What Works Clearinghouse

ies INSTITUTE OF EDUCATION SCIENCES

Saxon Elementary School Math

Program Description[2]

Saxon Elementary School Math, published by Houghton Mifflin Harcourt, is a core curriculum for students in kindergarten through grade 5. A distinguishing feature of *Saxon Elementary School Math* is its use of a distributed approach, as opposed to a chapter-based approach, for instruction and assessment. The program is built on the premise that students learn best when instruction is incremental and explicit, previously learned concepts are continually reviewed, and assessment is frequent and cumulative. At each grade level, math concepts are introduced, reviewed, and practiced over time in order to move students from understanding to mastery to fluency. For grades K–3, the *Saxon Elementary School Math* curriculum emphasizes hands-on activities and teacher-directed math conversations that engage students in learning. The curriculum for grades 4–5 also uses math conversations to introduce new concepts, and shifts the focus to student-directed learning.

Research[3]

One study of *Saxon Elementary School Math* that falls within the scope of the Elementary School Math review protocol meets What Works Clearinghouse (WWC) evidence standards, and two studies meet WWC evidence standards with reservations. The three studies included students in grades K–5 from 325 schools in 19 states.[4]

Based on these three studies, the WWC considers the extent of evidence for *Saxon Elementary School Math* on elementary school students to be medium to large for mathematics achievement.

1. This report has been updated to include reviews of 13 studies that have been released since 2005. Of the additional studies, 6 were not within the scope of the protocol and 5 were within the scope of the protocol but did not meet evidence standards. A complete list and disposition of all studies reviewed are provided in the references.
2. The descriptive information for this program was obtained from a publicly available source: the program's website (http://saxonpublishers.hmhco.com/en/saxonpublishers.htm, downloaded June 2010). The WWC requests developers to review the program description sections for accuracy from their perspective. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review.
3. The studies in this report were reviewed using WWC Evidence Standards, Version 1.0 (see the WWC Standards), as described in protocol Version 1.0.
4. The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

Effectiveness

Saxon Elementary School Math was found to have mixed effects on mathematics achievement.

	Mathematics achievement
Rating of effectiveness	Mixed effects
Improvement index[5]	Average: +5 percentile points
	Range: –1 to +12 percentile points

Absence of conflict of interest

The *Math Curricula* study summarized in this intervention report was prepared by staff of Mathematica Policy Research. Because the principal investigator for the WWC review of elementary school math interventions is also a Mathematica staff member, the study was rated by staff members from the University of Wisconsin and the Optimal Solutions Group. The intervention report was reviewed by the principal investigator, a WWC Quality Assurance reviewer, and an external peer reviewer.

Additional program information

Developer and contact

Saxon Elementary School Math was developed and is distributed by Saxon Publishers, an imprint of Houghton Mifflin Harcourt Supplemental Publishers. Address: 181 Ballardvale Street, Wilmington, MA 01887. Email: greatservice@hmhpub.com. Web: http://saxonpublishers.hmhco.com/. Telephone: (800) 289-3994.

Scope of use

The first Saxon textbook, *Saxon Algebra,* was published in 1979 by John Saxon for junior college students. In 1980, a high school version, *Algebra 1,* was published. In 1981, the program was tested by 20 teachers with approximately 1,400 students. By 1993, the company had become Saxon Publishers and had developed programs for kindergarten through high school. Information is not available on the numbers or demographics of students, schools, or districts using this intervention.

Teaching

Daily lessons in grades 1–3 consist of three components: (1) the meeting, (2) the math lesson, and (3) written practice, which includes guided class practice and homework. A typical lesson begins with the meeting, during which students engage in various practical activities (for example, understanding calendars) and enter into math conversations and dialogue with their classmates and teacher to communicate their understanding of math concepts. Following the meeting, the teacher introduces new concepts during the math lesson. Hands-on activities are incorporated into the math lesson to encourage student involvement and further the learning of new concepts. The math lesson is followed by written practice, which includes teacher-facilitated guided class practice of new and previously learned concepts. Students complete the day's homework independently. Cumulative and written assessments occur every five lessons.

In kindergarten, the same three components are used but may be separated into different sessions, and assessments are conducted as individual interviews between the teacher and individual students. For grades 4 and 5, a daily lesson consists of four components: (1) the warm-up; (2) the math lesson, which introduces a new math concept; (3) practice on the new concept; and (4) mixed practice, including new and previously learned concepts. Students are introduced to concepts incrementally, given opportunities for continual review and practice, and assessed cumulatively and frequently. An assessment score of 80% or lower indicates a need for remediation, and provision for remediation is part of the program.

5. These numbers show the average and range of student-level improvement indices for findings across two of the three studies. It was not possible to calculate improvement indices for Resendez and Manley (2005) due to the lack of student-level data.

