

Achievement Effects of Four Early Elementary School Math Curricula

Findings from First Graders in 39 Schools

Executive Summary

Achievement Effects of Four Early Elementary School Math Curricula

Findings from First Graders in 39 Schools

Executive Summary

February 2009

Roberto Agodini
Barbara Harris
Sally Atkins-Burnett
Sheila Heaviside
Timothy Novak
Mathematica Policy Research, Inc.

Robert Murphy
SRI International

Audrey Pendleton
Project Officer
Institute of Education Sciences

U.S. Department of Education

Arne Duncan
Secretary

Institute of Education Sciences

Sue Betka
Acting Director

National Center for Education Evaluation and Regional Assistance

Phoebe Cottingham
Commissioner

February 2009

This report was prepared for the Institute of Education Sciences under Contract No. ED-04-CO-0112/0003. The project officer was Audrey Pendleton in the National Center for Education Evaluation and Regional Assistance.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the reports.

This publication is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the citation should be: Roberto Agodini, Barbara Harris, Sally Atkins-Burnett, Sheila Heaviside, Timothy Novak, and Robert Murphy (2009). *Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools—Executive Summary* (NCEE 2009-4053). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

To order copies of this report,

- Write to ED Pubs, Education Publications Center, U.S. Department of Education, P.O. Box 1398, Jessup, MD 20794-1398.
- Call in your request toll free to 1-877-4ED-Pubs. If 877 service is not yet available in your area, call 800-872-5327 (800-USA-LEARN). Those who use a telecommunications device for the deaf (TDD) or a teletypewriter (TTY) should call 800-437-0833.
- Fax your request to 301-470-1244.
- Order online at www.edpubs.org.

This report also is available on the IES website at <http://ies.ed.gov/ncee>.

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

ACHIEVEMENT EFFECTS OF FOUR EARLY ELEMENTARY SCHOOL MATH CURRICULA: FINDINGS FROM FIRST GRADERS IN 39 SCHOOLS

EXECUTIVE SUMMARY

Many U.S. children start school with weak math skills and there are differences between students from different socioeconomic backgrounds—those from poor families lag behind those from affluent ones (Rathburn and West 2004). These differences also grow over time, resulting in substantial differences in math achievement by the time students reach the fourth grade (Lee, Gregg, and Dion 2007).

The federal Title I program provides financial assistance to schools with a high number or percentage of poor children to help all students meet state academic standards. Under the No Child Left Behind Act (NCLB), Title I schools must make adequate yearly progress (AYP) in bringing their students to state-specific targets for proficiency in math and reading. The goal of this provision is to ensure that all students are proficient in math and reading by 2014.

The purpose of this large-scale, national study is to determine whether some early elementary school math curricula are more effective than others at improving student math achievement, thereby providing educators with information that may be useful for making AYP. A small number of curricula dominate elementary math instruction (seven math curricula make up 91 percent of the curricula used by K-2 educators), and the curricula are based on different theories for developing student math skills (Education Market Research 2008). NCLB emphasizes the importance of adopting scientifically-based educational practices; however, there is little rigorous research evidence to support one theory or curriculum over another. This study will help to fill that knowledge gap. The study is sponsored by the Institute of Education Sciences (IES) in the U.S. Department of Education and is being conducted by Mathematica Policy Research, Inc. (MPR) and its subcontractor SRI International (SRI).

BASIS FOR THE CURRENT FINDINGS

This report presents results from the first cohort of 39 schools participating in the evaluation, with the goal of answering the following research question: What are the relative effects of different early elementary math curricula on student math achievement in disadvantaged schools? The report also examines whether curriculum effects differ for student subgroups in different instructional settings.

Curricula Included in the Study. A competitive process was used to select four curricula for the evaluation that represent many of the diverse approaches used to teach elementary school math in the United States:

- Investigations in Number, Data, and Space (Investigations) published by Pearson Scott Foresman (Russell, Economopoulos, Mokros, Kliman, Wright, Clements, Goodrow, Murray, and Sarama 2006)

- Math Expressions published by Houghton Mifflin Company (Fuson 2006a)
- Saxon Math (Saxon) published by Harcourt Achieve (Larson 2004)
- Scott Foresman-Addison Wesley Mathematics (SFAW) published by Pearson Scott Foresman (Charles, Crown, Fennel, Caldwell, Cavanagh, Chancellor, Ramirez, Ramos, Sammons, Schielack, Tate, Thompson, and Van de Walle 2005)

The process for selecting the curricula began with the study team inviting developers and publishers of early elementary school math curricula to submit a proposal to include their curricula in the evaluation. A panel of outside experts in math and math instruction then reviewed the submissions and recommended to IES curricula suitable for the study. The goal of the review process was to identify widely used curricula that draw on different instructional approaches and that hold promise for improving student math achievement.

