84
Asian	2.0	2.9	2.5	1.7	1.1	0.71
American Indian or Alaskan Native	2.4	0.1	0.1	0.5	6.1	0.48
Sample Size	**58**	**14**	**14**	**12**	**18**	

Source: Author calculations using the 2005–2006 and 2006–2007 CCD.

Note: The *p*-values are results from statistical tests that examine the joint equality of each school characteristic across curriculum groups. The statistical tests used regression models, regressing each school characteristic on an intercept, binary indicators for three of the four curricula, and an error term. The degrees of freedom used to calculate the statistical significance of the results were adjusted to reflect the number of blocks constructed when conducting random assignment.

Table A.14B. Baseline School Characteristics for the Pre-Attrition Sample

		Schools by Curriculum				
	All Schools	Investigations	Math Expressions	Saxon	SFAW	*p*-value
Title I–Eligible (Percentage)	83.3	82.4	87.5	86.7	77.8	0.86
Schoolwide Title I–Eligible (Percentage)	62.1	64.7	56.3	66.7	61.1	0.93
Students Eligible for Free or Reduced-price Meals (Percentage)	63.1	69.6	57.9	57.8	66.4	0.34
Student Enrollment (Average)						
1st grade	73	74	72	76	71	0.98
2nd grade	76	78	70	84	72	0.77
Student Gender (Percentage)						
Male	52.0	51.7	51.7	51.4	53.2	0.41
Female	48.0	48.3	48.3	48.6	46.8	0.41
Student Race/Ethnicity (Percentage)						
White	26.5	23.6	32.1	23.9	26.4	0.85
Non-Hispanic black	45.6	50.6	45.1	43.7	42.9	0.92
Hispanic	23.5	22.7	18.4	29.7	23.5	0.78
Asian	1.9	2.4	2.5	1.8	1.1	0.76
American Indian or Alaskan Native	2.5	0.6	1.9	0.9	6.1	0.47
Sample Size	**66**	**17**	**16**	**15**	**18**	

Source: Author calculations using the 2005–2006 and 2006–2007 CCD.

Note: The *p*-values are results from statistical tests that examine the joint equality of each school characteristic across curriculum groups. The statistical tests used regression models, regressing each school characteristic on an intercept, binary indicators for three of the four curricula, and an error term. The degrees of freedom used to calculate the statistical significance of the results were adjusted to reflect the number of blocks constructed when conducting random assignment.

Table A.15. Baseline Characteristics of 1st- and 2nd-Grade Teachers (Percentages Unless Stated Otherwise)

		Teachers by Curriculum				
	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value Comparing Curricula
Demographics						
Average Age						
1st grade	40.6	43.6	40.5	42.3	37.5	0.09
2nd grade	41.8	43.8	38.7	42.0	42.1	0.43
Female						
1st grade	91.7	93.3	95.9	86.8	91.3	0.78
2nd grade	92.3	95.9	90.0	91.8	91.8	0.71
Race						
1st Grade						
White	82.1	80.5	90.9	86.3	73.8	0.31
Other	17.9	19.5	9.1	13.7	26.2	
2nd Grade						
White	78.2	75.6	73.7	82.9	79.7	0.89
Other	21.8	24.4	26.3	17.1	20.2	
Hispanic						
1st grade	15.1	—	—	—	—	0.71
2nd grade	18.4	12.8	14.3	26.1	19.4	0.90
Average Years of Teaching Experience						
1st grade	12.7	15.0	12.3	13.1	10.7	0.31
2nd grade	13.0	15.2	13.2	12.5	11.8	0.59
Has a Regular or Standard Teaching Certificate						
1st grade	90.7	95.6	95.8	92.3	82.6	0.18
2nd grade	89.3	95.9	84.1	91.8	86.3	0.45
Education						
Highest Degree Earned						
1st Grade						
Bachelor's degree	43.8	38.1	48.9	35.3	50.0	0.35
Master's degree or higher	56.3	61.9	51.1	64.7	50.0	
2nd Grade						
Bachelor's degree	42.4	36.2	39.5	40.4	49.3	0.77
Master's degree or higher	57.6	63.8	60.5	59.6	50.6	
Field of Bachelor's Degree						
1st Grade						
Elementary education	62.9	83.3	56.5	60.0	56.7	0.44
Early childhood or K–12 education	14.1	—	—	—	—	
Other	22.9	—	—	—	—	
2nd Grade						
Elementary education	63.5	52.3	60.5	66.7	70.6	0.52
Early childhood or K–12 education	15.5	20.5	14.0	17.8	11.8	
Other	21.0	27.3	25.6	15.6	17.6	
Number of Math Education Courses Taken						
1st Grade						
None	4.9	—	—	—	—	0.56
One or two courses	59.3	—	—	—	—	
Three or more courses	35.8	41.5	26.7	31.4	41.8	

		Teachers by Curriculum				
	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value Comparing Curricula
2nd Grade						
None	3.0	—	—	—	—	0.59
One or two courses	46.3	—	—	—	—	
Three or more courses	50.7	43.5	50.0	48.9	57.4	
Number of Advanced Math Courses Taken						
1st Grade						
None	40.2	43.9	46.7	29.4	41.8	0.45
One or two courses	46.1	—	—	—	—	
Three or more courses	13.7	—	—	—	—	
2nd Grade						
None	38.9	37.0	33.3	43.5	40.6	0.67
One or two courses	41.9	50.0	50.0	30.4	39.1	
Three or more courses	19.2	13.0	16.7	26.1	20.3	
Sample Size						
1st grade	**214**	**45**	**48**	**52**	**69**	
2nd grade	**215**	**49**	**44**	**49**	**73**	
Teacher Assessment						
Average IRT Scale Score						
1st grade	-0.07	0.06	-0.18	-0.09	-0.05	0.78
2nd grade	0.00	-0.20	-0.22	0.19	0.13	0.18
Sample Size						
1st grade	**215**	**44**	**51**	**52**	**68**	
2nd grade	**204**	**46**	**43**	**45**	**70**	

Source: Author calculations using fall teacher survey data and the study-administered assessment of teacher math content and pedagogical knowledge.

Note: The statistical tests were conducted using two-level HLMs.

*An asterisk in the row labeled “1st Grade” or “2nd Grade” indicates that the measure is significantly different (at the 5 percent level) across the curriculum groups in that grade level; an asterisk in the row labeled “*p*-value comparing 1st and 2nd Grade” indicates that the measure is significantly different across curriculum groups and grade levels. Note that asterisks are not used to identify tests that were significantly different across grades for all teachers, or across grades within each curriculum group.

— Value is suppressed to protect respondent confidentiality.

Table A.16A. Baseline Characteristics of 2nd-Grade Students Who Experienced the Same Curriculum for Two Years

		Students by Curriculum				
	All Students	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value
Fall 1st-Grade Score (Average)	35.4	36.1	33.9	36.6	35.2	0.54
Age at Fall Test (Average)	6.5	6.5	6.6	6.5	6.5	0.32
Female (Percentage)	48.6	53.4	46.6	46.2	48.7	0.11
Race/Ethnicity (Percentage)[a]						
Hispanic	27.8	18.3	26.9	33.7	30.2	0.93
Non-Hispanic black	40.1	43.8	44.6	36.3	37.4	
Other Non-Hispanic	32.1	38.0	28.5	30.0	32.4	
LEP or ELL (Percentage)	16.6	13.4	11.4	21.8	18.5	0.54
Has IEP or Receives Special Services (Percentage)*	9.0	7.7	11.1	7.8	9.2	0.05
Days Between Start of School and Fall 1st-Grade Test (Average)	20	20	20	21	19	0.83
Days Between Fall 1st-Grade and Spring 2nd-Grade Tests (Average)	605	604	605	604	606	0.91
Sample Size	**2,045**	**430**	**479**	**452**	**684**	

Source: Author tabulations using school records data and the fall 1st-grade and spring 2nd-grade ECLS math test administered by the study.

Note: The *p*-values are results from statistical tests that examine the joint equality of each student characteristic across the curriculum groups. The statistical tests were conducted using three-level HLMs—see text in next section for a description of how these tests were conducted.

[a] Students classified as Hispanic in school records were coded as Hispanic regardless of race. Non-Hispanic students classified as black or black and other race were coded as non-Hispanic black. All other students were coded as other non-Hispanic.

