

Appendix

Appendix A1.1 Study characteristics: Resendez & Azin, 2006 (randomized controlled trial)

Characteristic	Description
Study citation	Resendez, M., & Azin, M. (2006). <i>Saxon Math randomized control trial: Final report</i> . Jackson, WY: PRES Associates, Inc.
Participants	The study sample included 543 sixth-, seventh-, and eighth-grade students in 25 classes (14 intervention classes with 303 students and 11 control classes with 240 students) from two Ohio schools during the 2005–06 school year. ¹ Of the total study sample, about 49% were male (49% intervention and 50% control), 8% were special education students (7% intervention and 10% control), and 25% received free or reduced-price lunch (19% intervention and 33% control). Approximately 81% were Caucasian (91% intervention and 70% control), 17% were African-American (8% intervention and 28% control), and 2% were of other racial/ethnic classifications (1% intervention and 3% control). None of the study sample was limited English proficient. Based on pretest percentile rankings, 18% were in the lowest quartile (10% intervention and 29% control), 29% were in the highest quartile (43% intervention and 11% control), and the remaining 53% were in the two middle quartiles (48% intervention and 60% control). ² The intervention and control groups were formed through random assignment at the classroom level. There were seven teachers across the two schools. Six of the seven participating teachers taught at least one intervention and one control class. The seventh teacher was randomly assigned to teach one <i>Saxon Math</i> class. The analysis sample included approximately 490 students for the TerraNova Math Total test (approximately 270 intervention and 220 control) and approximately 470 students for the TerraNova Math Computation Total test (approximately 260 intervention and 210 control). ^{3,4}
Setting	The study took place in one junior high school located in a suburban area of southwestern Ohio and one middle school located in a large city in northeastern Ohio. The junior high school served students in grades 7–9; the middle school served students in grades 5–8.
Intervention	Students were taught using one of four <i>Saxon Math</i> curricula: <i>Saxon Math 7/6</i> , <i>Saxon Math 8/7</i> , <i>Saxon Algebra ½</i> , or <i>Saxon Algebra 1</i> during the 2005–06 school year. Teachers were expected to implement key program components that included: warm-up activities; teaching a new lesson concept; lesson practice; “mixed practice” that reviewed and built upon previous concepts and prepared students for upcoming lessons; teacher-directed, whole-class investigations; and test day activities. ⁵ According to the study authors, two of the seven teachers typically did not follow the implementation guidelines; however, the majority of intervention classrooms (10 out of 14) covered at least 83% of the 120 or more <i>Saxon Math</i> lessons. ⁶
Comparison	The math curricula used in control classrooms varied by site. In the junior high school, control classrooms were taught using an unspecified traditional basal program for math instruction that used a modular approach, emphasized real-world application, and incorporated a variety of activities including exploration, modeling, and using tools such as technology to communicate math. In the middle school, control classrooms were taught using a variety of resources consisting primarily of teacher-created materials based on district and state guidelines; the curriculum also included an internet-based math program, traditional chapter-based textbooks, and other supplemental math resources.
Primary outcomes and measurement	The primary outcome measures were the TerraNova Math Total and the TerraNova Math Computation Total of the CTB McGraw Hill TerraNova Basic Multiple Assessment with Plus Test. The pretest administration occurred between September and October 2005; the posttest administration occurred between May and June 2006. For a more detailed description of these outcome measures, see Appendix A2.
Staff/teacher training	Teachers received about three hours of training before implementing their <i>Saxon Math</i> curricula; the <i>Saxon Math</i> trainer covered the program's philosophy, key components, and curriculum support materials. A follow-up training session conducted in October/November provided teachers with one-on-one suggestions for using the program.