Additional program information *(continued)*

Cost

Saxon Elementary School Math for grades K–3 can be ordered as a 24-student or 32-student kit that includes all the teacher, lesson, classroom, and student materials. The student kits range from more than $600 to more than $800, depending on the size of the kit. Individual kit components, such as manipulatives, workbooks, student texts, teacher manuals, and materials in Spanish, can be purchased separately. Grades 4 and 5 have a separate student edition ($50–$55) and a teacher manual set ($185). Other ancillary materials, such as blackline master books, practice workbooks, and a test-practice generator, can be purchased separately.

Research

Twenty studies reviewed by the WWC investigated the effects of *Saxon Elementary School Math.* One study (Agodini et al., 2009) is a randomized controlled trial that meets WWC evidence standards. Two studies (Good, Bickel, & Howley, 2006; Resendez & Manley, 2005) are randomized controlled trials or quasi-experimental designs that meet WWC evidence standards with reservations. The remaining 17 studies do not meet either WWC evidence standards or eligibility screens.

Meets evidence standards

Agodini et al. (2009) examined the effects of *Saxon Elementary School Math* compared to three other curricula using a randomized controlled design involving 39 schools and 1,309 first-grade students from four school districts in Connecticut, Minnesota, New York, and Nevada. Schools were randomly assigned to use one of four curricula—*Saxon Elementary School Math; Investigations in Number, Data, and Space; Math Expressions;* or *Scott Foresman–Addison Wesley Mathematics*—for the entire school year. Each district contained at least one treatment school (using *Saxon Elementary School Math*) and at least one school using each of the three respective comparison curricula.

Meets evidence standards with reservations

Good, Bickel, and Howley (2006) used a quasi-experimental design to investigate the impacts of *Saxon Elementary School Math* with a sample of 1,476 kindergarten through third-grade students in 57 schools from across the United States. The authors matched a randomly selected sample of elementary schools currently using *Saxon Elementary School Math* to a group of comparison schools based on school size, type, grade-level configuration, and student demographics. Teachers in the comparison schools used a range of other curricula.

Resendez and Manley (2005) conducted a retrospective study that included 170 intervention schools in Georgia and 172 comparison schools that were matched to the intervention schools based on student demographics, geographical location, and baseline math performance on Georgia's Criterion-Referenced Competency Test (CRCT). The intervention schools used the *Saxon Elementary School Math* program recommended for each grade level in grades 1–8 between 2000 and 2005. The comparison schools used a variety of other curricula. The majority of comparison schools used traditional basal math curricula. One third of the schools used a mix of basal, investigative, and other approaches, and a small percentage used an investigative approach to teaching math. This intervention report presents the study's findings for grades 1–5.

Extent of evidence

The WWC categorizes the extent of evidence in each domain as small or medium to large (see the WWC Procedures and Standards Handbook, Appendix G). The extent of evidence takes into account the number of studies and the total sample size across the studies that meet WWC evidence standards with or without reservations.[6]

The WWC considers the extent of evidence for *Saxon Elementary School Math* to be medium to large for mathematics achievement for elementary school students.

6. The extent of evidence categorization was developed to tell readers how much evidence was used to determine the intervention rating, focusing on the number and size of studies. Additional factors associated with a related concept (external validity, such as the students' demographics and the types of settings in which studies took place) are not taken into account for the categorization. Information about how the extent of evidence rating was determined for *Saxon Elementary School Math* is in Appendix A6.

Effectiveness

Findings

The WWC review of interventions for Elementary School Math addresses student outcomes in mathematics achievement. The findings below present the authors' estimates and WWC-calculated estimates of the size and the statistical significance of the effects of *Saxon Elementary School Math* on students.[7] Of the three studies reviewed, one reported statistically significant positive effects. The remaining two studies showed indeterminate effects.

Agodini et al. (2009) reported statistically significant greater achievement on the Early Childhood Longitudinal Study–Kindergarten (ECLS-K) mathematics assessment for schools using the *Saxon Elementary School Math* program compared to schools using two of the other three comparison curricula. The WWC confirmed those results and also found that impacts for *Saxon Elementary School Math* were significantly greater than the three comparison curricula considered jointly.

Good, Bickel, and Howley (2006) did not report statistical significance findings for intent-to-treat impacts. Using supplemental results supplied by the authors, the WWC calculations found no statistically significant effect of *Saxon Elementary School Math* on the performance of kindergarten through third-grade students on the mathematics subtest of the Stanford Achievement Test, Ninth Edition (SAT 9). The effect size of 0.07 on the SAT 9 does not meet the WWC criteria for substantively important effects (an effect size of 0.25 or greater).