Table A.16B. Baseline Characteristics of the Pre-Attrition Student Sample (Students who Should Have Experienced the Curriculum for Two Years)

		Students by Curriculum				
	All Students	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value
Fall 1st-Grade Score (Average)	35.0	34.8	33.8	36.2	35.0	0.53
Age at Fall Test (Average)	6.6	6.6	6.6	6.6	6.5	0.79
Female (Percentage)	48.7	50.6	47.6	48.6	48.1	0.65
Race/Ethnicity (Percentage)[a]						
Hispanic	29.5	26.3	26.9	35.8	28.9	0.97
Non-Hispanic black	40.3	42.9	45.4	36.6	37.3	
Other Non-Hispanic	30.2	30.9	27.8	27.6	33.8	
LEP or ELL (Percentage)	15.4	14.5	13.4	18.0	15.8	0.65
Has IEP or Receives Special Services (Percentage)*	8.8	7.4	9.2	7.8	10.3	0.44
Days Between Start of School and Fall 1st-Grade Test (Average)	20.5	20.9	20.0	22.0	19.5	0.78
Sample Size	**3,030**	**715**	**713**	**724**	**878**	

Source: Author tabulations using school records data and the fall 1st-grade ECLS math test administered by the study.

Note: The *p*-values are results from statistical tests that examine the joint equality of each student characteristic across the curriculum groups. The statistical tests were conducted using three-level HLMs—see text in next section for a description of how these tests were conducted.

[a] Students classified as Hispanic in school records were coded as Hispanic regardless of race. Non-Hispanic students classified as black or black and other race were coded as non-Hispanic black. All other students were coded as other non-Hispanic.

Table A.17. Average Student Math Scores, by Curriculum (Standard Deviations Are in Parentheses)

	Scale Score		
Curriculum	Fall 1st Grade	Spring 2nd Grade	Gain
Investigations	36.06 (13.04)	67.31 (18.47)	31.25 (13.38)
Math Expressions	33.90 (11.47)	67.99 (18.92)	34.09 (13.04)
Saxon	36.57 (12.83)	69.89 (16.75)	33.32 (12.27)
SFAW/enVision	35.24 (12.23)	68.87 (17.42)	33.63 (12.76)

Source: Author calculations using the study-administered student assessment.

In particular, a three-level hierarchical linear model (HLM) was used to estimate the relative effects of the study's curricula. The first (student) level of the HLM regressed the spring student 2nd-grade scale score on the following student characteristics:

- **Fall score**—Student scale score on the fall 1st-grade assessment.
- **Age**—Student age at the time of the fall 1st-grade assessment.
- **Gender**—Indicator of whether the student is female.
- **Race/ethnicity**—Indicators of whether the student is (1) Hispanic, or (2) non-Hispanic black. Non-Hispanic white students and non-Hispanic students of other races serve as the reference category.
- **LEP/ELL**—Student is limited English proficient or an English language learner.
- **IEP/special services**—Student has an individualized education plan or receives special services.
- **Days before fall assessment**—The number of days between the beginning of school and the student's fall assessment.
- **Days between assessments**—The number of days between the student's fall and spring assessments.

The second (classroom) level of the HLM regressed the intercept from the first-level equation on the following teacher characteristics:

- **Education**—Teacher has a master's degree. Teachers who do not have a master's degree, all of whom have a bachelor's degree, serve as the reference category.
- **Experience**—Years of teaching experience before the start of the first school year of the study.
- **Prior use of the assigned curriculum**—Teacher used the assigned curriculum at the K–3 level before joining the study.
- **Teacher assessment—**Teacher's overall scale score on the assessment of math content and pedagogical knowledge measured at baseline.

The third (school) level of the HLM regressed the intercept from the second-level equation on the following school characteristics:

- **Curriculum**—Indicators of whether the school was assigned to Investigations, Math Expressions, or Saxon. Schools assigned to Scott Foresman-Addison Wesley Math (SFAW) serve as the reference category.
- **Random assignment block**—Indicators for all but one of the blocks constructed for random assignment. Schools in the block without an indicator serve as the reference category.

Making pair-wise curriculum comparisons. With the four curricula included in the study, six unique pair-wise comparisons of effects can be made: (1) Investigations relative to Math Expressions, (2) Investigations relative to Saxon, (3) Investigations relative to SFAW, (4) Math Expressions relative to Saxon, (5) Math Expressions relative to SFAW, and (6) Saxon relative to SFAW. Because an SFAW indicator is not included in the model and thereby serves as the reference category, the coefficients on the Investigations, Math Expressions, and Saxon indicators indicate the effects of these curricula relative to SFAW. To make the pair-wise comparisons among Investigations, Math Expressions, and Saxon, the coefficients on the curriculum indicators are subtracted from one another.

The statistical significance of the curriculum differentials was calculated with and without adjusting for the six unique curriculum-pair comparisons that were made. For the multiple comparison adjustments, the Tukey-Kramer method was used to adjust the estimated *p*-values. When performing several statistical tests, the chance of finding a significant effect that is actually due to chance increases. For example, with the four curriculum groups in this study, six unique pair-wise comparisons can be made. If each comparison is made using a *t*-test with a 5 percent confidence level, then the probability that one of those six tests will be statistically significant, even when there are no real differences between groups, could be as high as [1 – (1-0.05)^6] = 26 percent. Tukey (1952) developed a method that adjusts for pair-wise comparisons by taking into account the dependencies between comparisons, while still maintaining a low probability of finding at least one false effect. Tukey (1953) and Kramer (1956) independently developed a modification that is appropriate for unequal sample sizes. The findings presented in the body of the report are based on the unadjusted tests.

HLM estimates. Table A.18 presents results for three specifications of the HLM: (1) a model that includes only the curriculum indicators and the block indicators used when conducting random assignment; (2) a model that adds the student's fall score to the first model; and (3) a model that adds the other student-, teacher-, and school-level controls to the second model. The results presented in the body of the report are based on the third model. The pattern of results for the curriculum indicators is similar across the second and third models, both of which contain students' fall scores.

Table A.19 presents the magnitude and statistical significance for the six unique pair-wise curriculum comparisons at each grade level. For example, the table presents the difference in average HLM-adjusted spring achievement between Investigations and Math Expressions students but not the opposite comparison, because the latter magnitude equals the former, just with the opposite sign.

The results are presented in effect-size units, which were calculated by dividing each pair-wise curriculum comparison by the pooled standard deviation of the spring score for the two curricula being compared; Hedges' g formula (with small-sample bias correction) was used to calculate the effect sizes. The *p*-value for each result was calculated with and without the Tukey-Kramer method to adjust for the six pair-wise comparisons.

Table A.18 Hierarchical Linear Model Estimates: Outcome Is Spring 2nd-Grade Math Scale Score

	Model Using Only Block Dummies		Model Using Only Fall Scale Score		Full Model	
Variable Name	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Student Level						
Intercept	73.94	2.11	33.97	2.12	49.89	38.43
Fall math scale score	.	.	0.98	0.02	0.96	0.02
Age at fall test	.	.	.	.	-3.73	0.63
Female	.	.	.	.	-2.79	0.50
Race/ethnicity						
Hispanic	.	.	.	.	-2.10	1.08
Non-Hispanic black	.	.	.	.	-5.19	0.91
LEP/ELL	.	.	.	.	-1.48	0.91
IEP/special services	.	.	.	.	-2.03	0.91
Days between start of school and fall assessment	.	.	.	.	-0.12	0.07
Days between assessments	.	.	.	.	0.03	0.06
Teacher Level						
Master's degree						
Experience	.	.	.	.	-0.02	0.04
Prior use of the assigned curriculum	.	.	.	.	-0.56	1.08
Teacher assessment overall score	.	.	.	.	0.11	0.45
School Level						
Curricula						
Investigations	-3.54	1.36	-4.38	1.21	-3.67	1.13
Math Expressions	-0.17	1.30	0.46	1.16	0.64	1.08
Saxon	1.15	1.30	0.03	1.17	0.00	1.07
Random assignment block						
Block 201	1.43	3.46	9.61	3.04	7.21	2.92
Block 202	5.80	2.73	7.63	2.47	5.67	2.43
Block 221	1.04	3.20	4.23	2.87	0.99	2.95
Block 222	3.93	2.70	8.76	2.46	5.97	2.50
Block 231	-13.47	3.16	-3.46	2.85	-4.84	2.70
Block 232	-12.17	2.42	-3.94	2.21	-3.82	2.09
Block 233	-13.05	2.50	-5.09	2.28	-6.45	2.13
Block 251	-10.00	3.13	-1.95	2.77	-1.15	2.80
Block 252	-5.99	2.90	0.02	2.59	0.50	2.54
Block 253	-3.94	2.98	1.16	2.65	0.13	2.55
Block 254	-6.83	2.66	-0.51	2.40	-0.59	2.36
Block 271	1.63	2.43	4.40	2.21	-0.69	2.41
Block 281	-16.07	2.78	-3.25	2.50	-4.15	2.62
Block 311	-12.75	2.86	-8.31	2.57	-6.57	2.31
Block 312	4.91	2.87	5.43	2.59	4.94	2.33
Residual Variance						
Student Level	250.38	.	123.87	.	117.17	.
Classroom Level	11.63	.	13.83	.	14.94	.
School Level	1.94	.	2.39	.	0.67	.