(continued)

Appendix A1.1 Study characteristics: Resendez & Azin, 2006 (continued)

1. Initial study participants included 549 students (307 intervention and 242 control) who were enrolled at the beginning of the fall semester. The study authors included in their final study sample only students who remained in the study throughout the year. Six students left during the school year (four intervention and two control), resulting in a study sample of 543 students still enrolled at the end of the spring semester.
2. Because random assignment was well executed, the WWC categorizes the study as a randomized controlled trial. The study authors statistically controlled for baseline differences between the intervention and control groups in their analysis.
3. The study authors did not report the sample sizes for the multilevel analyses reviewed here that controlled for baseline differences in pretest and demographic characteristics. Growth analyses of intervention students with both pre- and posttest scores includes 269 intervention students for the TerraNova Math Total and 261 intervention students for the TerraNova Math Computation Total tests; similar analyses were not reported for control students. The authors reported the number of control students missing pretest scores (11 for the Math Total test and 21 for the Math Computation Total test) and the number missing posttest scores (24 for the Math Total test and 28 for the Math Computation Total test), but did not indicate the number missing both pretest and posttest.
4. The study authors also conducted subgroup analyses. Because the authors did not report sufficient information to calculate effect sizes, the WWC excluded the subgroup analyses from this review.
5. In the junior high school, approximately 45 minutes of additional math instruction were provided to students in need of remediation. In particular, 38 seventh-grade remedial students received a “double-dose” of instruction: 22 students were in a *Saxon Math* class and a *Saxon Math* lab, 11 students were in a *Saxon Math* class and a control math lab, and five students were in a control math class and a control math lab. According to the study authors, this additional math instruction did not affect the findings for the two primary measures.
6. In the middle school, three of the four teachers stopped or reduced their use of their respective curricula for at least one to two months in order to focus on preparation for state testing in March; this disruption reduced use of both the *Saxon Math* and the control programs. The disruption in the *Saxon Math* group included stopping or reducing use of the *Saxon* program for 6 to 10 weeks.

Appendix A1.2 Study characteristics: Peters, 1992 (randomized controlled trial)

Characteristic	Description
Study citation	Peters, K. G. (1992). Skill performance comparability of two algebra programs on an eighth-grade population. <i>Dissertation Abstracts International</i> , 54(01), 77A. (UMI No. 9314428)
Participants	The study sample included 36 eighth-grade students in two classrooms from one junior high school during the 1991–92 school year. All of the students were “math talented” based on teacher recommendations and prior academic achievement. No information is provided on the specific thresholds that were used in delineating the math-talented criteria; however, all students scored at or above the 87th percentile on the California Achievement Test total math battery. Of the total sample, 56% were female (58% intervention and 53% control) and 44% were male (42% intervention and 47% control). Students were randomly assigned to one of two classrooms (one intervention classroom and one control classroom). However, the assignment of students was altered after random assignment; the analysis sample included 19 students in <i>Saxon Math</i> group and 17 students in the <i>University of Chicago School Mathematics Project (UCSMP) Algebra</i> group. ¹ The same teacher taught both classrooms. ²
Setting	The study took place in one junior high school in Nebraska. The district borders two large cities (Lincoln and Omaha) and has a mix of students living in rural and suburban locations.
Intervention	Students were taught using <i>Saxon Algebra 1</i> (1981) during the 1991–92 school year. Students in this group participated in daily sessions for one academic year. In each session, the teacher introduced a new concept incrementally, and students had opportunities to practice the new concept and past concepts during each session. Students were assessed every fifth lesson.
Comparison	Students were taught using the <i>UCSMP Algebra</i> curriculum. The <i>UCSMP Algebra</i> program was developed based on National Council of Teachers of Mathematics (NCTM) objectives that emphasized problem-solving skills, reading comprehension, use of technology, and relevant lessons with real-world applications. Each lesson is organized into an introduction of the concept, a reading section that explains the process, and real-life problem situations.
Primary outcomes and measurement	The primary outcome measure was the Orleans-Hanna Algebra Prognosis Test. ³ The pretest administration occurred in August 1991, and the posttest administration occurred in May 1992. For a more detailed description of this outcome measure, see Appendix A2.
Staff/teacher training	The teacher who taught both study groups did not have prior experience with the intervention or control curricula but had read extensively about both teaching formats. The teacher participated in a one-week summer workshop on <i>UCSMP Algebra</i> , and in two additional one-day workshops given by local consultants on the curricula used in this study.