Resendez and Manley (2005) reported no significant effects of the *Saxon Elementary School Math* program on overall math achievement in grades 1–5, as measured by Georgia's CRCT. Using school-level data provided by the authors, the WWC confirmed that *Saxon Elementary School Math* did not have a statistically significant effect on math achievement at each grade level from first to fifth grade. Due to the lack of student-level data, the effect size and improvement index could not be calculated.

Rating of effectiveness

The WWC rates the effects of an intervention in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative. The rating of effectiveness takes into account four factors: the quality of the research design, the statistical significance of the findings, the size of the difference between participants in the intervention and the comparison conditions, and the consistency in findings across studies (see the WWC Procedures and Standards Handbook, Appendix E).

The WWC found *Saxon Elementary School Math* to have mixed effects for mathematics achievement for elementary school students

Improvement index

The WWC computes an improvement index for each individual finding. In addition, within each outcome domain, the WWC computes an average improvement index for each study and an average improvement index across studies (see the WWC Procedures and Standards Handbook, Appendix F). The improvement index represents the difference between the percentile rank of the average student in the intervention condition and the percentile rank of the average student in the comparison condition. Unlike the rating of effectiveness, the improvement index is entirely based on the size of the effect, regardless of the statistical significance of the effect, the study design, or the analysis. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.

The student-level improvement index could not be computed for one of the three studies because student-level standard deviations were not available. Across the remaining two studies,

7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see the WWC Procedures and Standards Handbook, Appendix C for clustering and the WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the cases of Agodini et al. (2009) and Resendez and Manley (2005), no corrections for clustering or multiple comparisons were needed. In the case of Good, Bickel, and Howley (2006), a correction for clustering was needed, so the significance levels may differ from those reported in the original study.

The WWC found *Saxon Elementary School Math* to have mixed effects for mathematics achievement for elementary school students *(continued)*

the average improvement index for mathematics achievement is +5 percentile points, with a range of –1 to +12 percentile points across findings.

Summary

The WWC reviewed 20 studies on *Saxon Elementary School Math* for elementary school students. One of these studies meets WWC evidence standards; two studies meet WWC evidence standards with reservations; the remaining 17 studies do not meet either WWC evidence standards or eligibility screens. Based on the three studies, the WWC found mixed effects on mathematics achievement for elementary school students. The conclusions presented in this report may change as new research emerges.

References

Meets WWC evidence standards

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.

Meets WWC evidence standards with reservations

Good, K., Bickel, R., & Howley, C. (2006). Saxon Elementary Math *program effectiveness study.* Charlestown, WV: Edvantia.

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 fall outside the Elementary School Math review protocol or do not meet WWC evidence standards

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 the intervention and comparison groups are not shown to be equivalent at baseline.

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

Cummins-Colburn, B. J. L. (2007). Differences between state-adopted textbooks and student outcomes on the Texas Assessment of Knowledge and Skills examination. (Doctoral dissertation, Touro University International). *Dissertation Abstracts International, 68*(06A), 168–2299. The study does not meet WWC evidence standards because the intervention and comparison groups are not shown to be equivalent at baseline.

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.

Hansen, E., & Greene, K. (2000). *A recipe for math. What's cooking in the classroom: Saxon or Traditional?* Retrieved May 4, 2006, from http://www.secondaryenglish.com/recipeformath.html. The study does not meet WWC evidence standards because the intervention and comparison groups are not shown to be equivalent at baseline.

Harcourt Achieve, Inc. (2005). *Case study research summaries of* Saxon Math. Retrieved December 15, 2009, 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.

References *(continued)*

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 the intervention and comparison groups are not shown to be equivalent at baseline.

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.

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—since the intervention was not implemented as designed.

Plato, J. (1998). *An evaluation of* Saxon Math *at Blessed Sacrament School.* Retrieved May 4, 2006, from http://lrs.ed.uiuc.edu/students/plato1/Final.html. The study is ineligible for review because it does not use a comparison group.

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 the intervention and comparison groups are not shown to be equivalent at baseline.

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 the intervention and comparison groups are not shown to be equivalent at baseline.

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 the intervention and comparison groups are not shown to be equivalent at baseline.