Source: Author calculations using the study-administered student assessment, student records, fall teacher survey, and the study-administered teacher assessment.

Table A.19. Difference Between Pairs of Curricula in Average HLM-Adjusted Spring Student Math Achievement (in Effect Sizes), Students Who Used Their Assigned Curricula for Two Years (*p*-Values in Parentheses)

	Effect of					
	Investigations Relative to			Math Expressions Relative to		Saxon Relative to
	Math Expressions	Saxon	SFAW/ envision	Saxon	SFAW/ enVision	SFAW/ enVision
At the End of 1st Grade						
Effect Size	-0.18*+	-0.15*	-0.06	0.03	0.13*	0.10
Unadjusted *p*-Value	(0.00)	(0.02)	(0.35)	(0.58)	(0.03)	(0.09)
Adjusted *p*-Value	(0.02)	(0.08)	(0.78)	(0.94)	(0.11)	(0.33)
At the End of 2nd Grade						
Effect Size	-0.23*+	-0.21*+	-0.21*+	0.04	0.04	0.00
Unadjusted *p*-Value	(0.00)	(0.00)	(0.00)	(0.59)	(0.56)	(1.00)
Adjusted *p*-Value	(0.00)	(0.02)	(0.01)	(0.95)	(0.93)	(1.00)

Source: Authors' calculations based on data from the spring 1st- and 2nd-grade ECLS-K math test administered by the study, school records, the fall teacher survey, and school-level data from the 2005–2006 CCD.

Note: Effect sizes were calculated by dividing each pair-wise curriculum comparison by the Pooled standard deviation of the spring scale score of the two curricula being compared; Hedges' *g* formula (with the correction for small-sample bias) was used to calculate effect sizes. The unadjusted *p*-values do not account for the six pair-wise curriculum comparisons presented in the figure, whereas the adjusted *p*-values, which were calculated using the Tukey-Kramer method, account for the comparisons.

*Indicates that the effect size is statistically significant at the 5 percent level, according to the unadjusted *p*-value.
+Indicates that the effect size is statistically significant at the 5 percent level, according to the adjusted *p*-value.

The results show that the math curriculum experienced by 2nd-grade students for two years mattered. At the end of 2nd grade, students taught using Math Expressions, Saxon, and SFAW/enVision scored about 0.22 standard deviations higher than students taught using Investigations; none of the other curriculum differentials is statistically significant. When looking at results based on the adjusted statistical tests, all three differentials remain significant.

Sensitivity analyses. We explored whether the results are sensitive to (a) the few schools that dropped out of the study and, therefore, had to be excluded from the analyses; and (b) the students who moved between study schools that used a different study curriculum. As described below, the pattern of results is robust to these analyses.

Examining Attrition. In this study, schools were randomly assigned to curricula by district. In addition, within each district, random assignment was conducted within blocks that contained schools with similar baseline characteristics. Therefore, each block represents a separate randomized controlled trial (RCT), and analyses based on any collection of blocks that did not experience school attrition would provide causal evidence of relative curriculum effects.

The seven districts that are the basis of the results originally enrolled 66 schools in the study, of which 58 participated for a second year and implemented the curricula in 2nd grade; thus, the results presented in the brief and this appendix are based on 58 schools.

We examined whether the results are sensitive to the school attrition. In particular, we calculate relative curriculum effects based on random assignment blocks that did not lose a school. Working with blocks that did not lose a school drops the sample size from 58 to 39 schools. The pattern of results is robust to this sensitivity analysis. As mentioned above, results based on all 58 schools indicate that average achievement of Math Expressions, Saxon, and SFAW/enVision students was not significantly different, but average achievement of each of these groups was about 0.22 standard deviations higher than that of Investigations students. When based on the 39 (of 58) schools in random assignment blocks that did not lose a school for this analysis, the results are similar in both magnitude and statistical significance—in particular, average achievement of Math Expressions, Saxon, and SFAW/enVision students was not significantly different, but average achievement of each of these groups ranged from 0.23 to 0.31 standard deviations higher than that of Investigations students.

Examining Crossovers. In a study of this kind, in which study schools within a single district are using four curricula, it is possible that students move between schools that are assigned to different curricula during and between school years. Of the 2,045 students included in the analysis, 113 (about 28 students per curriculum) moved to a different study school with a different curriculum between fall 1st-grade and spring 2nd-grade testing. The crossover rate was not significantly different across curriculum groups. Although analytic techniques can be used to correct results for crossovers, those techniques cannot be used in this setting because the number of crossovers is too low to support the analysis. To explore whether the results are affected by the crossovers, we deleted them from the sample and reestimated the model. The results are nearly identical to the results based on the full sample of 2,045 students.

Subgroup analyses. We examined whether curriculum effects differ along six baseline characteristics: (1) school fall achievement, (2) school-level information about student eligibility for free or reduced-price meals, (3) teacher education, (4) teacher experience, (5) teacher math content and pedagogical knowledge, and (6) teacher prior use of the assigned curriculum at the K–3 level.

Separate HLMs were estimated for each characteristic by expanding the HLM described above to include interactions between the curriculum indicators and the subgroups defined by the characteristic. For example, to examine whether curriculum effects differ along a baseline characteristic that divides the sample into two subgroups, the model was expanded to include eight third-level interactions for that characteristic.

Because the study was not designed with sufficient statistical power for the subgroup analyses, the results are best viewed as exploratory, possibly raising policy-relevant questions to examine in studies with sufficient statistical power to address the questions. As shown in Table A.20, according to the unadjusted statistical tests, only 2 of the 12 subgroups have at least one pair-wise curriculum differential that is statistically significant:

- For schools with previous midrange math achievement, the average math achievement of Math Expressions students was significantly higher than that of Investigations students.
- For schools with greater than 40 percent eligibility for free and reduced-price meals, the average math achievement of SFAW/enVision students was significantly higher than that of Investigations students.

Table A.20. Difference Between Pairs of Curricula in Average HLM-Adjusted Spring Student Math Achievement (in Effect Sizes), Students Who Used Their Assigned Curricula for Two Years, by Subgroups

	Differential Effect of Curricula Effect Size and (unadjusted *p*-value, adjusted *p*-value)					
	Investigations Versus Math Expressions	Investigations Versus Saxon	Investigations Versus SFAW/enVision	Math Expressions Versus Saxon	Math Expressions Versus SFAW/enVision	Saxon Versus SFAW/enVision
School Fall Achievement						
Lowest Third	-0.01 (0.97, 1.00)	-0.26 (0.22, 0.93)	-0.06 (0.75, 1.00)	-0.25 (0.11, 0.74)	-0.05 (0.70, 1.00)	0.21 (0.22, 0.93)
Middle Third	-0.35 (**0.04**, 0.40)	-0.06 (0.73, 1.00)	-0.24 (0.08, 0.60)	0.32 (0.12, 0.76)	0.13 (0.44, 1.00)	-0.19 (0.26, 0.96)
Highest Third	-0.22 (0.16, 0.86)	0.04 (0.80, 1.00)	-0.07 (0.62, 1.00)	0.27 (0.07, 0.57)	0.16 (0.32, 0.98)	-0.11 (0.42, 1.00)
School Free/Reduced-Price Meals Eligibility						
Up to 40 Percent Eligibility	-0.24 (0.28, 0.91)	0.10 (0.63, 1.00)	0.09 (0.66, 1.00)	0.35 (0.16, 0.73)	0.34 (0.18, 0.77)	-0.01 (0.97, 1.00)
Greater Than 40 Percent Eligibility	-0.19 (0.07, 0.46)	-0.22 (0.06, 0.38)	-0.24 (**0.03**, 0.22)	-0.01 (0.91, 1.00)	-0.03 (0.74, 1.00)	-0.02 (0.83, 1.00)
Teacher Education						
Less Than a Master's Degree	-0.13 (0.27, 0.88)	0.04 (0.76, 1.00)	-0.13 (0.26, 0.88)	0.18 (0.15, 0.70)	0.01 (0.92, 1.00)	-0.18 (0.15, 0.69)
Master's Degree or Higher	-0.16 (0.15, 0.70)	-0.14 (0.21, 0.80)	-0.13 (0.24, 0.85)	0.03 (0.81, 1.00)	0.04 (0.71, 1.00)	0.02 (0.89, 1.00)
Teacher Experience						
Up to Five Years	-0.20 (0.32, 0.93)	-0.08 (0.69, 1.00)	-0.21 (0.27, 0.88)	0.12 (0.38, 0.96)	-0.01 (0.96, 1.00)	-0.14 (0.28, 0.90)
More Than Five Years	-0.16 (0.10, 0.53)	-0.13 (0.18, 0.77)	-0.15 (0.10, 0.55)	0.03 (0.75, 1.00)	0.01 (0.89, 1.00)	-0.02 (0.85, 1.00)
Teacher Math Content/Pedagogical Knowledge						
Lowest Quintile	-0.13 (0.40, 0.97)	0.02 (0.93, 1.00)	-0.18 (0.26, 0.88)	0.15 (0.34, 0.94)	-0.04 (0.76, 1.00)	-0.21 (0.21, 0.82)
Other Quintiles	-0.16 (0.13, 0.64)	-0.09 (0.36, 0.95)	-0.12 (0.21, 0.81)	0.07 (0.49, 0.99)	0.04 (0.65, 1.00)	-0.03 (0.79, 1.00)
Teacher Previously Used Curriculum						
No Prior Use	-0.16 (0.11, 0.59)	-0.11 (0.30, 0.91)	-0.13 (0.18, 0.77)	0.06 (0.56, 1.00)	0.04 (0.71, 1.00)	-0.02 (0.82, 1.00)
Prior Use	--	--	--	--	--	--