Study Design. An experimental design was used to evaluate the relative effects of the study’s four curricula. The design randomly assigned schools in each participating district to the four curricula, thereby setting up an experiment in each district. The relative effects of the curricula were calculated by comparing math achievement of students in the four curriculum groups.

The study does not include a control group of schools (or a “business as usual” group) that continue to use whatever math curriculum they were using before joining the study. The study team decided not to include such a control group because it would contain a variety of curricula used by the participating districts, thereby making it difficult to compare effects of the study’s curricula to effects for this group.

Participating Districts and Schools. The study compares the effects of the selected curricula on math achievement of students in disadvantaged schools. The study team identified and recruited districts that (1) have Title I schools, (2) are geographically dispersed, and (3) contain at least four elementary schools interested in study participation, so all four of the study’s curricula could be implemented in each district.

Participating sites are not a representative sample of districts and schools, because interested sites are likely to be unique in ways that make it difficult to select a representative sample. Interested districts were willing to use all four of the study’s curricula, allowed the curricula to be randomly assigned to their participating schools, and were willing to have the study team test students and collect other data required by the evaluation (as described below). It would have been extremely costly to recruit a representative sample of districts and schools that met these criteria.

The 39 schools examined in this report are contained in four districts that are geographically dispersed in four states and in three regions of the country. The districts also fall in areas with different levels of urbanicity—two districts are in urban areas, one is in a suburban area, and the other is in a rural area.

In this first cohort, curriculum implementation occurred in the first grade during the 2006-2007 school year. Data were collected from the 131 first-grade teachers in the study schools, and from 1,309 students—a random sample of about 10 students in each classroom was sufficient to support the analyses. Each of the four curricula was assigned about 10 schools with 33 classrooms and 325 students. The table below presents the exact number of schools, classrooms, and students included in the analysis, in total and by curriculum group.

NUMBER OF COHORT-ONE SCHOOLS, CLASSROOMS, AND STUDENTS,
IN TOTAL AND BY CURRICULUM

	All	Curriculum			
		Investigations	Math Expressions	Saxon	SFAW
Schools	39	10	9	9	11
Classrooms	131	33	31	31	36
Average # of classrooms/school	3.4	3.3	3.4	3.4	3.3
Students Both Fall and Spring Tested	1,309	332	314	304	359
Average # of students/classroom	10	10	10	10	10

An inspection of baseline school, teacher, and student characteristics shows that random assignment achieved its objective of creating four groups with similar characteristics before curriculum implementation began. The baseline characteristics include 7 school characteristics (see Table III.1 in the body of the report) 21 teacher characteristics (see Table II.1 in the body of the report), and 7 student characteristics (see Table III.2 in the body of the report), including student fall math achievement. Statistical tests indicate that none of the school and student characteristics are significantly different at the 5 percent level of confidence across the curriculum groups.¹ One of the 21 teacher characteristics (race) is significantly different across the curriculum groups;² however, as described in Chapter III, the approach for calculating curriculum effects adjusted for teacher race.

Statistical Power. The effect size that can be detected with the first cohort is as small as 0.22, where effect size is defined as a fraction of the standard deviation of the test score. Specifically, the minimum detectable effect (MDE) equals the difference in average student math scores of any two curriculum groups, divided by the pooled standard deviation of the score for the two curricula being compared.³

¹ The 5 percent level of confidence means there is no more than a 5 percent chance that the finding (that none of the school and student characteristics are different across the curriculum groups) could have occurred by chance.

² At least 93 percent of Investigations, Math Expressions, and Saxon teachers classified themselves as white, whereas 78 percent of SFAW teachers did so.

³ The MDE calculation accounts for the extent to which students in the first cohort are clustered in classrooms and schools according to their baseline achievement, after adjusting for other baseline student, teacher, and school

The MDE of 0.22 means that, when comparing student achievement of any two curriculum groups, it must differ by at least 15 percent of the gain made by the average first grader from a low income family to be detectable in this report. Chapter I provides more details about the computation of the MDE and what it represents.

Outcome Measure and Other Data Collection. To measure the achievement effects of the curricula, the study team tested students at the beginning and end of the school year using the math assessment developed for the Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K) (West, Denton, and Germino-Hausken 2000). The ECLS-K assessment is a nationally normed test that meets the study's requirements of: assessing knowledge and skills mathematicians and math educators feel are important for early elementary school students to develop; having accepted standards of validity and reliability; being administered to students individually; being able to measure achievement gains over the study's grade range (which ultimately will include the first, second, and third grades); and being able to accurately capture achievement of students from a wide range of backgrounds and ability levels.