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly, 43*(3), 290–322. The study is ineligible for review because it is not a primary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

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

Appendix A1.1 Study characteristics: Agodini et al., 2009

Characteristic	Description
Study citation	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.
Participants	The researchers recruited 40 schools from four geographically dispersed districts with Title I schools. Each district had to include at least four schools willing to participate in the study, to support implementation of the study's four curricula in each district. Within each of the participating districts, the schools were randomly assigned to one of the four curricula prior to the start of the school year, thereby setting up an experiment in each district. Roughly 10 students were randomly selected for assessment from each first-grade classroom in the study schools. The 40 schools included 1,457 first-grade students from 134 classrooms. One school dropped out of the study, leaving 39 in the analysis sample. The analysis sample included 1,309 first-grade students in 131 classrooms. The relative effects of the curricula were calculated by comparing math achievement of students in the four curriculum groups at the end of the 2006–07 academic year. Sixty-nine percent of students were eligible for free or reduced-price lunch. Fifty-four percent of schools in the study were schoolwide Title I eligible, compared to 41 percent nationwide.
Setting	The four districts were located in Connecticut, Minnesota, New York, and Nevada. They included two districts in urban areas, one in a suburban area, and one in a rural area. Each district contained Title I schools.
Intervention	First-grade teachers implemented the *Saxon Math* curriculum published by Harcourt Achieve.
Comparison	Three other curricula were used in the study: (1) *Investigations in Number, Data, and Space (Investigations);* (2) *Math Expressions*; and (3) *Scott Foresman–Addison Wesley Mathematics (SFAW).* The authors note that a "business-as-usual" control group was not included because it would have contained a variety of curricula used by the participating districts, making it difficult to interpret effects of the individual curricula in the study.
Primary outcomes and measurement	The authors measured math achievement using the assessment developed for the National Center for Education Statistics' Early Childhood Longitudinal Study–Kindergarten Class of 1998–99 (ECLS-K). For a more detailed description of the outcome measure, see Appendix A2.
Staff/teacher training	Teachers in the study received training by the publishers of their assigned curriculum. All teachers received a one-to-two-day training at the start of the school year and follow-up training during the school year. Ninety-six percent attended follow-up training on their assigned curriculum.

Appendix A1.2 Study characteristics: Good, Bickel, & Howley, 2006

Characteristic	Description
Study citation	Good, K., Bickel, R., & Howley, C. (2006). Saxon Elementary Math *program effectiveness study.* Charlestown, WV: Edvantia.
Participants	Participants were 1,476 students between kindergarten and third grade from 57 schools. In spring 2005, Harcourt Achieve sent Edvantia researchers a spreadsheet containing the names of U.S. schools implementing the *Saxon Elementary School Math* program. Edvantia staff randomly selected schools to participate in the study. Of the 40 Saxon schools asked, 33 agreed. Twenty-four comparison schools were selected based on their similarities to the experimental schools on several measures, including school size; grade-level configuration; percentage of students eligible for free and reduced-price school lunch (the conventional education-research proxy measure for poverty); percentage of racial and ethnic minority students; migrant percentages; charter school designation; Title I school designation; locale, for example, urban, rural, large town, or small town; and geographic location. Data with which to identify matches were obtained from the U.S. Department of Education's National Center on Educational Statistics Common Core of Data for public schools from the 2003–04 school year.
Setting	The experimental and comparison schools were located across 16 states, including Alabama (1 school), Arizona (5 schools), California (6 schools), Georgia (3 schools), Indiana (1 school), North Carolina (9 schools), Nebraska (5 schools), Nevada (2 schools), New York (2 schools), Oklahoma (9 schools), Oregon (2 schools), Tennessee (2 schools), Texas (2 schools), Utah (1 school), Virginia (6 schools), and Washington (1 school).
Intervention	The intervention condition occurred over the 2005–06 school year. Teachers implemented the *Saxon Elementary School Math* program.
Comparison	Comparison-group teachers implemented a variety of other curricula, and some reported using skills that were part of the Saxon curriculum. The publishers of the programs tended to be Harcourt Brace, Houghton Mifflin, Silver Burdett Ginn, McGraw-Hill, and Scott Foresman.
Primary outcomes and measurement	The Stanford Achievement Test, Ninth Edition (SAT 9) was administered as the pretest and posttest measure of math achievement. Participating students completed only the math subtest of the SAT 9. In the fall, students took the appropriate grade-level versions of the SAT 9: the SESAT 1, SESAT 2, abbreviated Primary 1, or abbreviated Primary 2 tests, respectively, for kindergarten through third grade. The tests administered to K–3 students in the spring included the SESAT 2, abbreviated Primary 1, abbreviated Primary 2, and abbreviated Primary 3. The tests were administered by either the classroom teacher or the site coordinator. For a more detailed description of these outcome measures, see Appendix A2.
Staff/teacher training	Training is not described in the study.