Source: Authors' calculations based on data from the first-grade ECLS-K math tests administered by the study, school records, the fall teacher survey, and school-level data from the 2005–2006 CCD.

Notes: The table shows the effect size for each curriculum-pair comparison, with the unadjusted and adjusted *p*-values in parentheses. Underlined effect sizes are considered substantively important according to the What Works Clearinghouse.

Effect sizes were calculated by dividing each pair-wise curriculum comparison by the pooled standard deviation of the spring scale score of the two curricula being compared; Hedges' *g* formula (with the correction for small-sample bias) was used to calculate effect sizes. The results were produced using a three-level HLM. We adjusted the "adjusted *p*-values" using the Tukey-Kramer method for the number of curriculum-pair comparisons made among subgroups defined by each characteristic, but not for the number made in other subgroups; we did not make the same adjustment for "unadjusted *p*-values."

*Indicates that the effect size is statistically significant at the 5 percent level, according to the unadjusted *p*-value.

+Indicates that the effect size is statistically significant at the 5 percent level, according to the adjusted *p*-value.

— Value is suppressed to protect respondent confidentiality.

No other pair-wise differences for these two subgroups are statistically significant. When based on adjusted statistical tests, no curriculum differentials are statistically significant in any of the subgroups examined.

However, some differences between pairs of curricula were noticeably large, though not statistically significant. The What Works Clearinghouse defines effect sizes greater than or equal to 0.25 as substantively important and considers them in its rating of interventions. A few subgroup comparisons yielded differences larger than 0.25 standard deviations and distinct from the overall findings. For schools with lower levels of free and reduced-price meal eligibility, Math Expressions outperformed both SFAW/enVision and Saxon. For schools with previous mid- and high-range math achievement, Math Expressions outperformed Saxon.

Table A.21 presents school, teacher, and student sample sizes for each subgroup, along with the average value of the characteristic used to define each subgroup. For example, the cell for the "lowest third" under the "School Fall Achievement" subgroup indicates the average value of school fall achievement for the schools included in that subgroup. The table also presents the minimum detectable effect size (MDE) for each subgroup. The MDEs were calculated assuming that the sample is distributed evenly across the curriculum groups.

Table A.21. Sample Sizes Used in Student Experience Subgroup Analyses

		Sample Size			
Subgroup	Average Value of Subgroup Characteristic	Schools	Teachers	Students	Minimum Detectable Effect Size Between Any Pair of Curricula
School Fall Achievement[a]					
Lowest third	29.57	19	72	651	0.31
Middle third	35.11	19	66	573	0.27
Highest third	40.19	20	84	821	0.30
School Free/Reduced-Price Meals Participation					
Up to 40 percent eligibility	22.96	10	49	482	0.41
Greater than 40 percent eligibility	69.50	48	168	1,563	0.21
Teacher Education					
Bachelor's degree	--	23	96	859	0.22
Master's degree	--	35	121	1,186	0.21
Teacher Experience					
Up to five years	2.19	33	59	578	0.26
Greater than five years	16.79	57	158	1,467	0.18
Teacher Math Content/Pedagogical Knowledge[a]					
1st (lowest) quintile	-1.30	26	43	419	0.33
2nd through 5th quintiles	0.24	58	174	1,626	0.19
Teacher Previously Used Curriculum					
No prior use	--	50	180	1,696	0.19
Prior use	--	--	--	--	--

[a] School Fall Achievement and Teacher Math Content/Pedagogical Knowledge are expressed in scale score units.

Implications of the switch from SFAW to enVision. Because some students used SFAW in 1st grade and enVision in 2nd, the results for the fourth curriculum group (SFAW/enVision) are difficult to interpret. Therefore, we also examined results for the three Group 1 districts in which students in schools assigned to SFAW used the curriculum in both 1st and 2nd grades. As Table A.22 shows, students taught using SFAW in both 1st and 2nd grades scored about 0.58 standard deviations higher than students taught using Investigations; none of the other comparisons involving SFAW is statistically significant. The results are based on both unadjusted and adjusted statistical tests.

Table A.22. Difference Between Pairs of Curricula in Average HLM-Adjusted Spring 2nd-Grade Student Math Achievement (in Effect Sizes), Three Districts That Used SFAW for Both Years of the Study (*p-Value*s in Parentheses)

	Effect of					
	Investigations Relative to			Math Expressions Relative to		Saxon Relative to
	Math Expressions	Saxon	SFAW	Saxon	SFAW	SFAW
At the end of 2nd Grade						
Effect Size	-0.38*+	-0.35*+	-0.58*+	-0.06	-0.18	-0.26
Unadjusted *p*-Value	(0.01)	(0.02)	(0.00)	(0.66)	(0.11)	(0.06)
Adjusted *p*-Value	(0.03)	(0.07)	(0.00)	(0.97)	(0.35)	(0.20)

Source: Authors' calculations based on data from the spring 1st- and 2nd-grade ECLS-K math test administered by the study, school records, the fall teacher survey, and school-level data from the 2005–2006 CCD.

Note: Effect sizes were calculated by dividing each pair-wise curriculum comparison by the pooled standard deviation of the spring scale score of the two curricula being compared; Hedges' *g* formula (with the correction for small-sample bias) was used to calculate effect sizes. The unadjusted *p*-values do not account for the six pair-wise curriculum comparisons presented in the figure, whereas the adjusted *p*-values, which were calculated using the Tukey-Kramer method, account for the comparisons. Table A.10 reports the sample sizes at the three levels—that is, the number of schools, classrooms, and students.

*Indicates that the effect size is statistically significant at the 5 percent level, according to the unadjusted *p*-value.
+Indicates that the effect size is statistically significant at the 5 percent level, according to the adjusted *p*-value.

E. Curriculum Implementation

To help set the context for the achievement effects, we used the teacher survey data to understand how the 1st- and 2nd-grade teachers of the students in Figure A.3 implemented their school's assigned curriculum.

Tables A.23 through A.27 present implementation information that could be measured consistently across the curricula. For each measure, an overall mean and means by curriculum group are provided. To assess content coverage, the spring survey asked teachers to indicate the number of lessons they taught in each of 20 math content areas by responding to a series of questions with the following categorical answers: 0 (none; I did not teach this topic); 1 (1–5 lessons); 2 (6–10 lessons); 3 (11–15 lessons); or 4 (more than 15 lessons). Teachers reported the number of lessons taught in each content area, regardless of whether they used their assigned curriculum or other materials. Table A.27 presents the mean response for each content area. A mean of 3, for example, indicates that 11 to 15 lessons focused on that content area. The items are arranged from topics most frequently taught in year one—when all the curriculum groups are pooled together—to those least frequently taught.

