What Works Clearinghouse™

INSTITUTE OF EDUCATION SCIENCES

Elementary School Mathematics Updated May 2013

Saxon Math

Report Contents

Program Description

Saxon Math, published by Houghton Mifflin Harcourt, is a core curriculum for students in grades K–5. A distinguishing feature of the curriculum is its use of an incremental approach for instruction and assessment. This approach limits the amount of new math content delivered to students each day and allows time for daily practice. New concepts are introduced gradually and integrated with previously introduced content so that concepts are developed, reviewed, and practiced over time rather than being taught during discrete periods of time, such as in chapters or units.

Instruction is built around math conversations that engage students in learning, as well as continuous practice with hands-on activities, manipulatives, and paper-pencil methods. The program includes frequent, cumulative assessments used to direct targeted remediation and support to struggling students. Starting in grade 3, the focus shifts from teacher-directed instruction to a more student-directed, independent learning approach, though math conversations continue to be used to introduce new concepts.

Research[2]

The What Works Clearinghouse (WWC) identified two studies of *Saxon Math* that both fall within the scope of the Elementary School Mathematics topic area and meet WWC evidence standards. One study meets standards without reservations, and the other study meets WWC evidence standards with reservations. Together, these studies included more than 8,060 students in grades 1–5 from 452 schools in 11 states.[3]

The WWC considers the extent of evidence for Saxon Math on the math performance of elementary school students to be medium to large for the mathematics achievement domain, the only outcome domain examined for studies reviewed under the Elementary School Mathematics topic area.

Effectiveness

Saxon Math was found to have potentially positive effects on mathematics achievement for elementary school students.

Table 1. Summary of findings[4]

Outcome domain	Rating of effectiveness	Improvement index *(percentile points)*: Average	Improvement index *(percentile points)*: Range	Number of studies	Number of students	Extent of evidence
Mathematics achievement	Potentially positive effects	+3	–2 to +7	2	> 8,060[5]	Medium to large

Program Information

Background

Saxon Math is distributed by Saxon Publishers, an imprint of Houghton Mifflin Harcourt Supplemental Publishers. Address: Specialized Curriculum Group, 9205 Southpark Center Loop, Orlando, FL 32819. Email: greatservice@hmhpub.com. Website: http://www.hmheducation.com/saxonmathk5/index.php. Telephone: (800) 289-4490. Fax: (800) 289-3994.

Program details

Saxon Math uses an incremental and integrated approach to instruction that includes three strategies: (a) fact-fluency practice that promotes recall when working with math operations and fractions, (b) mental math exercises intended to build number sense and problem-solving strategies, and (c) practice solving challenging, non-routine story problems in which problem solving strategies are emphasized.

The curriculum's main classroom activities draw on these strategies. The first classroom activity is a daily whole-group activity that provides an opportunity for students to review previously covered material, focusing on number sense, math life-skills, and problem solving. A second activity engages students in conversations that help them grasp new mathematical ideas introduced that day. Students then practice both the newly acquired skills and previously learned concepts during daily written practice sessions. Students complete a similar set of written practice problems at home with adult support. Beginning in first grade, the curriculum incorporates a third activity that allows students to practice basic math facts. This component aims to improve recall of facts and enable students to solve more complex problems.

Students complete written, cumulative assessments after every five lessons. The results of these assessments provide teachers with data for instructional decision making and provide feedback for students and parents.

Cost

For *Saxon's Primary Math* curricula (available for grades K–4), each set of teacher's materials costs between $225.75 and $299.95, and student kits cost between $761.65 and $884.55 for 24 students and $889.42 and $1,086.00 for 32 students. For the *Saxon Math Intermediate 3-5* curricula (available for grades 3–5), the teacher's manual costs $238.30 and the student edition costs $68.65 per student.[6] Other available materials include posters, manipulatives, and guides for adapting the *Saxon Math* curriculum for special education students.

Research Summary

The WWC identified 26 studies that investigated the effects of *Saxon Math* on the math performance of elementary school students.

The WWC reviewed 14 of those studies against group design evidence standards. One study (Agodini, Harris, Thomas, Murphy, & Gallagher, 2010) is a randomized controlled trial that meets WWC evidence standards without reservations, and one study (Resendez & Manley, 2005) is a quasi-experimental design that meets WWC evidence standards with reservations. Those two studies are summarized in this report. Twelve studies do not meet WWC evidence standards.

The remaining 12 studies do not meet WWC eligibility screens for review in this topic area. Citations for all 26 studies are in the References section, which begins on p. 5.

Table 2. Scope of reviewed research[7]

Grade	1, 2, 3, 4, 5
Delivery method	Whole class
Program type	Curriculum

Summary of study meeting WWC evidence standards without reservations

Agodini et al. (2010) presented results for 110 elementary schools that had been randomly assigned to one of four conditions: *Investigations in Number, Data, and Space*® (28 schools), *Math Expressions* (27 schools), *Saxon Math* (26 schools), and *Scott Foresman–Addison Wesley Elementary Mathematics* (29 schools). The analysis included 4,716 first-grade students and 3,344 second-grade students who were evenly divided among the four conditions. The study authors compared average spring math achievement of students in each condition after one school year of program implementation. Student outcomes were measured by the Early Childhood Longitudinal Study–Kindergarten (ECLS-K) math assessment.