Appendix A1.3 Study characteristics: Resendez & Manley, 2005

Characteristic	Description
Study citation	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.
Participants	The participants in this study were students in grades 1–8 in 170 intervention schools and 172 comparison schools that were matched based on student demographics. This intervention report focuses only on findings for grades 1–5, because grades 6–8 are outside of the scope of this review.[1] The authors selected Georgia schools that used the *Saxon Elementary School Math* curriculum between 2000 and 2005. The sample was obtained from the Georgia Department of Education. The authors note that per state policy, only school-level data could be released. 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 numbers of schools per grade are not mutually exclusive. Some of the schools contained multiple grades, so the numbers presented do not represent distinct clusters of schools.
Setting	The sample schools were distributed across the state of Georgia and represented a mixture of rural, urban, and suburban communities. The gender and racial compositions of the schools were similar in the intervention schools and comparison schools, with roughly equal gender distribution and more than half of the students white. Both study conditions were also similar in terms of the percent of students with disabilities, students with limited English proficiency, and students categorized as gifted.
Intervention	The *Saxon Elementary School Math* curriculum was used as a core curriculum in the intervention schools. The elementary schools in the sample used the version of the *Saxon Elementary School Math* program that was appropriate for each grade level, and participating schools had used the program for an average of three years (with a range of 1–15 years).
Comparison	The schools in the comparison group 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 third of the schools used curricula that were a mix of basal, investigative, and computer-based approaches. The authors reported no significant differences in baseline math performance between the Saxon and non-Saxon schools.
Primary outcomes and measurement	The outcome measure was 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. Fourth-grade students were tested in each school year from 1999–00 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 A2.
Staff/teacher training	No information was provided regarding the teacher training for the intervention.

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

Appendix A2 Outcome measures for the mathematics achievement domain

Outcome measure	Description
Early Childhood Longitudinal Study–Kindergarten (ECLS-K), Math Assessment	This is an individually administered, nationally normed assessment capable of measuring math achievement gains from kindergarten through grade 8. It was developed for the National Center for Education Statistics' Early Childhood Longitudinal Study–Kindergarten Class of 1998–99 (ECLS-K).
Stanford Achievement Test, Ninth Edition (SAT 9), Math Subtest	The SAT 9 math subtest is a nationally normed assessment published by Pearson Education. It is composed of two parts: problem solving and mathematics procedures. The SAT 9 math subtest was developed in alignment with the National Council of Teachers of Mathematics' *Curriculum and Evaluation Standards for School Mathematics*.[1]
Georgia's Criterion-Referenced Competency Test (CRCT),[2] Mathematics	As cited in Resendez and Manley (2005), the CRCT is a criterion-referenced test which is referenced 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.[3] 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. Five objectives are covered by the test: (1) numbers and number sense; (2) geometry and measurement; (3) patterns, relationships, and algebra; (4) computation and estimation; and (5) problem solving. The cut points are set by the state and take into account the difficulty of each specific objective.

1. See the product description at http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=E139A.
2. The original CRCT scores shown in the report are by objective. Upon request from the WWC, the author calculated the mean overall score across all objectives, controlling for pretest, for each grade.
3. Georgia Department of Education. (n.d.). *Criterion-referenced competency tests.* Retrieved November 17, 2009 from http://www.doe.k12.ga.us/ci_testing.aspx?PageReq=CI_TESTING_CRCT.

Appendix A3 Summary of study findings included in the rating for the mathematics achievement domain[1]

Outcome measure	Study sample	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Saxon Math* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Saxon Math* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Agodini et al., 2009 (randomized controlled trial)[7]								
ECLS-K	Grade 1 (versus *Investigations*)	19/636	47.36[8] (7.62)	44.87 (8.64)	2.49	0.30	Statistically significant	+12
ECLS-K	Grade 1 (versus *Math Expressions*)	18/618	45.27[8] (7.62)	45.45 (8.97)	–0.18	–0.02	ns	–1
ECLS-K	Grade 1 (versus *SFAW*)	20/663	46.21[8] (7.62)	44.28 (8.27)	1.93	0.24	Statistically significant	+10
Average for mathematics achievement (Agodini et al., 2009)[9]						**0.17**	**Statistically significant**	**+7**
Good, Bickel, & Howley, 2006[7]								
SAT 9	Grades K–3	57/1476	580.10[10] (63.37)	575.82[10] (58.66)	4.28	0.07	ns	+3
Average for mathematics achievement (Good, Bickel, & Howley, 2006)[9]						**0.07**	**ns**	**+3**
Resendez & Manley, 2005[7]								
CRCT	Grade 1	229/nr	86.26[11] (nr)	85.20[11] (nr)	1.06	na[12]	ns	na[12]
CRCT	Grade 2	229/nr	88.31[11] (nr)	86.86[11] (nr)	1.45	na[12]	ns	na[12]
CRCT	Grade 3	218/nr	86.94[11] (nr)	85.93[11] (nr)	1.01	na[12]	ns	na[12]
CRCT	Grade 4	210/nr	73.92[11] (nr)	71.39[11] (nr)	2.53	na[12]	ns	na[12]
CRCT	Grade 5	208/nr	82.86[11] (nr)	81.66[11] (nr)	0.80	na[12]	ns	na[12]
Average for mathematics achievement (Resendez & Manley, 2005)[9]						**na[12]**	**ns**	**na[12]**
Domain average for mathematics achievement across all studies[9]						**0.12**	**na**	**+5**