Table A.23. Curricula Previously Used by Teachers (Percentages Unless Stated Otherwise)

		Teachers by Curriculum				
	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value Comparing Curricula
Used the Assigned Curriculum at the K–3 Level Prior to the Study						
1st grade*	14.4	—	—	—	—	0.01
2nd grade*	17.9	—	—	—	—	0.00
p-value comparing 1st and 2nd grade	0.35	na	na	na	Na	0.30
Taught Math in K–3 During the Previous Year						
1st grade	86.2	88.4	80.9	85.4	89.1	0.62
2nd grade*	90.7	93.3	100.0	84.4	87.5	0.00
p-value comparing 1st and 2nd grade*	0.17	0.44	0.00	0.90	0.76	0.00
Curriculum Used in Previous Year (among those who taught K–3 previously)						
1st grade						
Everyday Math	9.7	—	—	—	—	0.62
Harcourt Math	10.2	—	—	—	—	
Saxon Math	25.0	—	—	—	—	
SFAW Math	30.1	—	—	—	—	
Other	25.0	—	—	—	—	
2nd Grade						
Everyday Math	8.2	—	—	—	—	0.99
Harcourt Math	12.3	—	—	—	—	
Saxon Math	29.2	—	—	—	—	
SFAW Math	29.8	—	—	—	—	
Other	20.5	—	—	—	—	
p-Value Comparing 1st and 2nd Grade	0.41	0.35	0.50	0.82	0.23	na
Number of Years Used Prior Curriculum (among those who taught K–3 previously)						
1st grade	4.5	4.5	5.3	4.4	4.0	0.35
2nd grade	4.2	5.1	3.9	3.7	4.0	0.41
p-Value Comparing 1st and 2nd Grade	0.45	0.37	0.04	0.16	0.74	0.07
Sample Size						
1st grade	**203**	**43**	**47**	**48**	**65**	
2nd grade	**196**	**45**	**38**	**46**	**67**	

Source: Author calculations using fall teacher survey data.

Note: The statistical tests were conducted using two-level HLMs as described in the text.

*An asterisk in the row labeled “1st Grade” or “2nd Grade” indicates that the measure is significantly different (at the 5 percent level) across the curriculum groups; an asterisk in the row labeled “*p*-Value Comparing 1st and 2nd Grade” indicates that the measure is significantly different across curriculum groups and grade levels. Note that asterisks are not used to identify tests that were significantly different across grades for all teachers, or across grades within each curriculum group.

— Value is suppressed to protect respondent confidentiality.

na indicates that the data do not support a statistical test.

Table A.24. Initial and Refresher Teacher Training on the Assigned Curriculum (Percentages Unless Stated Otherwise)

		Teachers by Curriculum				
	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value Comparing Curricula
Attended Initial/Refresher Training						
1st grade*	90.7	100.0	91.7	89.6	84.6	0.00
2nd grade	84.7	82.6	80.4	81.3	91.3	0.35
p-value comparing 1st and 2nd grade*	0.07	0.00	0.15	0.26	0.26	0.00
Publisher-Specified Initial Training Length	1–2 days	1 day	2 days	1 day	1 day	
Publisher-Specified Refresher Training Length	0.5–1 day	0.5–1 day	0.5 day	0.5 day	0.5–1 day	
Number of Days Attended (Among Those Who Attended)						
1st grade*	1.2	1.0	2.0	1.0	1.0	0.00
2nd grade*	0.9	0.9	1.1	0.6	1.1	0.00
p-value comparing 1st and 2nd grade*	0.00	0.00	0.00	0.00	0.01	0.00
How Well Prepared After Training (among those who attended)						
1st grade*						
Very well	46.7	35.7	37.2	59.1	52.8	0.05
Adequate	35.7	57.1	34.9	15.9	35.8	
Somewhat or not at all	17.6	7.1	27.9	25.0	11.3	
2nd grade*						
Very well	47.0	66.7	51.6	40.5	36.7	0.00
Adequate	40.9	—	—	—	—	
Somewhat or not at all	12.2	—	—	—	—	
p-value comparing 1st and 2nd grade*	0.17	0.00	0.42	0.13	0.43	0.00
Sample Size						
1st grade	**205**	**43**	**48**	**49**	**65**	
2nd grade	**209**	**46**	**46**	**48**	**69**	

Source: Author calculations using data from the fall teacher surveys.

Note: The statistical tests were conducted using two-level HLMs as described in the text.

*An asterisk in the row labeled "1st grade" or "2nd grade" indicates that the measure is significantly different (at the 5 percent level) across the curriculum groups; an asterisk in the row labeled "*p*-value comparing 1st and 2nd grade" indicates that the measure is significantly different across curriculum groups and grade levels. Note that asterisks are not used to identify tests that were significantly different across grades for all teachers, or across grades within each curriculum group.

— Value is suppressed to protect respondent confidentiality.

Table A.25. Follow-Up and Total Teacher Training on the Assigned Curriculum (Percentages Unless Stated Otherwise)

		Teachers by Curriculum				
	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value Comparing Curricula
			Follow-Up Training			
Follow-up Training Available						
1st grade	92.5	93.0	93.2	85.4	96.9	0.55
2nd grade*	76.7	82.6	90.0	34.1	90.9	0.00
p-value comparing 1st and 2nd grade	0.00	0.17	0.63	0.00	0.20	0.09
Participated in Follow-up Training						
1st grade	86.9	88.1	88.4	76.0	93.7	0.30
2nd grade*	62.0	68.9	75.6	17.4	79.4	0.00
p-value comparing 1st and 2nd grade*	0.00	0.05	0.19	0.00	0.04	0.05
Number of Follow-Up Days Attended (Among Those Who Attended)						
1st grade*	1.5	2.9	0.5	0.4	1.9	0.00
2nd grade*	1.3	1.7	0.5	0.3	1.7	0.00
p-value comparing 1st and 2nd grade*	0.00	0.00	0.57	0.72	0.13	0.00
Sample Size						
1st grade	200	43	44	49	64	
2nd grade	200	45	41	46	68	
			Total Training			
Attended any training						
1st grade*	97.6	100.0	97.9	98.0	95.5	0.00
2nd grade	93.8	93.5	95.6	89.1	95.9	0.49
p-value comparing 1st and 2nd grade*	0.07	0.00	0.56	0.12	0.89	0.00
Total Days Attended (Among Those Who Attended)						
1st grade*	2.3	3.1	2.3	1.2	2.6	0.00
2nd grade*	1.6	2.1	1.3	0.7	2.2	0.00
p-value comparing 1st and 2nd grade*	0.00	0.00	0.00	0.00	0.09	0.03
Sample Size						
1st grade	**209**	**45**	**47**	**51**	**66**	
2nd grade	**210**	**46**	**45**	**46**	**73**	

Source: Author calculations using data from the fall and spring teacher surveys.

Note: The statistical tests were conducted using two-level HLMs as described in the text.

*An asterisk in the row labeled “1st grade” or “2nd grade” indicates that the measure is significantly different (at the 5 percent level) across the curriculum groups; an asterisk in the row labeled “*p*-value comparing 1st and 2nd grade” indicates that the measure is significantly different across curriculum groups and grade levels. Note that asterisks are not used to identify tests that were significantly different across grades for all teachers, or across grades within each curriculum group.

Table A.26. Teacher-Reported Instruction in the Spring (Percentages Unless Stated Otherwise)

		Teachers by Curriculum				
	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-value Comparing Curricula
Used Assigned Curriculum As Core Curriculum						
1st grade*	99.5	100.0	97.7	100.0	100.0	0.00
2nd grade*	98.6	95.8	100.0	100.0	98.6	0.00
p-value comparing 1st and 2nd grade*	1.00	0.00	0.00	1.00	0.00	0.00
Average Preparation Per Week (Hours)						
1st grade	2.6	2.6	2.6	2.5	2.7	0.93
2nd grade	2.5	2.7	2.8	2.5	2.2	0.60
p-value comparing 1st and 2nd grade	0.62	0.82	0.83	1.00	0.12	0.61
Hours Per Week of Math Instruction (Average)[a]						
1st grade	5.2	5.2	5.1	5.6	5.1	0.28
2nd grade*	5.4	5.1	5.3	6.1	5.2	0.03
p-value comparing 1st and 2nd grade	0.26	0.45	0.23	0.14	1.00	0.28
Percentage of Time Spent Practicing Math Procedures and Recall of Math Facts						
1st grade*	32.8	21.0	42.1	34.0	33.5	0.04
2nd grade	33.1	26.2	38.5	38.6	30.8	0.33
p-value comparing 1st and 2nd grade	0.91	0.32	0.58	0.27	0.42	0.43
Completed at Least 80 Percent of Lessons from Assigned Curriculum						
1st grade	85.1	86.0	81.8	84.6	87.1	0.96
2nd grade	71.5	56.3	79.1	80.4	71.4	0.14
p-value comparing 1st and 2nd grade	0.00	0.01	0.59	0.65	0.05	0.27
Supplemented the Assigned Curriculum with Other Materials						
1st grade	31.5	20.9	38.6	36.5	29.7	0.42
2nd grade	42.4	41.3	46.5	41.3	41.4	0.92
p-value comparing 1st and 2nd grade	0.02	0.03	0.26	0.71	0.24	0.43
Frequency of Supplementation[b]						
1st grade*						
At least once per week	77.4	—	—	—	—	0.00
Twice per month or less	22.6	—	—	—	—	
2nd grade						
At least once per week	53.9	52.9	63.2	56.3	45.8	0.73
Twice per month or less	46.1	47.1	36.8	43.8	54.2	
p-value comparing 1st and 2nd grade	0.01	0.00	0.09	0.92	0.13	0.44
Reasons for Supplementation[b]						
1st grade						
Remediation with a small group	39.1	—	—	—	—	na
Remediation with the entire class	32.8	—	—	—	—	
Enrichment with a small group	25.0	—	—	—	—	
Enrichment with the entire class	50.0	—	—	—	—	