Another important feature of the ECLS-K assessment is that it is an adaptive test, which is an approach used to measure achievement that is tailored to a student's achievement level. In particular, the test begins by administering to each student a short, first-stage routing test used to broadly measure each examinee's achievement level. Depending on the score on the routing test, the student is then administered one of three longer, second-stage tests: (1) an easy test, (2) a middle-difficulty test, or (3) a difficult test. Some of the items on the second-stage tests overlap, and this overlap is used to place the scores on the different tests on the same scale. Item response theory (IRT) techniques (Lord 1980) were used to develop the scale score, which, according to the test developers, are the appropriate scores to analyze for our purposes (Rock and Pollack 2002).⁴ Adaptive tests are useful for measuring achievement because they limit the amount of time children are away from their classrooms and reduce the risk of ceiling or floor effects in the test score distribution—something that can have adverse effects on measuring achievement gains.

The assessment includes questions in the five math 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 items in each of the second-stage tests administered to the study's first graders can primarily be classified as

(continued)

characteristics. The calculation also uses the Tukey-Kramer method (Tukey 1952, 1953; Kramer 1956) to account for the six unique pair-wise comparisons that can be made with the study's four curricula: (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.

⁴ Student answers on the assessment were sent to the Educational Testing Service (ETS) for scoring—ETS was a developer of the ECLS-K Mathematics Assessment. A three-parameter IRT model was used to place scores from the different tests students took on the same scale. Reliabilities for the study's sample (0.93 for the fall score and 0.94 for the spring score) were consistent with the national ECLS-K sample (Rock and Pollack 2002, pp. 5-7 through 5-9)—reliabilities are based on the internal consistency (alpha) coefficients. Also, there were no floor or ceiling effects observed in either the fall or spring scores.

Number Sense, Properties, and Operations, with the remainder from the other areas. The easy test contained only a few items from each of the remaining areas, whereas the middle-difficulty and difficult tests contained more such items. On the middle-difficulty test, the remaining items were mainly about Patterns, Algebra, and Functions, whereas those on the difficult test were mainly about Data Analysis, Statistics, and Probability.

To help interpret the measured effects of the curricula, teachers were surveyed about curriculum implementation. The survey data are useful for assessing teacher participation in curriculum training, usage of the assigned curriculum, and any supplementation with other materials. Teachers also reported their usage of the essential and secondary features of their assigned curriculum, which was useful for assessing adherence to each curriculum. Demographic information about teachers also was collected through the surveys, and student demographics were obtained from school records.

MAIN FINDINGS

The study's main findings include information about curriculum implementation and the relative effects of the curricula on student math achievement. Statistical tests were used to assess the significance of all the results. Hierarchical linear modeling (HLM) techniques—which account for the extent to which students are clustered in classrooms and schools according to achievement—were used to conduct the statistical tests. When comparing results for pairs of curricula, the Tukey-Kramer method (Tukey 1952, 1953; Kramer 1956) was used to adjust the statistical tests for the six unique pair-wise comparisons that can be made with four curricula, as described above. Only results that are statistically significant at the 5 percent level of confidence are discussed.⁵

Before presenting the main findings, it is worth mentioning the information that is and is not provided by the study. The relative effects of the curricula presented below reflect differences between the curricula, including differences in teacher training, instructional strategies, content coverage, and curriculum materials. Of course, the relative effects ultimately depend on how teachers implemented the curricula, and implementation reflects what publishers and teachers achieved, not some level of implementation specified by the study. Information about curriculum implementation presented in this report is based only on teacher reports—the study team is observing classrooms and plans to present that information in a future report.⁶ Also, the relative effects of the curricula are based only on the ECLS-K math assessment administered by the study team—in the third grade and perhaps even the second grade, districts administer their own math assessments to students and the study team is investigating the possibility of obtaining those scores for our future analyses of second and third graders. Lastly, because the participating

⁵ As mentioned above, the 5 percent level of confidence means there is no more than a 5 percent chance that any finding discussed could have occurred by chance.

⁶ Each classroom in the current sample was observed once during the 2006-2007 school year. Those observations are not presented in this report because the reliability of those data cannot be assessed until observations have been completed in all the study schools.

sites are not a representative sample of districts and schools, the design does not support making statements about effects for districts and schools outside of the study.