(continued)

Appendix A3 Summary of study findings included in the rating for the mathematics achievement domain[1] *(continued)*

ns = not statistically significant
na = not applicable
nr = not reported
ECLS-K = Early Childhood Longitudinal Survey–Kindergarten
SAT 9 = Stanford Achievement Test, Ninth Edition
CRCT = Georgia's Criterion-Referenced Competency Test
Investigations = *Investigations in Number, Data, and Space*
SFAW = *Scott Foresman–Addison Wesley Mathematics*

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the mathematics achievement domain. Subgroup and subtest findings from the same studies are not included in these ratings but are reported in Appendices A4.1 and A4.2, respectively.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see the WWC Procedures and Standards Handbook, Appendix B.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see the WWC Procedures and Standards Handbook, Appendix C for clustering and the WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the cases of Agodini et al. (2009) and Resendez and Manley (2005), no corrections for clustering or multiple comparisons were needed. In the case of Good, Bickel, and Howley (2006), a correction for clustering was needed, so the significance levels may differ from those reported in the original study.
8. The treatment group coefficient represents the sum of the unadjusted control group mean and the hierarchical linear modeling (HLM) coefficient for the difference between the two groups in the study.
9. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.
10. These figures represent difference-in-differences adjusted means not reported in the original study. They are based on results provided by the author(s) in response to a request by the WWC. The difference-in-differences adjustment subtracts baseline differences between the study groups from the post-intervention differences between the groups. The author query for additional information was required because the original study presented only analyses of the impact of the amount of treatment received, rather than intent-to-treat effects. The means for the *Saxon* and comparison groups differed by 0.07 standard deviations at baseline.
11. The original study reported only means for subtests. The value reported here is the mean across those subtests. For subtest results, see Appendix A4.2.
12. Student-level standard deviations were not available for this study. School-level standard deviations for the intervention group were 6.60 for grade 1, 6.39 for grade 2, 6.50 for grade 3, 8.51 for grade 4, and 6.94 for grade 5. School-level standard deviations for the comparison group were 6.80 for grade 1, 7.35 for grade 2, 7.15 for grade 3, 11.83 for grade 4, and 8.93 for grade 5. Because the student-level effect sizes and improvement indices could not be computed, the magnitude of the effect size was not considered for rating purposes. Note, however, that the average school-level effect size for the study is zero, and student-level effect sizes are typically smaller than school-level effect sizes. The statistical significance for this study is comparable to other studies and is included in the intervention rating. For further details, please see the WWC Procedures and Standards Handbook, Appendix B.

Appendix A4.1 Summary of subgroup findings for the mathematics achievement domain[1]

Outcome measure	Study sample[4]	Sample size (students)[5]	Authors' findings from the study[2]: Mean outcome (standard deviation[3]): *Saxon Math* group	Authors' findings from the study[2]: Mean outcome (standard deviation[3]): Comparison group	WWC calculations: Mean difference (*Saxon Math* – comparison)	WWC calculations: Effect size[6]	WWC calculations: Statistical significance[7] (at $\alpha = 0.05$)	WWC calculations: Improvement index[8]
Agodini et al., 2009[9]								
Comparison 1: *Saxon Math* compared with *Investigations in Number, Data, and Space*								
ECLS-K	Lowest third	179	nr[10]	nr[10]	nr[10]	0.71	Statistically significant	+26
ECLS-K	Middle third	159	nr[10]	nr[10]	nr[10]	0.17	ns	+7
ECLS-K	Highest third	298	nr[10]	nr[10]	nr[10]	0.15	ns	+6
ECLS-K	Up to 40% FRP	378	nr[10]	nr[10]	nr[10]	0.31	ns	+12
ECLS-K	Greater than 40% FRP	258	nr[10]	nr[10]	nr[10]	0.37	ns	+14
Comparison 2: *Saxon Math* compared with *Math Expressions*								
ECLS-K	Lowest third	206	nr[10]	nr[10]	nr[10]	0.32	ns	+13
ECLS-K	Middle third	205	nr[10]	nr[10]	nr[10]	–0.20	ns	–8
ECLS-K	Highest third	207	nr[10]	nr[10]	nr[10]	–0.08	ns	–3
ECLS-K	Up to 40% FRP	316	nr[10]	nr[10]	nr[10]	–0.01	ns	0
ECLS-K	Greater than 40% FRP	302	nr[10]	nr[10]	nr[10]	–0.02	ns	–1
Comparison 3: *Saxon Math* compared with *Scott Foresman–Addison Wesley Elementary Mathematics*								
ECLS-K	Lowest third	201	nr[10]	nr[10]	nr[10]	0.56	Statistically significant	+21
ECLS-K	Middle third	195	nr[10]	nr[10]	nr[10]	–0.01	ns	0
ECLS-K	Highest third	267	nr[10]	nr[10]	nr[10]	0.18	ns	+7
ECLS-K	Up to 40% FRP	346	nr[10]	nr[10]	nr[10]	0.30	ns	+12
ECLS-K	Greater than 40% FRP	317	nr[10]	nr[10]	nr[10]	0.20	ns	+8