Table A.26 *(continued)*

		Teachers by Curriculum				
	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value Comparing Curricula
As a replacement for selected units or lessons	12.5	—	—	—	—	
As a supplement for selected units or lessons	64.1	—	—	—	—	
Other	26.6	—	—	—	—	
2nd Grade						
Remediation with a small group	31.0	—	—	—	—	
Remediation with the entire class	29.9	—	—	—	—	
Enrichment with a small group	31.0	—	—	—	—	
Enrichment with the entire class	33.3	—	—	—	—	
As a replacement for selected units or lessons	10.3	—	—	—	—	
As a supplement for selected units or lessons	63.2	—	—	—	—	
State standards	28.7	—	—	—	—	
Other	20.7	—	—	—	—	
p-value comparing 1st and 2nd grade	na	na	na	na	na	na
Materials Used for Supplementation[b]						
1st Grade						
Saxon Math	13.1	—	—	—	—	0.99
Teacher-created materials	24.6	—	—	—	—	
Other curriculum materials	24.6	—	—	—	—	
Other supplemental materials	37.7	—	—	—	—	
2nd Grade						
Saxon Math	9.9	—	—	—	—	0.39
Teacher-created materials	22.2	—	—	—	—	
Test prep materials	9.9	—	—	—	—	
Other curriculum materials	11.1	—	—	—	—	
Other supplemental materials	46.9	—	—	—	—	
p-value comparing 1st and 2nd grade	na	na	na	na	na	na
Likelihood of Using Assigned Curriculum Again						
1st grade						
Very likely	52.5	58.1	51.2	46.2	54.7	0.16
Likely	33.2	30.2	23.3	34.6	40.6	
Not at all likely	14.4	11.6	25.6	19.2	4.7	
2nd grade						
Very likely	39.0	27.7	37.2	60.9	33.3	0.30
Likely	30.7	31.9	37.2	21.7	31.9	
Not at all likely	30.2	40.4	25.6	17.4	34.8	
p-value comparing 1st and 2nd grade*	0.00	0.01	0.92	0.81	0.00	0.01
Sample Size						
1st grade	**203**	**43**	**44**	**52**	**64**	
2nd grade	**208**	**48**	**43**	**46**	**71**	

Source: Author calculations using spring teacher survey data.

Note: The statistical tests were conducted using two-level HLMs as described in the text.

Table A.26 *(continued)*

[a] Teachers reported the number of days per week and number of minutes per day devoted to math instruction. The study team used the information to construct a measure of the hours per week spent on math instruction.

[b] Percentage calculated among teachers who reported supplementing.

*An asterisk in the row labeled “1st grade” or “2nd grade” indicates that the measure is significantly different (at the 5 percent level) across the curriculum groups; an asterisk in the row labeled “*p*-value comparing 1st and 2nd grade” indicates that the measure is significantly different across curriculum groups and grade levels. Note that asterisks are not used to identify tests that were significantly different across grades for all teachers, or across grades within each curriculum group.

— Value is suppressed to protect respondent confidentiality.

na indicates that the data do not support a statistical test.

Table A.27. Teacher-Reported Average Number of Lessons Taught in Various Math Content Areas

		Teachers by Curriculum				
Topic of Lesson[a]	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value
Adding and Subtracting with Whole Numbers						
1st Grade	3.56	3.55	3.61	3.85	3.29	*0.10*
2nd Grade*	3.42	3.47	3.44	3.80	3.13	*0.03*
p-Value Comparing 1st and 2nd Grade	0.08	0.60	0.26	0.97	0.16	*0.86*
Word Problems						
1st Grade*	3.33	3.37	3.57	3.65	2.86	*0.02*
2nd Grade*	3.37	3.45	3.76	3.73	2.86	*0.00*
p-Value Comparing 1st and 2nd Grade	0.39	0.86	0.16	0.34	0.96	*0.49*
Addition and Subtraction Facts with Whole Numbers						
1st Grade	3.40	2.90	3.52	3.81	3.32	*0.08*
2nd Grade*	3.35	3.28	3.45	3.87	3.01	*0.03*
p-Value Comparing 1st and 2nd Grade	0.17	0.80	0.54	0.88	0.14	*0.70*
Counting with Whole Numbers						
1st Grade	3.37	3.40	3.39	3.69	3.05	*0.12*
2nd Grade*	2.83	3.34	2.74	3.55	2.09	*0.00*
p-Value Comparing 1st and 2nd Grade	0.00	0.78	0.04	0.37	0.00	*0.28*
Understanding Numbers Less than 10						
1st Grade	2.97	2.93	3.02	3.42	2.59	*0.14*
2nd Grade*	2.31	2.84	2.14	2.98	1.64	*0.00*
p-Value Comparing 1st and 2nd Grade	0.00	0.79	0.01	0.17	0.00	*0.10*
Creating, Continuing, or Predicting Patterns						
1st Grade	2.74	2.60	2.77	3.25	2.38	*0.20*
2nd Grade*	2.48	2.57	2.07	3.56	1.96	*0.00*
p-Value Comparing 1st and 2nd Grade	0.13	0.88	0.09	0.30	0.08	*0.12*
Collecting or Analyzing Data						
1st Grade	2.56	3.00	2.50	2.65	2.24	*0.12*
2nd Grade*	2.49	2.85	2.56	2.87	1.91	*0.02*
p-Value Comparing 1st and 2nd Grade	0.70	0.88	1.00	0.15	0.63	*0.53*
Graphs						
1st Grade	2.64	2.60	2.68	3.08	2.27	*0.11*
2nd Grade*	2.47	2.64	2.60	3.16	1.83	*0.01*
p-Value Comparing 1st and 2nd Grade	0.49	0.62	0.67	0.76	0.17	*0.58*
Money						
1st Grade*	2.55	1.49	3.02	3.35	2.29	*0.00*
2nd Grade	2.79	3.00	2.65	3.51	2.26	*0.14*
p-Value Comparing 1st and 2nd Grade*	0.00	0.00	0.32	0.92	0.37	*0.00*
Place Value with Whole Numbers						
1st Grade*	2.34	1.35	2.45	2.94	2.48	*0.04*
2nd Grade	2.76	2.77	2.69	3.31	2.46	*0.13*
p-Value Comparing 1st and 2nd Grade*	0.00	0.00	0.17	0.03	0.73	*0.02*

Table A.27 *(continued)*

		Teachers by Curriculum				
Topic of Lesson[a]	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value
Geometric Shapes or Spatial Relationships						
1st Grade	2.40	2.86	2.14	2.40	2.27	*0.18*
2nd Grade*	2.26	2.72	2.05	2.73	1.77	*0.01*
p-Value Comparing 1st and 2nd Grade	0.59	0.90	0.45	0.04	0.18	*0.07*
Time						
1st Grade*	2.31	1.44	2.18	2.88	2.52	*0.00*
2nd Grade*	2.49	2.89	2.56	2.87	1.97	*0.00*
p-Value Comparing 1st and 2nd Grade*	0.17	0.00	0.15	0.97	0.01	*0.00*
Measurement with Standard Tools						
1st Grade	1.72	1.35	1.77	2.37	1.40	*0.17*
2nd Grade*	2.05	1.26	2.21	2.96	1.90	*0.00*
p-Value Comparing 1st and 2nd Grade	0.06	0.50	0.10	0.10	0.18	*0.23*
Nonstandard Measurement						
1st Grade	1.48	2.07	1.30	1.67	1.05	*0.15*
2nd Grade	1.54	1.34	1.50	1.89	1.46	*0.33*
p-Value Comparing 1st and 2nd Grade*	0.36	0.05	0.42	0.09	0.26	*0.04*
Fractions						
1st Grade*	1.61	0.70	1.93	1.85	1.83	*0.01*
2nd Grade*	1.87	1.96	1.12	2.93	1.59	*0.00*
p-Value comparing 1st and 2nd Grade*	0.02	0.00	0.02	0.00	0.73	*0.00*
Probability						
1st Grade	1.11	1.00	1.27	1.13	1.05	*0.47*
2nd Grade	1.22	1.17	1.16	1.51	1.10	*0.64*
p-Value Comparing 1st and 2nd Grade	0.50	0.68	0.41	0.04	0.82	*0.12*
Multiplying and Dividing with Whole Numbers						
1st Grade	0.16	0.19	0.20	0.16	0.13	*0.97*
2nd Grade*	0.90	0.87	0.69	1.80	0.47	*0.00*
p-Value Comparing 1st and 2nd Grade*	0.00	0.00	0.04	0.00	0.03	*0.01*
Multiplication and Division Facts with Whole Numbers						
1st Grade	0.10	0.14	0.18	0.06	0.05	*0.72*
2nd Grade*	0.84	0.76	0.67	1.80	0.36	*0.00*
p-Value Comparing 1st and 2nd Grade*	0.00	0.01	0.02	0.00	0.01	*0.02*
Decimals						
1st Grade	0.15	0.16	0.34	0.04	0.10	*0.20*
2nd Grade*	0.63	0.35	0.71	1.29	0.34	*0.00*
p-Value Comparing 1st and 2nd Grade*	0.00	0.04	0.06	0.00	0.04	*0.05*
Percentages						
1st Grade	0.08	0.09	0.19	0.02	0.05	*0.25*
2nd Grade	0.22	0.13	0.23	0.14	0.29	*0.52*
p-Value Comparing 1st and 2nd Grade	0.01	0.17	0.82	0.28	0.04	*0.45*