1. The author indicates that a random selection of numbers was used to divide participants between the intervention and control groups. However, the assignment of students was altered to accommodate scheduling difficulties and student requests for other course offerings. The study author demonstrated the baseline equivalence of the *Saxon Math* and *UCSMP Algebra* groups at pretest.
2. Because both the intervention and control curricula were monitored on a weekly basis by the researcher to help maintain the integrity of implementation, and because there is no indication in the study that the teacher was biased toward one of the conditions, this design was accepted for review.
3. The study author described only the Orleans-Hanna Algebra Prognosis Test as the measure of student math achievement. The study also examined four study-generated criterion unit tests, not from the Orleans-Hanna Algebra Prognosis Test, designed to descriptively measure student understanding of algebraic components. However, the author did not provide information on the reliability or validity of these four tests. Accordingly, analyses based on these four unit tests were not considered in this version of the report.

Appendix A1.3 Study characteristics: Crawford & Raia, 1986 (quasi-experimental design)

Characteristic	Description
Study citation	Crawford, J., & Raia, F. (1986). <i>Analyses of eighth grade math texts and achievement</i> . Oklahoma City, OK: Oklahoma City Public Schools, Planning, Research, and Evaluation Department.
Participants	The study sample included 78 eighth-grade students (39 intervention and 39 comparison) taught by four teachers in four Oklahoma middle schools during the 1984–85 school year. ¹ The authors did not report demographic information. To create a comparison group that was similar at baseline to the intervention group, the researchers conducted a stratified matching procedure based on pretest total math score on the California Achievement Test (CAT) at the student level within teachers to match a comparison student with each student in the intervention group. When more than one student from the comparison group matched a student in the treatment group, the student match was selected at random.
Setting	The study took place in four middle schools in Oklahoma City Public Schools.
Intervention	Students were taught using the <i>Saxon Algebra ½</i> (1983) textbook during the 1984–85 school year. Information about the level of implementation was not provided. The intervention was implemented by four teachers. Each of these teachers taught both intervention classes and comparison classes.
Comparison	Students were taught using the math textbook in place prior to the pilot study. The comparison group used the textbook Scott-Foresman <i>Mathematics</i> (1980).
Primary outcomes and measurement	The primary outcome measure was the total math score on the CAT. Pretest data was from the year-end administration of the CAT in 1984, and posttest data came from the end-of-year test administration in 1985. For a more detailed description of this outcome measure, see Appendix A2.
Staff/teacher training	Information on teacher training was not provided.

1. The study authors reported on three “studies”: one that compared *Saxon* students to all other eighth-grade students in Oklahoma City Public Schools, one that compared *Saxon* students in the pilot schools to non-*Saxon* students attending the same schools, and one that compared *Saxon* students taught by teachers who used both the *Saxon* and non-*Saxon* textbooks to non-*Saxon* students taught by the same teachers. The third study included an analysis in which the authors matched students on pretest within strata formed by teachers who used both *Saxon* and non-*Saxon* textbooks. This WWC review focuses only on this third, within-teacher matched comparison analysis because it is the only analysis for which the authors demonstrated baseline equivalence.