Summary of study meeting WWC evidence standards with reservations

Resendez and Manley (2005) conducted a study of available school-level test results that included 170 intervention schools and 172 comparison schools in Georgia. Comparison schools were matched to intervention schools based on student demographics. The intervention schools used the *Saxon Math* program recommended for each grade level in grades 1–8 between 2000 and 2005. The comparison schools used a variety of other curricula. About three-fifths of comparison schools used traditional basal math curricula; one-third of the schools used a mix of basal, investigative, and other approaches; and 5% used an investigative approach to teaching math. This intervention report presents the study's findings for grades 1–5.

Effectiveness Summary

The WWC review of *Saxon Math* for the Elementary School Mathematics topic area includes student outcomes in one domain: mathematics achievement. The two studies of *Saxon Math* that meet WWC evidence standards reported findings in this domain. The findings below present the authors' estimates and WWC-calculated estimates of the size and statistical significance of the effects of *Saxon Math* on the mathematics achievement of elementary school students. For a more detailed description of the rating of effectiveness and extent of evidence criteria, see the WWC Rating Criteria on p. 19.

Summary of effectiveness for the mathematics achievement domain

Two studies reported findings in the mathematics achievement domain.

Agodini et al. (2010) reported, and the WWC confirmed, statistically significant positive effects of the *Saxon Math* program on the ECLS-K math assessment when compared to *Scott Foresman–Addison Wesley Elementary Mathematics* in grade 2. The study reports no significant effects of *Saxon Math* on the ECLS-K math assessment when compared to *Investigations in Number, Data, and Space®* and *Math Expressions*. The average effect size across the curricula and both grades (first and second) was not large enough to be considered substantively important according to WWC criteria (an effect size of at least 0.25). Based on the one statistically significant finding, the WWC characterizes this study as having statistically significant positive effects.

Resendez and Manley (2005) reported significant effects of the *Saxon Math* program on school-level math achievement in grades 2, 4, and 5, but reported no significant effects in grades 1 and 3. These findings control for schools' baseline math achievement levels. Due to the lack of student-level data, the student-level effect size and improvement index could not be calculated for this study. Based on WWC calculations, the average effect across grades 1–5 is not statistically significant. Therefore, this study is characterized as having indeterminate effects.

Thus, for the mathematics achievement domain, one study showed statistically significant positive effects and one study showed indeterminate effects. This results in a rating of potentially positive effects, with a medium to large extent of evidence.

Table 3. Rating of effectiveness and extent of evidence for the mathematics achievement domain

Rating of effectiveness	Criteria met
Potentially positive effects *Evidence of a positive effect with no overriding contrary evidence.*	In the two studies that reported findings, the estimated impact of the intervention on outcomes in the *mathematics achievement* domain was positive and statistically significant in one study and indeterminate in one study.
Extent of evidence	**Criteria met**
Medium to large	Two studies that included more than 8,060 students in 452 schools reported evidence of effectiveness in the *mathematics achievement* domain.

References

Study that meets WWC evidence standards without reservations

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders* (NCEE 2011-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20114001/pdf/20114001.pdf

Additional sources:

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary school math curricula. *Journal of Research on Educational Effectiveness, 3*(3), 199–253.

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools* (NCEE 2009-4052). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., Murphy, R., & Institute of Education Sciences (ED), National Center for Education Evaluation and Regional Assistance. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools* (NCEE 2009-4053). Executive summary. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Study that meets WWC evidence standards with reservations

Resendez, M., & Manley, M. A. (2005). *The relationship between using Saxon Elementary and Middle School Math and student performance on Georgia statewide assessments*. Orlando, FL: Harcourt Achieve.

Studies that do not meet WWC evidence standards

Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: The case of elementary mathematics in Indiana. *Educational Evaluation and Policy Analysis, 34*(4), 391–412. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Calvery, R., Bell, D., & Wheeler, G. (1993, November). *A comparison of selected second and third graders' math achievement: Saxon vs. Holt.* Paper presented at the meeting of the Mid-South Educational Research Association, New Orleans, LA. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Cummins-Colburn, B. J. L. (2007). Differences between state-adopted textbooks and student outcomes on the Texas Assessment of Knowledge and Skills examination. *Dissertation Abstracts International, 68*(06A), 168–2299. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Educational Research Institute of America. (2009). *A longitudinal analysis of state mathematics scores for Indiana schools using Saxon Math No. 362.* Bloomington, IN: Author. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Good, K., Bickel, R., & Howley, C. (2006). *Saxon Elementary Math program effectiveness study.* Charlestown, WV: Edvantia. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Hansen, E., & Greene, K. (2000). *A recipe for math. What's cooking in the classroom: Saxon or Traditional?* The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Hook, W., Bishop, W., & Hook, J. (2007). A quality math curriculum in support of effective teaching for elementary schools. *Educational Studies in Mathematics, 65*(2), 125–148. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Nguyen, K., Elam, P., & Weeter, R. (1993). *The 1992–93 Saxon Mathematics program evaluation report*. Oklahoma City: Oklahoma City Public Schools. The study does not meet WWC evidence standards because the measures of effectiveness cannot be attributed solely to the intervention—the intervention was not implemented as designed.