Curriculum Implementation. The main findings from the implementation analysis are:

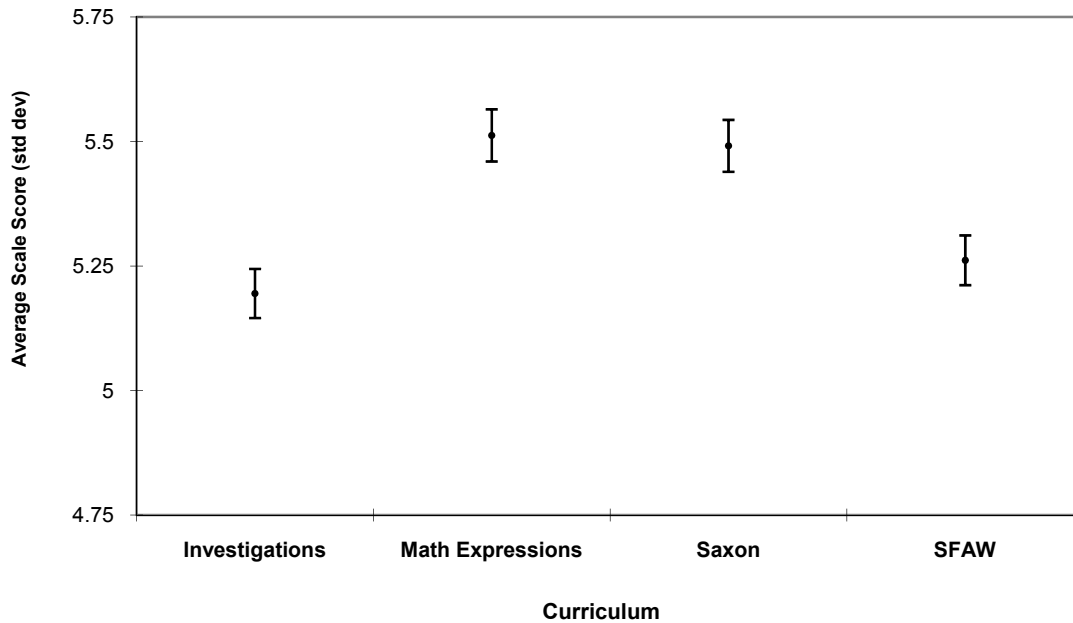
- All teachers received initial training from the publishers and 96 percent received follow-up training. Taken together, training varied by curriculum, ranging from 1.4 to 3.9 days.
- Nearly all teachers (99 percent in the fall, 98 percent in the spring) reported using their assigned curriculum as their core math curriculum according to the fall and spring surveys, and about a third (34 percent in fall and 36 percent in spring) reported supplementing their curriculum with other materials.
- Eighty-eight percent of teachers reported completing at least 80 percent of their assigned curriculum.⁷
- On average, Saxon teachers reported spending one more hour on math instruction per week than did teachers of the other curricula.

Achievement Effects. The figure below illustrates the relative effects of the study's curricula on student math achievement. The figure includes a symbol for each of the four curricula, where the dot in the middle of each symbol indicates the average spring math score of students in the respective curriculum groups. The average scores are adjusted for baseline measures of several student, teacher, and school characteristics related to student spring achievement (such as student fall math scores) to improve the precision of the results. The bars that extend from each dot represent the 95 percent confidence interval around each average score. HLM techniques were used to calculate the average scores and confidence intervals.

Curricula with non-overlapping confidence intervals have average scores that are significantly different at the 5 percent level of confidence. The results are presented in standard deviations, which means that subtracting the average values (the dots) for any two curricula indicates the effect size of using the first curriculum instead of the second. The effect sizes discussed below were calculated by dividing each pair-wise curriculum comparison by the pooled standard deviation of the spring score for the two curricula being compared, and Hedges' *g* formula (with the correction for small-sample bias) was used to calculate the pooled standard

⁷ Adherence to the essential features of each curriculum also was examined and is presented in Chapter II. Several analytical approaches can be used to examine adherence, but only one approach could be supported by the relatively small teacher sample sizes that are currently available for each curriculum. We do not make any general statements about adherence in the executive summary because it would be useful to examine whether the results are sensitive to the other analytical approaches, and instead encourage readers interested in the adherence analysis we were able to conduct at this point to see Chapter II. A future planned report (described at the end of the executive summary) will have larger teacher sample sizes that can support the other analyses.

**Average HLM-Adjusted Spring Math Score with Confidence Interval, by Curriculum
(in standard deviations)**



Note: The dots in each symbol represent the average HLM-adjusted spring math score (in standard deviations) for each curriculum, and the bars that extend from each dot represent the 95 percent confidence interval around each average. Curricula with non-overlapping confidence intervals have significantly different average scores at the 5 percent level of confidence.

deviations. Appendix D presents averages of the unadjusted math scores (see Table D.3). The relative effects of the curricula described below are similar when based on the simple averages, although the confidence intervals are wider than those based on the HLM-adjusted averages, as expected.