Appendix A1.4 Study characteristics: Resendez, Fahmy, & Manley, 2005, Cohort A (quasi-experimental design)

Characteristic	Description
Study citation	Resendez, M., Fahmy, A., & Manley, M. A. (2005). Cohort A. <i>The relationship between using Saxon Middle School Math and student performance on Texas statewide assessments</i> . Available online from Harcourt Achieve: http://saxonpublishers.hmhco.com/HA/correlations/pdf/s/SXMath_Middle_TX_research_web.pdf . (Cohort A).
Participants	Cohort A was from Sample 1 of the study. Sample 1 included three cohorts of over 16,000 sixth-, seventh-, and eighth-grade students from 30 schools (15 intervention and 15 comparison). Of the total sample, about 46% were Caucasian (46% intervention and 45% comparison), 43% were Hispanic (42% intervention and 43% comparison), 11% were African-American (11% intervention and 10% comparison), 6% were limited English proficient (5% intervention and 7% comparison), 15% were special education (15% intervention and 14% comparison), 49% were female (49% intervention and 48% comparison), and 46% were economically disadvantaged (43% intervention and 48% comparison). To create the matched comparison group of schools for the 15 <i>Saxon Math</i> schools in the initial sample, the Texas Education Agency (TEA) identified 40 matched comparison schools from which 15 were randomly selected. The intervention and comparison schools were matched on demographic characteristics including race, ethnicity, poverty, English language proficiency, and percentage of mobile students. The analysis sample for Cohort A (grade 8 of Sample 1) included 3,054 students (1,472 intervention and 1,582 comparison) from 25 schools (12 intervention and 13 comparison) in grades 6–8 during the 1998–99 through 2000–01 school years. ¹
Setting	The study took place in Texas middle schools serving sixth-, seventh-, and eighth-grade students between 1998–99 and 2000–01.
Intervention	Students in Cohort A (grade 8 of Sample 1) were taught using <i>Saxon Math</i> curricula (<i>Saxon 7/6</i> , <i>Saxon 8/7</i> , or <i>Saxon Algebra ½</i>) during the 1998–99 (grade 6), 1999–2000 (grade 7), and 2000–01 (grade 8) school years.
Comparison	The majority of comparison schools used core basal math curricula, which generally consist of a chapter-based approach to math instruction. Two schools used an investigative approach with an emphasis on making connections among various mathematical topics and between math and problems in other disciplines.
Primary outcomes and measurement	The primary outcome for Cohort A (grade 8 of Sample 1) was the Texas Assessment of Academic Skills (TAAS) – Texas Learning Index (TLI) math score. The pretest measure was the TLI math score from grade 5 (taken in the spring of 1998). The posttest measure was an average of the TLI math scores from grade 6 (taken in the spring of 1999), grade 7 (taken in the spring of 2000), and grade 8 (taken in the spring of 2001). For a more detailed description of this outcome measure, see Appendix A2.
Staff/teacher training	Information on teacher training was not provided.

1. The study authors excluded from the analysis three intervention schools that were not using *Saxon Math* during the 1998–99 school year; two comparison schools were subsequently dropped. In addition, the WWC excluded Cohorts B and C from this review because they were included only in an analysis of tenth-grade math performance. Because it is unknown whether the intervention and comparison groups for these cohorts attended similar schools in ninth- and tenth-grade, it is impossible to determine whether the effects can be attributed solely to *Saxon Math*; therefore, the WWC excluded these cohorts from this review.

Appendix A1.5 Study characteristics: Resendez, Fahmy, & Manley, 2005, Cohort F (quasi-experimental design)