Table A.27 *(continued)*

		Teachers by Curriculum				
Topic of Lesson[a]	All Teachers	Investigations	Math Expressions	Saxon	SFAW/ enVision	*p*-Value
Sample Size						
1st grade	**202**	**43**	**44**	**52**	**63**	
2nd grade	**205**	**47**	**43**	**45**	**70**	

Source: Author calculations using spring teacher survey data.

Note: The statistical tests were conducted using two-level HLMs as described in the text.

[a] Possible responses: 0 (none), 1 (1–5 lessons), 2 (6–10 lessons), 3 (11–15 lessons), and 4 (more than 15 lessons). A mean of 4 indicates that teachers covered at least 15 lessons in the content area.

*An asterisk in the row labeled “1st grade” or “2nd grade” indicates that the measure is significantly different (at the 5 percent level) across the curriculum groups in that grade level; an asterisk in the row labeled “*p*-Value Comparing 1st and 2nd Grade” indicates that the measure is significantly different across curriculum groups and grade levels. Note that asterisks are not used to identify tests that were significantly different across grades for all teachers, or across grades within each curriculum group.

In addition, for each implementation measure, the results of eight statistical tests are presented. The first two tests (that appear in the last column of each table, for the rows labeled "1st Grade" and "2nd Grade") indicate whether the implementation measures are significantly different across curriculum groups within each grade level. A two-level hierarchical linear model (HLM) was used to calculate these tests, which accounted for the clustering of teachers within schools. The other six tests (that appear in the rows labeled "*p*-value comparing 1st and 2nd Grade") indicate whether the implementation measures differ significantly across grades, for all teachers combined, for teachers in each curriculum group, and whether the curriculum group differentials within each grade differ across grades. A two-level HLM was used to calculate these tests, which accounted for the clustering of teachers within schools. The first (teacher-level) equation regressed each implementation measure on an intercept and a teacher-level error term; the second (school-level) equation regressed the intercept from the first equation on an intercept, binary indicators for three of the four curricula, a binary indicator for grade, interaction terms between each curriculum indicator and the grade indicator, and a school-level error term. As above, the degrees of freedom used to calculate the statistical significance of the results were adjusted to reflect the information (number of blocks to which schools were assigned) used to conduct random assignment. The 5 percent level of confidence was used to determine statistical significance.[18]

Table A.28 presents information about curriculum adherence based on measures that are specific to each curriculum.

F. Effects of Switching Curricula

Before entering the study, some schools were using either Saxon or SFAW and, upon entrance, were randomly assigned to either continue using their pre-study curriculum or switch to one of the other study curricula. These schools enable us to examine how staying with Saxon and SFAW, instead of switching to another study curriculum, affects student achievement during the first year a new curriculum is used. We cannot examine the switching-staying issue for Investigations, because only two schools were using that curriculum before joining the study and both schools were assigned to continue using Investigations. We also cannot examine the switching-staying issue for Math Expressions, because no study school was using that curriculum before joining the study.

[18] To assess whether content coverage differs across the four curriculum groups and across the two grades, we were hoping to estimate a two-level hierarchical logistic regression for each of the five-category responses that account for the clustering of teachers in schools and includes interaction terms for curriculum and grade. Unfortunately, there were issues with model convergence for many of the five-category responses. Therefore, we used a different approach for this analysis, in which each content area was transformed to a binary measure, where teachers who reported a value equal to or above the mean were coded as 1, and teachers who reported below the mean were coded as 0. We then estimated the statistical model described above—a two-level hierarchical logistic regression—to assess whether the binary measure for each content area differed across the curriculum groups and grades.

Table A.28. Summary of 1st- and 2nd-Grade Teachers' Reported Curriculum Adherence

	Teachers by Curriculum			
	Investigations	Math Expressions	Saxon	SFAW/envision
All Teachers, by Grade				
Number of Features in Adherence Measure	15	14	12	8
Percentage of Features Implemented by Teacher with Expected Frequency				
1st Grade				
0–50	21.1	31.7	20.8	21.7
51–75	31.6	31.7	31.3	36.7
76–100	47.4	36.6	47.9	41.7
2nd Grade				
0–50	30.0	37.2	18.2	—
51–75	42.1	37.2	27.3	—
76–100	36.8	25.6	54.6	—
p-value comparing 1st- and 2nd-grade distributions	0.97	0.62	0.62	0.09
Average Percentage of Features Implemented by Teacher with Expected Frequency				
1st grade	70.4	62.5	72.6	70.4
2nd grade	67.4	59.8	77.8	—
p-value comparing 1st and 2nd grade	0.59	0.66	0.11	0.06
Sample Size				
1st grade	**38**	**41**	**48**	**60**
2nd grade	**38**	**43**	**44**	**59**
2nd-Grade Teachers in Group 1 Schools				
	Investigations	Math Expressions	Saxon	SFAW
Number of Features in Adherence Measure	15	14	12	8
Percentage of Features Implemented by Teacher with Expected Frequency				
0–50	25.0	50.0	31.3	47.1
51–75	62.5	50.0	25.0	17.7
76–100	12.5	0.0	43.8	35.3
Average Percentage of Features Implemented by Teacher with Expected Frequency	60.4	50.7	72.9	58.8
Sample Size	**16**	**20**	**16**	**17**
2nd-Grade Teachers in Group 2 Schools				
	Investigations	Math Expressions	Saxon	enVision
Number of Features in Adherence Measure	15	14	12	15
Percentage of Features Implemented by Teacher with Expected Frequency				
0–50	18.2	26.1	10.7	33.3
51–75	27.3	26.1	28.6	33.3
76–100	54.6	47.8	60.7	33.3
Average Percentage Of Features Implemented By Teacher With Expected Frequency	72.4	67.7	80.7	61.6
Sample Size	**22**	**23**	**28**	**42**

Source: Author calculations using spring teacher survey data.

Note: The statistical tests were conducted using two-level HLMs as described in the text.

For both Saxon and SFAW, we examine the effects of switching curricula after 1st grade. During the first year of study participation, some of the schools that were previously using Saxon or SFAW implemented their assigned curriculum in both the 1st and 2nd grade; the rest of the schools implemented their assigned curriculum in only the 1st grade during the first year of study participation. We focus on the effects of switching curricula between 1st and 2nd grade and not on the effects of switching between kindergarten and 1st grade, because kindergarten math curricula are often less structured or defined. For example, we compare 2nd-grade achievement of students who stayed with SFAW with that of students who switched to another study curriculum after 1st grade.

The Saxon analysis is based on 12 schools, and the SFAW analysis on 25 schools. Tables A.29 and A.30 present the number of schools, classrooms, and students included in the analyses. A separate three-level hierarchical linear model (HLM) was used to estimate the relative achievement effects of staying with Saxon and SFAW versus switching to another study curriculum. The HLM included the student, teacher, and school characteristics described in Section D, though the curriculum indicators in the school-level equation were replaced with variables that indicated whether the school was randomly assigned to switch to one of the other study curricula. Parameter estimates on the indicators equal the difference in average adjusted achievement between students whose school switched curriculum and those whose school stayed with its pre-study curriculum. The differences are presented in standard deviations (or effect size units) below, which were calculated by dividing each pair-wise comparison by the pooled standard deviation of the spring score for the two groups being compared; Hedges' g formula (with small-sample bias correction) was used to calculate the effect sizes. Only unadjusted *p*-values are reported, and values that are less than or equal to 0.05 are considered statistically significant, although the one result that is statistically significant remains significant even when a Bonferroni correction is used to adjust for the multiple comparisons that were made.