Resendez, M., & Azin, M. (2007). *The relationship between using Saxon Elementary and Middle School Math and student performance on California statewide assessments*. Jackson, WY: PRES Associates. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Additional source:

Resendez, M., & Azin, M. (2007). *Saxon Math and California English Learners' math performance: Research brief.* Jackson, WY: PRES Associates.

Resendez, M., & Azin, M. (2008). *The relationship between using Saxon Math at the elementary and middle school levels and student performance on the North Carolina statewide assessment: Final report.* Jackson, WY: PRES Associates. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Resendez, M., Sridharan, S., & Azin, M. (2006). *The relationship between using Saxon Elementary School Math and student performance on Texas statewide assessments*. Jackson, WY: PRES Associates. The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Roan, C. (2012). *A comparison of elementary mathematics achievement in everyday math and Saxon Math schools in Illinois* (Doctoral dissertation). Available from ProQuest Dissertations and Theses. (UMI No. 3507509) The study does not meet WWC evidence standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Studies that are ineligible for review using the Elementary School Mathematics Evidence Review Protocol

Bell, G. (2011). *The effects of Saxon Math instruction on middle school students' mathematics achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. *(UMI No. 3488140)* The study is ineligible for review because it does not use a sample within the age or grade range specificed in the protocol.

Christofori, P. (2005). *The effect of direct instruction math curriculum on higher-order problem solving* (Unpublished doctoral dissertation). University of South Florida, Sarasota. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Doe, C. (2006). Marvelous math products. *MultiMedia & Internet@Schools, 13*(3), 30–33. The study is ineligible for review because it does not examine the effectiveness of an intervention.

Fahsl, A. J. (2001). An investigation of the effects of exposure to Saxon Math textbooks, socioeconomic status and gender on math achievement scores. *Dissertation Abstracts International, 62*(08), 2681A. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Educational Research Institute of America. (2009). *A longitudinal analysis of state mathematics scores for Florida schools using Saxon Math No. 365.* The study is ineligible for review because it does not use a comparison group design or a single-case design.

Educational Research Institute of America. (2009). *A longitudinal analysis of state mathematics scores for Oklahoma schools using Saxon Math No. 363.* Bloomington, IN: Author. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Harcourt Achieve, Inc. (2005). *Case study research summaries of Saxon Math.* Retrieved from http://saxonpublishers.hmhco.com/en/resources/result_c.htm?ca=Research%3a+Efficacy&SRC1=4 The study is ineligible for review because it does not use a comparison group design or a single-case design.

Klein, D. (2000). *High achievement in mathematics: Lessons from three Los Angeles elementary schools*. Washington, DC: Brookings Institution Press. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Plato, J. (1998). *An evaluation of Saxon Math at Blessed Sacrament School.* Retrieved from http://lrs.ed.uiuc.edu/students/plato1/Final.html The study is ineligible for review because it does not use a comparison group design or a single-case design.

Slavin, R. E., & Lake, C. (2007). Effective programs in elementary mathematics: A best-evidence synthesis. *The Best Evidence Encyclopedia, 1*(2). Retrieved from http://www.bestevidence.org/word/elem_math_Feb_9_2007.pdf The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Additional source:

Slavin, R. E., & Lake, C. (2009). *Effective programs for elementary mathematics: A best evidence synthesis. Educator's summary.* Retrieved from http://www.bestevidence.org/word/elem_math_Mar_11_2009_sum.pdf

Viadero, D. (2009). Study gives edge to 2 math programs. *Education Week, 28*(23), 1–13. The study is ineligible for review because it does not examine the effectiveness of an intervention.

Vinogradova, E., King, C., & Rhoades, T. (2008, April). *Success for all students: What works? Best practices in Maryland public schools*. Paper presented at the annual meeting of the American Sociological Association, Boston, MA. The study is ineligible for review because it does not examine the effectiveness of an intervention.

Appendix A.1: Research details for Agodini et al. (2010)

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders* (NCEE 2011-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20114001/pdf/20114001.pdf

Table A1. Summary of findings

Meets WWC evidence standards without reservations

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	110 schools/8,060 students	+3	Yes

Setting The study took place in elementary schools in 12 districts across 10 states, including Connecticut, Florida, Kentucky, Minnesota, Mississippi, Missouri, Nevada, New York, South Carolina, and Texas. Of the 12 districts, three were in urban areas, five were in suburban areas, and four were in rural areas.