Characteristic	Description
Study citation	Resendez, M., Fahmy, A., & Manley, M. A. (2005). Cohort F. <i>The relationship between using Saxon Middle School Math and student performance on Texas statewide assessments</i> . Available online from Harcourt Achieve: http://saxonpublishers.hmhco.com/HA/correlations/pdf/s/SXMath_Middle_TX_research_web.pdf . (Cohort F).
Participants	Cohort F was from Sample 3 of the study. Sample 3 included three cohorts of over 18,000 sixth-, seventh-, and eighth-grade students from 30 schools (15 intervention and 15 comparison). Of the total sample, about 42% were Caucasian (38% intervention and 45% comparison), 44% were Hispanic (47% intervention and 41% comparison), 13% were African-American (14% intervention and 12% comparison), 6% were limited English proficient (5% intervention and 6% comparison), 15% were special education (14% intervention and 15% comparison), 49% were female (49% intervention and 49% comparison), and 52% were economically disadvantaged (48% intervention and 55% comparison). To create the matched comparison group of schools for the 15 <i>Saxon Math</i> schools in the initial sample, the Texas Education Agency (TEA) identified 40 matched comparison schools from which 15 were randomly selected. The intervention and comparison schools were matched on demographic characteristics including race, ethnicity, poverty, English language proficiency, and percentage of mobile students. The analysis sample for Cohort F (grade 6 of Sample 3) included 2,933 students (1,526 intervention and 1,407 comparison) from 20 schools (10 intervention and 10 comparison) in grade 6 during the 2003–04 school year. ¹
Setting	The study took place in Texas middle schools serving sixth-, seventh-, and eighth-grade students during the 2003–04 school year.
Intervention	Students in Cohort F (grade 6 of Sample 3) were taught using <i>Saxon Math</i> curriculum during the 2003–04 school year; the majority used <i>Saxon 7/6</i> , and the remainder used <i>Saxon 8/7</i> .
Comparison	The majority of comparison schools used core basal math curricula, which generally consist of a chapter-based approach to math instruction. Two schools used an investigative approach with an emphasis on making connections among various mathematical topics and between math and problems in other disciplines.
Primary outcomes and measurement	The primary outcome for Cohort F (grade 6 of Sample 3) was the Texas Assessment of Knowledge and Skills (TAKS). The pretest measure was the TAKS math performance scores from grade 5 (taken in the spring of 2003). The posttest measure was the TAKS math performance scores from grade 6 (taken in the spring of 2004). For a more detailed description of this outcome measure, see Appendix A2.
Staff/teacher training	Information on teacher training was not provided.

1. The study authors excluded from the analysis five intervention schools that were not using *Saxon Math* during the 2003–04 school year; five comparison schools were subsequently dropped. In addition, the WWC excluded Cohorts G and H from review because pre-*Saxon* math achievement data were not available for these students and, consequently, baseline equivalence could not be established.