ns = not statistically significant
nr = not reported
ECLS-K = Early Childhood Longitudinal Study–Kindergarten
FRP = Free/reduced-price meal eligibility

(continued)

Appendix A4.1 Summary of subgroup findings for the mathematics achievement domain[1] *(continued)*

1. This appendix presents subgroup findings for measures that fall in the mathematics achievement domain. Total group scores were used for rating purposes and are presented in Appendix A3.
2. The subgroup sample sizes were obtained through communication with the study authors.
3. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
4. Subgroups were defined using school characteristics. Subgroups defined using baseline student achievement data are defined as students in schools with average math scores in the lowest, middle, and highest third of the study's school-level distribution. Subgroups based on socioeconomic status are examined for students in schools with up to 40% of students eligible for free or reduced-price meals, compared to schools with more than 40% of students eligible for free or reduced-price meals.
5. The authors provided only the number of students, not the number of teachers or schools in each subgroup.
6. Positive effect sizes favor the intervention group; negative effect sizes favor the comparison group. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
8. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
9. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Agodini et al. (2009), no corrections for clustering or multiple comparisons were needed.
10. The study provided effect sizes and statistical significance for subgroup outcomes produced though hierarchical linear modeling (HLM) that were calculated in accordance with WWC standards. Adjusted means were not available and are consequently omitted in this table. The table includes the effect sizes and statistical significance reported in the study, along with improvement index values calculated by the WWC based on the study-reported effect sizes.

Appendix A4.2 Summary of subscale findings for the mathematics achievement domain[1]

Outcome measure	Study sample	Sample size (schools)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Saxon Math* group[3]	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group[3]	WWC calculations: Mean difference[4] (*Saxon Math* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Resendez & Manley, 2005 (quasi-experimental design)[8]								
CRCT: Numbers and number sense	Grade 1	229	89.53 (nr)	88.52 (nr)	1.01	na[9]	ns	na[9]
CRCT: Geometry and measurement	Grade 1	229	90.34 (nr)	90.29 (nr)	0.05	na[9]	ns	na[9]
CRCT: Patterns, relations, and algebra	Grade 1	229	87.88 (nr)	86.28 (nr)	1.60	na[9]	ns	na[9]
CRCT: Computation and estimation	Grade 1	229	78.93 (nr)	77.43 (nr)	1.50	na[9]	ns	na[9]
CRCT: Problem solving	Grade 1	229	84.64 (nr)	83.49 (nr)	1.15	na[9]	ns	na[9]
CRCT: Numbers and number sense	Grade 2	229	88.57 (nr)	86.62 (nr)	1.95	na[9]	ns	na[9]
CRCT: Geometry and measurement	Grade 2	229	91.46 (nr)	92.36 (nr)	–0.90	na[9]	ns	na[9]
CRCT: Patterns, relations, and algebra	Grade 2	229	87.05 (nr)	83.58 (nr)	3.47	na[9]	Statistically significant	na[9]
CRCT: Computation and estimation	Grade 2	229	86.93 (nr)	85.83 (nr)	1.10	na[9]	ns	na[9]
CRCT: Problem solving	Grade 2	229	87.54 (nr)	85.93 (nr)	1.61	na[9]	ns	na[9]
CRCT: Numbers and number sense	Grade 3	218	89.74 (nr)	88.24 (nr)	1.50	na[9]	ns	na[9]
CRCT: Geometry and measurement	Grade 3	218	93.60 (nr)	92.24 (nr)	1.36	na[9]	ns	na[9]
CRCT: Patterns, relations, and algebra	Grade 3	218	86.26 (nr)	85.90 (nr)	0.36	na[9]	ns	na[9]
CRCT: Statistics and computation	Grade 3	218	87.13 (nr)	85.83 (nr)	1.30	na[9]	ns	na[9]

(continued)

Appendix A4.2 Summary of subscale findings for the mathematics achievement domain[1] *(continued)*