The figure shows that:

- **Student math achievement was significantly higher in schools assigned to Math Expressions and Saxon, than in schools assigned to Investigations and SFAW.** Average HLM-adjusted spring math achievement of Math Expressions and Saxon students was 0.30 standard deviations higher than Investigations students, and 0.24 standard deviations higher than SFAW students. For a student at the 50th percentile in math achievement, these effects mean that the student's percentile rank would be 9 to 12 points higher if the school used Math Expressions or Saxon, instead of Investigations or SFAW.

- **Math achievement in schools assigned to the two more effective curricula (Math Expressions and Saxon) was not significantly different, nor was math achievement in schools assigned to the two less effective curricula (Investigations and SFAW).** The Math Expressions-Saxon and Investigations-SFAW differentials equal 0.02 and -0.07 standard deviations, respectively, and neither is statistically significant.

We also examined whether the relative effects of the curricula differ along six characteristics that differentiate instructional settings: (1) participating districts, (2) school fall achievement, (3) school free/reduced-price meals eligibility, (4) teacher education, (5) teacher experience, and (6) teacher math content/pedagogical knowledge that was measured before curriculum training began using an assessment administered by the study team. These characteristics were used to create 15 subgroups—one for each of the four districts, three based on school fall achievement, and two subgroups for each of the other four characteristics.

Eight of the fifteen subgroup analyses found statistically significant differences in student math achievement between curricula. The significant curriculum differences ranged from 0.28 to 0.71 standard deviations, and all of the significant differences favored Math Expressions or Saxon over Investigations or SFAW. There were no subgroups for which Investigations or SFAW showed a statistically significant advantage.

NEXT STEPS FOR THE STUDY

Another 71 schools joined the study during the 2007-2008 school year (the year after the 39 schools examined in this report joined), and curriculum implementation occurred in both the first and second grades in all participating schools. A follow-up report is planned that will present results based on all 110 schools participating in the evaluation, and for both the first and second grades. The study also is supporting curriculum implementation and data collection during the 2008-2009 school year in a subset of schools, in which implementation will be expanded to the third grade. A third report is planned that will present those results.

REFERENCES

- Charles, Randall, Warren Crown, Francis (Skip) Fennell, Janet H. Caldwell, Mary Cavanagh, Dinah Chancellor, Alma B. Ramirez, Jeanne F. Ramos, Kay Sammons, Jane F. Schielack, William Tate, Mary Thompson, and John A. Van de Walle. *Scott Foresman-Addison Wesley Mathematics. Grade 1*. Glenview, IL: Pearson Scott Foresman, 2005.
- Education Market Research. *Mathematics Market, Grades K-12, 2008: Teaching Methods, Textbooks/Materials Used and Needed, and Market Size*. Rockaway Park, NY: EMR, 2008.
- Fuson, Karen C. *Math Expressions. Grade 1*. Boston, MA: Houghton Mifflin Company, 2006a.
- Kramer C.Y. "Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications." *Biometrika*, vol. 12, 1956, pp. 307-310.
- Larson, Nancy. *Saxon Math 1*. Austin, TX: Harcourt Achieve, 2004.
- Lee, J., W. Grigg, and G. Dion. *The Nation's Report Card: Mathematics 2007*. Publication No. NCES 2007-494. Washington, D.C: U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences, 2007.
- Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Publishers, 1980.
- Rathburn, A., and J. West. *From Kindergarten Through Third Grade: Children's Beginning School Experiences*. Publication no. NCES-2004-007. U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office, 2004.
- Rock, Donald A., and Judith M. Pollack. "Early Childhood Longitudinal Study—Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for Kindergarten Through First Grade." Publication No. NCES 2002-05. Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2002.
- Russell, Susan J., Karen Economopoulos, Jan Mokros, Marlene Kliman, Tracey Wright, Douglas H. Clements, Anne Goodrow, Megan Murray, and Julie Sarama. *Investigations in Number, Data, and Space. Grade 1*. Glenview, IL: Pearson Scott Foresman, 2006.
- Tukey, J.W. "The Problem of Multiple Comparisons." "Allowances for Various Types of Error Rates." Unpublished IMS address, Chicago, IL, 1952.
- Tukey, J.W. "The Problem of Multiple Comparisons." Unpublished manuscript, 1953.
- West, Jerry, Kristin Denton, and Elvira Germino-Hausken. "America's Kindergartners." Publication No. NCES 2000-070. Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2000.