Appendix A2 Outcome measures for the math achievement domain

Outcome measure	Description
TerraNova Math Total (MC/CR) Scale Score	The TerraNova Math Total (MC/CR) scale score is based on a portion of CTB McGraw Hill's TerraNova Basic Multiple Assessment (Level 16-6th grade, Level 17-7th grade, and Level 18-8th grade) with Plus Test. This 90-minute portion of the TerraNova standardized test contains 41 or 42 multiple-choice and constructed-response items, consisting of a few computational problems but mostly word problems. The objectives tested include: number and number relations; computation and estimation; measurement; geometry and spatial sense; data, statistics, and probability; patterns, function, and algebra (seventh grade and eighth grade only); problem solving and reasoning; and communication (as cited in Resendez & Azin, 2006).
TerraNova Math Computation Total Scale Score	The TerraNova Math Computation Total scale score is based on a portion of CTB McGraw Hill's TerraNova Basic Multiple Assessment (Level 16-6th grade, Level 17-7th grade, and Level 18-8th grade) with Plus Test. This 20-minute portion of the TerraNova standardized test has 20 multiple-choice, computational problems. The objectives tested include: multiplying whole numbers (sixth grade only), dividing whole numbers (sixth grade only), decimals (sixth and seventh grade), fractions, integers (seventh and eighth grade), percents (seventh and eighth grade), and order of operations (seventh and eighth grade) (as cited in Resendez & Azin, 2006).
TerraNova Objective Performance Indices (OPI)	In addition to reporting the two total math scale scores noted above, CTB McGraw Hill provides objective performance indices (OPI) for each objective measured on the TerraNova Basic Multiple Assessment with Plus Test. The OPI is an estimate of the number of items a student or group of students could be expected to answer correctly if there had been 100 such items on the test for that objective (as cited in Resendez & Azin, 2006).
Orleans-Hanna Algebra Prognosis Test	The nationally normed Orleans-Hanna Algebra Prognosis test consists of 60 items and is used to predict student success in future algebra study by comparing the actual test score with the students' most recent math grades. In contrast to an achievement test, students are required to answer questions by following a procedure or set of operations using mathematical or verbal expressions parallel to but different from those contained in the model lessons. This test is often used to predict the ability to succeed in a first-year algebra course of study. Raw scores converted by the teacher to standard scores were used in the analysis (as cited in Peters, 1992).
California Achievement Test (CAT) General Mathematics Exam	The California Achievement Test is a standardized achievement test. The mathematics section includes subtests on mathematics computation and mathematics concepts and applications. Normal Curve Equivalent (NCE) scores were used in the analysis (outcome measure used in Crawford & Raia, 1986; information obtained from the test publisher website; edition of the test was not reported).
California Achievement Test (CAT) Math Concepts and Applications	The Mathematics Concepts and Applications section is a subtest of the California Achievement Test General Mathematics Exam that assesses the ability to perform fundamental mathematics operations, apply mathematical concepts, and use a variety of problem-solving strategies. Normal Curve Equivalent (NCE) scores were used in the analysis (outcome measure used in Crawford & Raia, 1986; information obtained from the test publisher website; edition of the test was not reported).
California Achievement Test (CAT) Math Computation	The Mathematics Computation section is a subtest of the California Achievement Test General Mathematics Exam that assesses the ability to perform fundamental mathematics operations, apply mathematical concepts, and use a variety of problem-solving strategies. Normal Curve Equivalent (NCE) scores were used in the analysis (outcome measure used in Crawford & Raia, 1986; information obtained from the test publisher website; edition of the test was not reported).
Texas Learning Index (TLI) Math Score (based on the Texas Assessment of Academic Skills)	The Texas Assessment of Academic Skills (TAAS) is a criterion-referenced state test that measures problem-solving and critical-thinking skills. The Texas Learning Index (TLI) is an outcome metric, based on student performance on the TAAS, allowing for comparisons between administrations and between grades. The TAAS was used in Texas from 1990–2002; it was replaced by the Texas Assessment of Knowledge and Skills in 2003 (as cited in Resendez, Fahmy, & Manley, 2005, Cohort A).
Texas Assessment of Knowledge and Skills (TAKS) Math Test	The Texas Assessment of Knowledge and Skills (TAKS) was used in the analysis of Cohort F (grade 6 of Sample 3). This test covers numbers, operations, and quantitative reasoning; patterns, relationships, and algebraic reasoning; geometry and spatial reasoning; concepts and uses of measurement; probability and statistics; and mathematical processes and tools. A scaled score was used in the analysis (as cited in Resendez, Fahmy, & Manley, 2005, Cohort F).

Appendix A3 Summary of study findings included in the rating for the math achievement domain¹