Study sample Following district and school recruitment and collection of consent from all teachers in the participating grades, 111 participating schools were randomly assigned to one of four curricula: (a) *Investigations in Number, Data, and Space*®, (b) *Math Expressions*, (c) *Saxon Math*, and (d) *Scott Foresman–Addison Wesley Mathematics*. Blocked random assignment of the schools was conducted separately within each district. In each district, participating schools were grouped together into blocks of four to seven schools based on characteristics such as Title I eligibility, free or reduced-price lunch eligibility status, grade enrollment size, math proficiency, and proportion of White and Hispanic students. Two districts had an additional blocking variable (magnet school status in one district and year-round school schedule in another district). One district required that all schools that fed into the same middle school receive the same condition. Schools in each block were randomly assigned among the four curricula. On average, 11 students were randomly sampled from each participating classroom for assessment. One school with three teachers and 32 students assigned to *Math Expressions* withdrew from the study and did not permit follow-up data collection.

The analysis sample included a total of 110 schools, 461 first-grade classrooms, 4,716 first graders, 328 second-grade classrooms, and 3,344 second graders. In the first grade sample, on average, 27 schools, 116 classrooms, and 1,180 students were assigned to each condition. In the second grade sample, on average, 18 schools, 82 classrooms, and 835 students were assigned to each condition.

Seventy-six percent of the schools in the study were eligible for Title I funding. Approximately half of the students in the sample were eligible for free or reduced-price lunch. Among students in the sample, 39% were White, 32% were non-Hispanic Black, 26% were Hispanic, 2% were Asian, and 1% were American Indian or Alaskan Native.

Intervention group

Students used *Saxon Math* as their core math curriculum. Study authors reported that about six out of seven teachers self-reported completing at least 80% of the curriculum.

Comparison group

The study included three comparison groups: (a) *Investigations in Number, Data, and Space®*, (b) *Math Expressions*, and (c) *Scott Foresman–Addison Wesley Elementary Mathematics.* Each curriculum was implemented by comparison teachers for one school year.

Investigations in Number, Data, and Space® is published by Pearson Scott Foresman. It uses a student-centered approach that encourages reasoning and understanding and draws on constructivist learning theory. The lessons build on students' existing knowledge and focus on understanding math concepts rather than simply learning computational methods. The curriculum is organized in nine thematic units, each lasting 5–5.5 weeks. Study authors reported that about four out of five teachers self-reported completing at least 80% of the curriculum.

Math Expressions is published by Houghton Mifflin Harcourt and uses a blend of student-centered and teacher-directed instructional approaches. Students using the curriculum question and discuss mathematics and are explicitly taught problem solving strategies. There is an emphasis on using multiple specified objects, drawings, and language to represent concepts, and on learning through the use of real-world situations. Students are expected to explain and justify their solutions. Study authors reported that about nine out of 10 teachers self-reported completing at least 80% of the curriculum.

Scott Foresman–Addison Wesley Mathematics is published by Pearson Scott Foresman and is a 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. Study authors reported that about nine out of 10 teachers self-reported completing at least 80% of the curriculum.

Outcomes and measurement

Mathematics achievement was measured using the mathematics assessment developed for the ECLS-K class of 1998–99. The assessment is individually administered, nationally normed, and adaptive. The assessment meets accepted standards of validity and reliability. Scale scores from an item response theory (IRT) model were used in the analysis. The test was administered in the fall of the implementation year (within 4 weeks of the first day of classes) to assess students' baseline math achievement. The test was also administered in the spring—that is, from 1–6 weeks before the end of the school year of program implementation. For a more detailed description of the outcome measure, see Appendix B.

Support for implementation

Teachers in all four groups were provided training by the curriculum publisher. Teachers assigned to *Saxon Math* were provided 1 day of initial training in the summer before the school year began. One follow-up training session, tailored to meet each district's needs, was offered during the school year.

Teachers assigned to *Investigations in Number, Data, and Space®* (comparison group 1) were provided 1 day of initial training in the summer before the school year began. Follow-up sessions were typically 3–4 hours long and held after school.

Teachers assigned to *Math Expressions* (comparison group 2) were provided 2 days of initial training in the summer before the school year began. Two follow-up trainings were offered during the school year. Follow-up sessions typically consisted of classroom observations followed by short feedback sessions with teachers.

Teachers assigned to *Scott Foresman–Addison Wesley Elementary Mathematics* (comparison group 3) received 1 day of initial training in the summer before the school year began. Follow-up training was offered about every 4–6 weeks throughout the school year. Follow-up sessions were typically 3–4 hours long and held after school.

Appendix A.2: Research details for Resendez and Manley (2005)

Resendez, M., & Manley, M. A. (2005). ***The relationship between using Saxon Elementary and Middle School Math and student performance on Georgia statewide assessments.*** **Orlando, FL: Harcourt Achieve.**

Table A2. Summary of findings — **Meets WWC evidence standards with reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	342 schools	na	No

na = not applicable

Setting The schools included in the study were distributed across the state of Georgia and represented a mixture of rural, urban, and suburban communities.