Results from the Saxon analyses indicate that average 2nd-grade achievement was similar among students who stayed with Saxon and those who switched to another study curriculum after 1st grade. Specifically, the difference in achievement between staying with Saxon versus switching to Investigations equals 0.09 standard deviations (*p*-value = 0.57), switching to Math Expression equals 0.00 standard deviations (*p*-value = 0.98), and switching to SFAW equals -0.01 standard deviations (*p*-value = 0.97)

Results from the SFAW analyses indicate that average 2nd-grade achievement was similar among students who stayed with SFAW and those who switched to Investigations or Math Expressions after 1st grade, whereas switching to Saxon resulted in higher achievement. Specifically, the difference in achievement between staying with SFAW versus switching to Investigations equals -0.07 standard deviations (*p*-value = 0.55), switching to Math Expression equals -0.06 standard deviations (*p*-value = 0.60), and switching to Saxon equals -0.35 standard deviations (*p*-value = 0.01)

Table A.29. Sample Sizes for Analyses that Compare Staying with Saxon Versus Switching to Another Study Curriculum

			Switched to		
	Total	Stayed with Saxon	Investigations	Math Expressions	SFAW
Staying with Saxon Through 2nd Grade Versus Switching to Another Curriculum After 1st Grade					
Schools	12	3	3	3	3
Classrooms	69	11	17	25	16
Students	701	136	165	240	160

Table A.30. Sample Sizes for Analyses that Compare Staying with SFAW Versus Switching to Another Study Curriculum

			Switched to		
	Total	Stayed with SFAW	Investigations	Math Expressions	Saxon
Staying with SFAW Through 2nd Grade Versus Switching to Another Curriculum After 1st Grade					
Schools	25	6	6	6	7
Classrooms	89	31	16	19	23
Students	934	306	190	216	222

G. Supplemental Analyses: Curriculum Effects during a Second Year of Implementation in the 1st Grade

Curriculum implementation was repeated in the 1st grade during the second year of the study in the 58 schools examined in the previous analysis, and 2 others (for a total of 60 schools). In each of these 60 schools, the study administered fall and spring tests to first graders each year, allowing us to examine whether curriculum effects in a particular grade change as schools gain experience with a curriculum in that grade. The analysis is based on 2,475 first graders across 199 classrooms who participated in the second year of the study.

Table A.31 shows the number of schools that participated in the first year and implemented their assigned curriculum in the 1st grade, along with the number that participated a second year and repeated implementation in the 1st grade. Table A.32 shows the number of schools, classrooms, and students in the analysis sample (after all attrition was accounted for).

Some teacher turnover occurred between the first and second years. All replacement teachers agreed to participate in the study and were included in the analysis. About 80 percent of the teachers who participated during the first year also participated during the second (not shown).

Table A.31. Schools that Implemented their Assigned Curriculum in 1st Grade for Two Years

		Schools by Curriculum			
	All Schools	Investigations	Math Expressions	Saxon	SFAW/ enVision
Schools in First Year (Number)	66	17	16	15	18
Schools in Second Year (Number)	60	14	14	14	18
Attrition rate (Percentage)	9.1	17.6	12.5	6.7	0.0

Table A.32. Number of Schools, Classrooms, and Students in the Analysis

		Samples by Curriculum			
	All	Investigations	Math Expressions	Saxon	SFAW/ enVision
Analysis Sample, Second Year					
Schools	60	14	14	14	18
Classrooms	199	45	42	53	59
Students	2,475	512	502	666	795

Data collection response rates. The fraction of students that were tested in both the fall and spring exceeded 80 percent for each curriculum group. Parent refusals accounted for approximately one-third of student nonresponse, and another 50 percent was due to students moving to a nonstudy school.

For the teacher assessment and fall teacher survey, response rates exceeded 90 percent for each curriculum group. For the spring survey, response rates for Saxon and SFAW exceeded 90 percent and equaled 89 and 88 percent for Investigations and Math Expressions, respectively.

Baseline equivalence. We examined the comparability of the curriculum groups along baseline school, teacher, and student characteristics for the new 1st-grade cohort and found the following:

- None of the school characteristics differ significantly across the curriculum groups.
- Nearly all measures of teacher demographics, education, experience, and scores on the pre-curriculum training teacher assessment do not differ significantly across the curriculum groups, with one exception: The percentage of teachers with a degree in education is significantly different across the curriculum groups, ranging from 96 to 100 percent.[19]

[19] Given the number of teacher characteristics examined (20 characteristics), our 5 percent threshold for statistical significance means that one characteristic could differ significantly across the curriculum groups by chance.

- None of the student characteristics is significantly different across the curriculum groups.

The approach for calculating curriculum effects adjusts for school, teacher, and student characteristics.

Curriculum effects. Like the prior analyses, we estimated three specifications of the HLM to calculate the curriculum effects: (1) a model that includes only the curriculum indicators and the block indicators used when conducting random assignment; (2) a model that adds the student's fall score to the first model; and (3) a model that adds as many of the other student-, teacher-, and school-level controls as possible to the second model.

Table A.33 summarizes the relative effects of the curricula for the two cohorts of first graders who participated in the study. The results are based on the third HLM. The pattern of results for the curriculum indicators is similar across the second and third models, both of which contain students' fall scores. The statistical significance of the curriculum differentials was calculated with and without adjusting for the six unique curriculum-pair comparisons that were made, as described earlier.

Results based on unadjusted statistical tests indicate that after a second year of curriculum implementation (when teachers and schools have experience), students taught using Math Expressions and Saxon scored an average of 0.11 and 0.13 standard deviations higher than students taught using Investigations, respectively; none of the other curriculum differentials is statistically significant. These differences in test scores are the equivalent of moving a student from the 50th to the 54th to 55th percentile. Based on the adjusted statistical tests, none of the curriculum-pair differentials is statistically significant.

Sensitivity analyses. We explored whether the results are sensitive to (a) the few schools that dropped out of the study and, therefore, had to be excluded from the analyses; and (b) the students who moved between study schools that used a different study curriculum. The pattern of results is robust to this sensitivity analysis and not affected by crossovers.

Table A.33. Difference Between Pairs of Curricula in Average HLM-Adjusted Spring Student Math Achievement (in Effect Sizes), Two Cohorts of 1st-Grade Students (*p*-Values In Parentheses)

	Effect of					
	Investigations Relative to			Math Expressions Relative to		Saxon Relative to
	Math Expressions	Saxon	SFAW/ enVision	Saxon	SFAW/ enVision	SFAW/ enVision
First Year of 1st-Grade Implementation						
Effect Size	-0.20*+	-0.23*+	-0.05	-0.01	0.15*+	0.18*+
Unadjusted *p*-Value	(0.00)	(0.00)	(0.33)	(0.81)	(0.01)	(0.00)
Adjusted *p*-Value	(0.01)	(0.00)	(0.76)	(1.00)	(0.04)	(0.01)
Second Year of 1st-Grade Implementation						
Effect Size	-0.11*	-0.13*	-0.06	-0.01	0.07	-0.08
Unadjusted *p*-Value	(0.03)	(0.02)	(0.28)	(0.90)	(0.21)	(0.14)
Adjusted *p*-Value	(0.13)	(0.09)	(0.70)	(1.00)	(0.59)	(0.43)

Source: Authors' calculations based on data from the spring 1st- and 2nd-grade ECLS-K math test administered by the study, school records, the fall teacher survey, and school-level data from the 2005–2006 CCD.

Note: Effect sizes were calculated by dividing each pair-wise curriculum comparison by the pooled standard deviation of the spring scale score of the two curricula being compared; Hedges' *g* formula (with the correction for small-sample bias) was used to calculate effect sizes. The unadjusted *p*-values do not account for the six pair-wise curriculum comparisons presented in the figure, whereas the adjusted *p*-values, which were calculated using the Tukey-Kramer method, account for the comparisons.

*Indicates that the effect size is statistically significant at the 5 percent level, according to the unadjusted *p*-value.
+Indicates that the effect size is statistically significant at the 5 percent level, according to the adjusted *p*-value.

For more information on the full study, please visit:

http://ies.ed.gov/ncee/projects/evaluation/math_curricula.asp

To read the evaluation brief, please visit:

http://ies.ed.gov/ncee/pubs/20134019/pdf/20134019.pdf

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

This brief was prepared and based on analysis for NCEE by Roberto Agodini, Barbara Harris, Neil Seftor, Janine Remillard, and Melissa Thomas under contract with Mathematica Policy Research (contract number ED-04-CO-0112/0003). All authors are affiliated with Mathematica Policy Research, except Janine Remillard who is affiliated with the University of Pennsylvania. The NCEE project officer was Audrey Pendleton.