Outcome measure	Study sample	Sample size (schools)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Saxon Math* group[3]	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group[3]	WWC calculations: Mean difference[4] (*Saxon Math* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
CRCT: Computation and estimation	Grade 3	218	86.81 (nr)	85.71 (nr)	1.10	na[9]	ns	na[9]
CRCT: Problem solving	Grade 3	218	78.11 (nr)	77.64 (nr)	0.47	na[9]	ns	na[9]
CRCT: Numbers and number sense	Grade 4	210	71.47 (nr)	70.85 (nr)	0.62	na[9]	ns	na[9]
CRCT: Geometry and measurement	Grade 4	210	79.22 (nr)	78.16 (nr)	1.06	na[9]	ns	na[9]
CRCT: Patterns, relations, and algebra	Grade 4	210	69.76 (nr)	67.70 (nr)	2.06	na[9]	ns	na[9]
CRCT: Statistics and computation	Grade 4	210	82.15 (nr)	80.17 (nr)	1.98	na[9]	ns	na[9]
CRCT: Computation and estimation	Grade 4	210	73.12 (nr)	67.65 (nr)	5.47	na[9]	Statistically significant	na[9]
CRCT: Problem solving	Grade 4	210	67.81 (nr)	63.83 (nr)	3.98	na[9]	Statistically significant	na[9]
CRCT: Numbers and number sense	Grade 5	208	79.74 (nr)	77.31 (nr)	2.43	na[9]	ns	na[9]
CRCT: Geometry and measurement	Grade 5	208	80.77 (nr)	81.54 (nr)	–0.77	na[9]	ns	na[9]
CRCT: Patterns, relations, and algebra	Grade 5	208	76.16 (nr)	74.56 (nr)	1.60	na[9]	ns	na[9]
CRCT: Statistics and computation	Grade 5	208	79.82 (nr)	81.52 (nr)	–1.70	na[9]	ns	na[9]
CRCT: Computation and estimation	Grade 5	208	88.74 (nr)	86.62 (nr)	2.12	na[9]	ns	na[9]
CRCT: Problem solving	Grade 5	208	89.55 (nr)	88.43 (nr)	1.12	na[9]	ns	na[9]

ns = not statistically significant
na = not applicable
nr = not reported

(continued)

Appendix A4.2 Summary of subscale findings for the mathematics achievement domain[1] *(continued)*

1. This appendix presents subscale findings for measures that fall in the mathematics achievement domain. Total scale scores were used for rating purposes and are presented in Appendix A3.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The intervention group and control group means are pretest adjusted and provided by the authors. They may differ from the means reported in the original study.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see the WWC Procedures and Standards Handbook, Appendix B.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors. No correction was required for clustering within classrooms or schools, or for multiple comparisons.
9. Student-level standard deviations and improvement indices were not available for this study. School-level standard deviations, which were requested by the WWC and provided by the first study author, ranged from 4.50 to 10.32 across grade levels and subtests in the intervention group and from 5.41 to 14.75 across grade levels and subtests in the comparison group. Because student-level standard deviations were not available, student-level effect sizes and improvement indices could not be computed. However, the statistical significance of the findings in Resendez and Manley (2005) is comparable to other studies and is reported in this appendix. For further details, see the WWC Procedures and Standards Handbook, Appendix B.

Appendix A5 *Saxon Elementary School Math* rating for the mathematics achievement domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of mathematics achievement, the WWC rated *Saxon Elementary School Math* as having mixed effects for elementary school students. The remaining ratings (no discernable effects, potentially negative effects, and negative effects) were not considered, as *Saxon Elementary School Math* was assigned the highest applicable rating.

Rating received

Mixed effects: Evidence of inconsistent effects as demonstrated through either of the following criteria.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing 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.

  Not met. *Saxon Elementary School Math* had no studies showing negative effects on achievement.

OR

- Criterion 2: At least one study showing a statistically significant or substantively important effect, and more studies showing an *indeterminate* effect than showing a statistically significant or substantively important effect.

  Met. One study of *Saxon Elementary School Math* showed a statistically significant positive effect, and two studies showed indeterminate effects.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a *strong* design.

  Not met. Only one study of *Saxon Elementary School Math* showed a statistically significant positive effect.

AND

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. No studies of *Saxon Elementary School Math* showed negative effects.

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Met. One study of *Saxon Elementary School Math* showed a statistically significant positive effect.

AND

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Not met. Among the three studies of *Saxon Elementary School Math* that met WWC evidence standards, more showed indeterminate effects (two studies) than positive effects (one study).

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A6 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Mathematics achievement	3	325	na	Medium to large

na = not applicable/not studied. Total number of students not reported in all of the relevant studies.

1. A rating of "medium to large" requires at least two studies and two schools across studies in one domain and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small." For more details on the extent of evidence categorization, see the WWC Procedures and Standards Handbook, Appendix G.