Study sample Using information provided by the Georgia Department of Education, the study authors identified Georgia schools that used the *Saxon Math* curricula between 2000 and 2005, as well as schools that did not use *Saxon Math* but had similar student demographics to those who did. The study sample included students in grades 1–8 in 170 intervention schools and 172 comparison schools. This intervention report focuses only on findings for grades 1–5, because grades 6–8 are outside of the scope of this review.[8] Data for the intervention group came from 85 schools for first grade, 85 schools for second grade, 83 schools for third grade, 79 schools for fourth grade, and 79 schools for fifth grade. Data for the comparison group came from 144 schools for first grade, 144 schools for second grade, 135 schools for third grade, 131 schools for fourth grade, and 129 schools for fifth grade. The authors reported no significant differences in baseline math performance between the *Saxon* and non-*Saxon* schools.

Intervention group The *Saxon Math* curricula were used as a core curriculum in the intervention schools. These schools used the version of the *Saxon Math* program that was appropriate for each grade level. Participating schools had used the program for an average of three years.

Comparison group Comparison group schools were selected from among all Georgia schools that did not implement *Saxon Math* based on propensity score matching methods. Schools were matched based on the their percentages of students who were female, African American, White, Hispanic, Native American, limited English proficient, educationally disadvantaged, migrant, disabled, gifted, and having left school during the prior year. The comparison group schools used a mixture of non-*Saxon* curricula. Sixty-two percent of the schools in the comparison group used basal math curricula with chapter-based approaches to teaching math. Five percent of the schools used curricula with an investigative approach. The remaining 33% of the schools used curricula that were a mix of basal, investigative, and computer-based approaches. No additional information was provided by the authors about the specific components of the basal, investigative, or computer-based approaches.

Outcomes and measurement

Study authors measured outcomes using Georgia's Criterion-Referenced Competency Test (CRCT), which assesses competency in number sense and numeration, geometry and measurement, patterns and relations/algebra, statistics and probability, computation and estimation, and problem solving. The authors note that per state policy, only school-level data could be released. Fourth-grade students were tested in each school year from 1999–2000 to 2004–05. First-grade, second-grade, third-grade, and fifth-grade students were tested in the spring of school years 2001–02, 2003–04, and 2004–05. All posttest scores are from spring 2005. For a more detailed description of this outcome measure, see Appendix B.

Support for implementation

The intervention and comparison schools in the study were all using their curricula as part of business-as-usual operations and did not receive additional implementation support as a part of the study. Therefore, teachers received the training and implementation support normally provided with their school's curriculum. The study does not provide additional details on implementation support that schools may have received from curricula developers or other parties.

Appendix B: Outcome measures for the mathematics achievement domain

Mathematics achievement	
Early Childhood Longitudinal Study–Kindergarten (ECLS-K) Math Assessment	This assessment was developed for the ECLS-K class of 1998–99. The ECLS-K is a nationally normed adaptive test. The assessment measures understanding and skills in five content areas: (a) number sense, properties, and operations; (b) measurement; (c) geometry and spatial sense; (d) data analysis, statistics, and probability; and (e) patterns, algebra, and functions. On the first-grade test, approximately three-quarters of the items focused on number sense, properties, and operations, with the remaining items predominantly drawn from the areas of data analysis, statistics, and probability; and patterns, algebra, and functions. An ECLS-K math assessment for the second grade did not exist, so the study authors worked with the developer of the ECLS-K, Educational Testing Service, to select appropriate items from existing ECLS-K math assessments (including the K–1, third-, and fifth-grade instruments). Half of the items in the second-grade test were related to number sense, properties, and operations, with the other half covering measurement; geometry and spatial sense; and patterns, algebra, and functions (as cited in Agodini et al., 2010).
Georgia's Criterion-Referenced Competency Test (CRCT), Mathematics[9]	As cited in Resendez and Manley (2005), the CRCT is a criterion-referenced test linked to Georgia's Quality Core Curriculum Goals. According to the Georgia Department of Education, the CRCT is a multiple-choice test that is valid and reliable for Georgia's public school students.[10] The CRCT math scores range from 150 to 450, with scores below 300 not meeting standards and scores above 350 exceeding standards. The criteria for meeting the standards vary by objective and grade level. The test includes subscales that cover six objectives: (a) numbers and number sense; (b) geometry and measurement; (c) patterns, relationships, and algebra; (d) statistics and probability; (e) computation and estimation; and (f) problem solving. The cut points are set by the state and take into account the difficulty of each specific objective.