Outcome measure	Study sample	Sample size (clusters/ students)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ³ (<i>Saxon Math</i> – comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
			<i>Saxon Math</i> group	Comparison group				
Resendez & Azin, 2006⁷								
TerraNova Math Total	Grades 6–8	25/~490	679.11 ⁸ (40.47)	673.32 ⁹ (38.11)	5.7 ⁹	0.15 ¹⁰	ns	+6
TerraNova Math Computation Total	Grades 6–8	25/~470	674.00 ⁸ (49.37)	662.84 ⁹ (46.47)	11.16	0.23 ¹⁰	ns	+9
Average for math achievement (Resendez & Azin, 2006)¹¹						0.19	ns	+8
Peters, 1992⁷								
Orleans-Hanna Prognosis Test	Grade 8 (math talented)	2/36	95.67 ¹² (4.53)	95.06 ¹³ (4.09)	0.61	0.14	ns	+6
Average for math achievement (Peters, 1992)¹¹						0.14	ns	+6
Crawford & Raia, 1986⁷								
California Achievement Test	Grade 8	4/78	55.56 (11.86)	50.72 (11.75)	4.84	0.41	ns ¹⁴	+16
Average for math achievement (Crawford & Raia, 1986)¹¹						0.41	ns	+16
Resendez, Fahmy, & Manley, 2005, Cohort A⁷								
Texas Learning Index Score	Grade 8	25/3,054	83.95 (6.99) ¹⁵	82.98 (7.62) ¹⁵	0.97	0.13	ns	+5
Average for math achievement (Resendez, Fahmy, & Manley, 2005, Cohort A)¹¹						0.13	ns	+5
Resendez, Fahmy, & Manley, 2005, Cohort F⁷								
Texas Assessment of Knowledge and Skills Math Scale Score	Grade 6	20/2,933	2,229.02 (225.89) ¹⁶	2,174.49 (205.10) ¹⁶	54.53	0.25	ns	+10
Average for math achievement (Resendez, Fahmy, & Manley, 2005, Cohort F)¹¹						0.25	ns	+10
Domain average for math achievement across all studies¹¹						0.22	na	+9

ns = not statistically significant

na = not applicable

(continued)

Appendix A3 Summary of study findings included in the rating for the math achievement domain¹ (continued)

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the math achievement domain. Subtest findings from the same studies are not included in these ratings, but are reported in Appendix A4.
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 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 WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the cases of Resendez and Azin (2006) and Peters (1992), no corrections for clustering were needed. In the cases of Crawford and Raia (1986), Resendez, Fahmy, and Manley (2005, Cohort A), and Resendez, Fahmy, and Manley (2005, Cohort F), a correction for clustering was needed, so the significance levels may differ from those reported in the original studies. For the *Saxon Math* studies summarized here, no corrections for multiple comparisons were needed.
8. The *Saxon Math* value from Resendez and Azin (2006) is the unadjusted comparison group mean plus the program coefficient from the hierarchical linear modeling (HLM) analysis.
9. The comparison group mean from Resendez and Azin (2006) is unadjusted.
10. The effect sizes reported here differ from those reported in the original study due to differences in the effect-size formulas used.
11. 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. The domain improvement index is calculated from the average effect size.
12. The *Saxon Math* group value from Peters (1992) is the unadjusted comparison group mean plus the difference in mean gains between the intervention (*Saxon Math*) and comparison groups.
13. The comparison group mean from Peters (1992) is unadjusted.
14. The previous intervention report accepted the author's assertion that this finding was statistically significant. However, WWC calculations do not confirm the statistical significance in either the study or the previous intervention report.
15. In the case of Resendez, Fahmy, and Manley (2005, Cohort A), the means for the *Saxon Math* group and comparison group are repeated measures ANCOVA adjusted means during grade 8 for Cohort A (grade 8 of Sample 1). The mean difference between these two scores represents the effect of three years' exposure to *Saxon Math*. The standard deviations are the unadjusted standard deviations for grade 8 provided to the WWC by the study authors in response to a query.
16. The study authors provided the WWC with unadjusted standard deviations for the *Saxon Math* and comparison groups in response to a query.