Appendix C: Findings included in the rating for the mathematics achievement domain

			Mean (standard deviation)		WWC calculations			
Outcome measure	Study sample	Sample size	Intervention group	Comparison group	Mean difference	Effect size	Improvement index	*p*-value
Agodini et al., 2010[a]								
ECLS-K	Grade 1 (vs. *Investigations in Number, Data, and Space*)	54 schools/ 2,235 students	45.05 (7.32)	44.51 (8.04)	0.54	0.07	+3	0.15
ECLS-K	Grade 1 (vs. *Math Expressions*)	52 schools/ 2,320 students	44.36 (7.32)	44.74 (8.52)	–0.38	–0.05	–2	0.31
ECLS-K	Grade 1 (vs. *Scott Foresman–Addison Wesley*)	55 schools/ 2,377 students	44.94 (7.32)	44.43 (8.15)	0.51	0.07	+3	0.16
ECLS-K	Grade 2 (vs. *Investigations in Number, Data, and Space*)	36 schools/ 1,711 students	71.25 (16.16)	69.85 (15.75)	1.40	0.09	+3	0.09
ECLS-K	Grade 2 (vs. *Math Expressions*)	35 schools/ 1,721 students	72.24 (16.16)	71.38 (16.70)	0.86	0.05	+2	0.28
ECLS-K	Grade 2 (vs. *Scott Foresman–Addison Wesley*)	36 schools/ 1,706 students	73.06 (16.16)	70.31 (15.74)	2.75	0.17	+7	0.00
Domain average for mathematics achievement (Agodini et al., 2010)						**0.07**	**+3**	**Statistically significant**
Resendez & Manley, 2005[b]								
CRCT	Grade 1	229 schools/ nr students	86.26 (6.60)	85.20 (6.80)	1.06	na	na	0.19
CRCT	Grade 2	229 schools/ nr students	88.31 (6.39)	86.86 (7.35)	1.45	na	na	0.00
CRCT	Grade 3	218 schools/ nr students	86.94 (6.50)	85.93 (7.15)	1.01	na	na	0.12
CRCT	Grade 4	210 schools/ nr students	73.92 (8.51)	71.39 (11.83)	2.53	na	na	0.00
CRCT	Grade 5	208 schools/ nr students	82.46 (6.94)	81.66 (8.93)	0.80	na	na	0.00
Domain average for mathematics achievement (Resendez & Manley, 2005)						**na**	**na**	**na**
Domain average for mathematics achievement across all studies						**0.07**	**+3**	**na**

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the average change expected for all students who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is given the intervention. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. The statistical significance of each study's domain average was determined by the WWC. nr = not reported by the authors. na = not applicable. ECLS-K = Early Childhood Longitudinal Study–Kindergarten. CRCT = Criterion-Referenced Competency Test.

[a] For Agodini et al. (2010), the unit of assignment is the school. The *p*-values presented here were reported in the original study. The intervention group mean is the unadjusted comparison mean plus the program coefficients from the hierarchical linear modeling (HLM) analysis. The comparison group mean is the unadjusted comparison group mean. A correction for multiple comparisons was needed but did not affect the statistical significance of the findings. This study is characterized as having a statistically significant positive effect because the effect for at least one measure within the domain is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons.

[b] For Resendez & Manley (2005), no corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study. The original study reported only means for *CRCT* subtests. The value reported here is the mean across those subtests as reported by the author to the WWC. The means presented here adjust for differences in the groups at pretest. For subtest results, see Appendix D. Standard deviations are measured at the school level and were provided by the author to the WWC. Because student-level standard deviations were not available for this study, the student-level effect sizes and improvement indices could not be computed and the magnitude of the effect size was not considered for rating purposes. For further details, please see the *WWC Procedures and Standards Handbook*, Appendix B.

Appendix D: Summary of supplemental findings for the mathematics achievement domain

			Mean (standard deviation)		WWC calculations			
Outcome measure	Study sample	Sample size (schools)	Intervention group	Comparison group	Mean difference	Effect size	Improvement index	*p*-value
Resendez & Manley, 2005[a]								
CRCT: Numbers and number sense	Grade 1	229	89.53 (6.31)	88.52 (7.00)	1.01	na	na	0.16
CRCT: Geometry and measurement	Grade 1	229	90.34 (5.83)	90.29 (5.70)	0.05	na	na	0.94
CRCT: Patterns, relations, and algebra	Grade 1	229	87.88 (6.99)	86.28 (6.61)	1.60	na	na	0.02
CRCT: Computation and estimation	Grade 1	229	78.93 (9.54)	77.43 (10.10)	1.50	na	na	0.14
CRCT: Problem solving	Grade 1	229	84.64 (7.30)	83.49 (8.39)	1.15	na	na	0.17
CRCT: Numbers and number sense	Grade 2	229	88.57 (6.80)	86.62 (8.38)	1.95	na	na	0.02
CRCT: Geometry and measurement	Grade 2	229	91.46 (6.18)	92.36 (5.41)	–0.90	na	na	0.13
CRCT: Patterns, relations, and algebra	Grade 2	229	87.05 (7.43)	83.58 (9.63)	3.47	na	na	0.00
CRCT: Computation and estimation	Grade 2	229	86.93 (7.13)	85.83 (7.82)	1.10	na	na	0.15
CRCT: Problem solving	Grade 2	229	87.54 (7.48)	85.93 (8.28)	1.61	na	na	0.04
CRCT: Numbers and number sense	Grade 3	218	89.74 (6.29)	88.24 (7.02)	1.50	na	na	0.03
CRCT: Geometry and measurement	Grade 3	218	93.60 (4.50)	92.24 (6.22)	1.36	na	na	0.03
CRCT: Patterns, relations, and algebra	Grade 3	218	86.26 (6.67)	85.90 (7.12)	0.36	na	na	0.59
CRCT: Statistics and probability	Grade 3	218	87.13 (7.21)	85.83 (7.98)	1.30	na	na	0.09
CRCT: Computation and estimation	Grade 3	218	86.81 (7.80)	85.71 (8.02)	1.10	na	na	0.19
CRCT: Problem solving	Grade 3	218	78.11 (10.12)	77.64 (10.69)	0.47	na	na	0.63
CRCT: Numbers and number sense	Grade 4	210	71.47 (10.32)	70.85 (14.39)	0.62	na	na	0.65
CRCT: Geometry and measurement	Grade 4	210	79.22 (8.93)	78.16 (11.13)	1.06	na	na	0.33
CRCT: Patterns, relations, and algebra	Grade 4	210	69.76 (8.77)	67.70 (11.07)	2.06	na	na	0.06

CRCT: Statistics and probability	Grade 4	210	82.15 (9.05)	80.17 (10.82)	1.98	na	na	0.07
CRCT: Computation and estimation	Grade 4	210	73.12 (10.30)	67.65 (14.75)	5.47	na	na	0.00
CRCT: Problem solving	Grade 4	210	67.81 (9.87)	63.83 (14.44)	3.98	na	na	0.00
CRCT: Numbers and number sense	Grade 5	208	79.74 (9.55)	77.31 (11.51)	2.43	na	na	0.03
CRCT: Geometry and measurement	Grade 5	208	80.77 (9.01)	81.54 (9.88)	–0.77	na	na	0.44
CRCT: Patterns, relations, and algebra	Grade 5	208	76.16 (9.37)	74.56 (12.11)	1.60	na	na	0.15
CRCT: Statistics and probability	Grade 5	208	79.82 (7.71)	81.52 (9.65)	–1.70	na	na	0.07
CRCT: Computation and estimation	Grade 5	208	88.74 (6.56)	86.62 (8.55)	2.12	na	na	0.01
CRCT: Problem solving	Grade 5	208	89.55 (6.85)	88.43 (7.34)	1.12	na	na	0.14

Table Notes: The supplemental findings presented in this table are additional findings from the studies in this report that do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the average change expected for all students who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is given the intervention. na = not applicable. CRCT = Criterion-Referenced Competency Test.

[a] For Resendez & Manley (2005), no corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study. The means presented here adjust for differences in the groups at pretest. Standard deviations are measured at the school level and were provided by the author to the WWC. Because student-level standard deviations were not available for this study, the student-level effect sizes and improvement indices could not be computed, and the magnitude of the effect size was not considered for rating purposes. For further details, see the *WWC Procedures and Standards Handbook*, Appendix B.

Endnotes

[1] The descriptive information for this program was obtained from a publicly available source: the program's website (http://www.hmheducation.com/saxonmathk5/index.php, downloaded June 2010). The WWC requests developers review the program description sections for accuracy from their perspective. The program description was provided to the developer in February 2012, and the WWC incorporated feedback from the developer. Following internal review, the program description was provided again to the developer in January 2013, and the WWC incorporated additional feedback from the developer. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review. The literature search reflects documents publicly available by December 2012.

[2] The previous report was released in September 2010. This report has been updated to include reviews of six studies released since that report. Of the additional studies, one was within the scope of the Elementary School Mathematics review protocol and meets WWC evidence standards. The remaining five studies do not meet either WWC eligibility screens or evidence standards. A complete list and disposition of all studies reviewed are provided in the references. The studies in this report were reviewed using the Evidence Standards from the WWC Procedures and Standards Handbook (version 2.1) along with those described in the the Elementary School Mathematics review protocol (version 2.0). When intervention reports are updated, all studies are re-reviewed under the current WWC standards. In this report, a study that met standards with reservations (Good, Bickel, & Howley, 2006) in the September 2010 report was re-reviewed, and it does not meet standards under version 2.1 of the WWC Evidence Standards; the intervention and comparison groups in that study were not shown to be equivalent at baseline. The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

[3] Absence of conflict of interest: One of the studies summarized in this intervention report, Agodini et al. (2010), was prepared by staff of one of the WWC contractors. Because the principal investigator for the WWC review of Elementary School Mathematics is also a staff member of that contractor and an author of this study, the study was rated by staff members from a different organization. The report was then reviewed by the principal investigator, a WWC Quality Assurance reviewer, and an external peer reviewer.

[4] For criteria used in the determination of the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 19. These improvement index numbers show the average and range of student-level improvement indices for all findings in Agodini et al. (2010); it was not possible to calculate improvement indices for Resendez and Manley (2005) because student-level data were not provided.

[5] One study, Resendez and Manley (2005), reported school sample size but did not report student sample size.

[6] Both the primary and intermediate math curricula are available for grades 3 and 4.

[7] Grade, delivery method, and program type refer to the studies that meet WWC evidence standards without or with reservations.

[8] Results from grades 6–8 are being reviewed as part of the WWC Middle School Math review.

[9] The original CRCT scores shown in the report are by objective. Upon request from the WWC, the authors calculated the mean overall score across all objectives, controlling for pretest, for each grade.

[10] Georgia Department of Education. (n.d.). *Criterion-referenced competency tests.* Retrieved from http://www.doe.k12.ga.us/Curriculum-Instruction-and-Assessment/Assessment/Pages/CRCT.aspx

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, May). *Elementary School Mathematics intervention report: Saxon Math*. Retrieved from http://whatworks.ed.gov

WWC Rating Criteria

Criteria used to determine the rating of a study

Study rating	Criteria
Meets WWC evidence standards without reservations	A study that provides strong evidence for an intervention's effectiveness, such as a well-implemented RCT.
Meets WWC evidence standards with reservations	A study that provides weaker evidence for an intervention's effectiveness, such as a QED or an RCT with high attrition that has established equivalence of the analytic samples.

Criteria used to determine the rating of effectiveness for an intervention

Rating of effectiveness	Criteria
Positive effects	Two or more studies show statistically significant positive effects, at least one of which met WWC evidence standards for a strong design, AND No studies show statistically significant or substantively important negative effects.
Potentially positive effects	At least one study shows a statistically significant or substantively important positive effect, AND No studies show a statistically significant or substantively important negative effect AND fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects.
Mixed effects	At least one study shows a statistically significant or substantively important positive effect AND at least one study shows a statistically significant or substantively important negative effect, but no more such studies than the number showing a statistically significant or substantively important positive effect, OR At least one study shows a statistically significant or substantively important effect AND more studies show an indeterminate effect than show a statistically significant or substantively important effect.
Potentially negative effects	One study shows a statistically significant or substantively important negative effect and no studies show a statistically significant or substantively important positive effect, OR Two or more studies show statistically significant or substantively important negative effects, at least one study shows a statistically significant or substantively important positive effect, and more studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects.
Negative effects	Two or more studies show statistically significant negative effects, at least one of which met WWC evidence standards for a strong design, AND No studies show statistically significant or substantively important positive effects.
No discernible effects	None of the studies shows a statistically significant or substantively important effect, either positive or negative.

Criteria used to determine the extent of evidence for an intervention

Extent of evidence	Criteria
Medium to large	The domain includes more than one study, AND The domain includes more than one school, AND The domain findings are based on a total sample size of at least 350 students, OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.
Small	The domain includes only one study, OR The domain includes only one school, OR The domain findings are based on a total sample size of fewer than 350 students, AND, assuming 25 students in a class, a total of fewer than 14 classrooms across studies.

Glossary of Terms

Attrition Attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups. The WWC considers the total attrition rate and the difference in attrition rates across groups within a study.

Clustering adjustment If intervention assignment is made at a cluster level and the analysis is conducted at the student level, the WWC will adjust the statistical significance to account for this mismatch, if necessary.

Confounding factor A confounding factor is a component of a study that is completely aligned with one of the study conditions, making it impossible to separate how much of the observed effect was due to the intervention and how much was due to the factor.

Design The design of a study is the method by which intervention and comparison groups were assigned.

Domain A domain is a group of closely related outcomes.

Effect size The effect size is a measure of the magnitude of an effect. The WWC uses a standardized measure to facilitate comparisons across studies and outcomes.

Eligibility A study is eligible for review and inclusion in this report if it falls within the scope of the review protocol and uses either an experimental or matched comparison group design.

Equivalence A demonstration that the analysis sample groups are similar on observed characteristics defined in the review area protocol.

Extent of evidence An indication of how much evidence supports the findings. The criteria for the extent of evidence levels are given in the WWC Rating Criteria on p. 19.

Improvement index Along a percentile distribution of students, the improvement index represents the gain or loss of the average student due to the intervention. As the average student starts at the 50th percentile, the measure ranges from –50 to +50.

Multiple comparison adjustment When a study includes multiple outcomes or comparison groups, the WWC will adjust the statistical significance to account for the multiple comparisons, if necessary.

Quasi-experimental design (QED) A quasi-experimental design (QED) is a research design in which subjects are assigned to intervention and comparison groups through a process that is not random.

Randomized controlled trial (RCT) A randomized controlled trial (RCT) is an experiment in which investigators randomly assign eligible participants into intervention and comparison groups.

Rating of effectiveness The WWC rates the effects of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 19.

Single-case design A research approach in which an outcome variable is measured repeatedly within and across different conditions that are defined by the presence or absence of an intervention.

Standard deviation The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample tend to be spread out over a large range of values.

Statistical significance Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < 0.05$).

Substantively important A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.

Please see the WWC Procedures and Standards Handbook (version 2.1) for additional details.