Appendix A4 Summary of subscale findings for the math achievement domain¹

Outcome measure	Study sample	Sample size (clusters/students)	Authors' findings from the study					
			Mean outcome (standard deviation) ²		WWC calculations			
			Saxon Math group	Comparison group	Mean difference ³ (Saxon Math—comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
Resendez & Azin, 2006⁷								
TerraNova OPI Number and Number Relations	Grades 6–8	25/499	66.22 ⁸ (20.71)	64.53 ⁹ (22.17)	1.69	0.08	ns	+3
TerraNova OPI Computation and Estimation	Grades 6–8	25/498	59.45 ⁸ (18.65)	57.00 ⁹ (19.06)	2.45	0.13	ns	+5
TerraNova OPI Measurement	Grades 6–8	25/502	43.29 ⁸ (22.52)	38.27 ⁹ (22.52)	5.02	0.22	ns	+9
TerraNova OPI Geometry & Spatial Sense	Grades 6–8	25/503	48.84 ⁸ (20.63)	46.72 ⁹ (19.94)	2.12	0.10	ns	+4
TerraNova OPI Data, Statistics & Probability	Grades 6–8	25/503	54.58 ⁸ (19.81)	52.23 ⁹ (20.43)	2.35	0.12	ns	+5
TerraNova OPI Patterns, Functions & Algebra	Grades 7 & 8	nr/396	55.30 ⁸ (18.10)	53.24 ⁹ (18.53)	2.06	0.11	ns	+4
TerraNova OPI Problem Solving and Reasoning	Grades 6–8	25/503	46.08 ⁸ (20.65)	41.03 ⁹ (19.57)	5.05	0.25	ns	+10
TerraNova OPI Communication	Grades 6–8	25/484	44.07 ⁸ (19.63)	43.80 ⁹ (20.57)	0.27	0.01	ns	+1
TerraNova OPI Multiplication	Grade 6	nr/105	67.27 ⁸ (25.10)	50.78 ⁹ (27.20)	16.49	0.62	ns	+23
TerraNova OPI Division	Grade 6	nr/102	59.69 ⁸ (27.48)	40.84 ⁹ (25.10)	18.85	0.71	ns	+26
TerraNova OPI Decimals	Grades 6 & 7	nr/212	66.00 ⁸ (19.13)	56.84 ⁹ (20.09)	9.16	0.47	ns	+18
TerraNova OPI Fractions	Grades 6–8	25/491	46.59 ⁸ (26.48)	39.75 ⁹ (25.30)	6.84	0.26	ns	+10

(continued)

Appendix A4 Summary of subscale findings for the math achievement domain (continued)

Outcome measure	Study sample	Sample size (clusters/ students)	Authors' findings from the study					
			Mean outcome (standard deviation) ²		Mean difference ³ (Saxon Math —comparison)	WWC calculations		
			Saxon Math group	Comparison group		Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
TerraNova OPI Integers	Grades 7 & 8	nr/387	60.45 ⁸ (24.71)	50.84 ⁹ (27.73)	9.61	0.37	ns	+14
TerraNova OPI Percents	Grades 7 & 8	nr/384	57.69 ⁸ (26.42)	46.25 ⁹ (23.62)	11.44	0.45	ns	+17
TerraNova OPI Order of Operations	Grades 7 & 8	nr/387	72.60 ⁸ (17.76)	68.99 ⁹ (23.53)	3.61	0.18	ns	+7
Crawford & Raia, 1986								
CAT Math Computation	Grade 8	4/78	57.66 (13.35)	51.44 (14.14)	6.22	0.45	ns	+17
CAT Math Concepts	Grade 8	4/78	53.18 (12.44)	50.00 (12.40)	3.18	0.25	ns	+10

ns = not statistically significant
nr = not reported
CAT = California Achievement Test
OPI = Objective Performance Indices

1. This appendix presents subscale findings for measures that fall in math achievement. 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. 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 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 results favorable to 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 WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Resendez and Azin (2006), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study. In the case of Crawford and Raia (1986), a correction for clustering was needed, so the significance levels may differ from those reported in the original study.
8. The Saxon Math group value from Resendez and Azin (2006) is the unadjusted comparison group mean plus the program coefficient from the hierarchical linear modeling (HLM) analysis.
9. The comparison group mean from Resendez and Azin (2006) is unadjusted.

Appendix A5 Saxon Math rating for the math achievement domain

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

For the outcome domain of math achievement, the WWC rated *Saxon Math* as having mixed effects. The remaining ratings (no discernible effects, potentially negative, and negative) were not considered, as *Saxon 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. No studies showed a statistically significant or substantively important negative effect.

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. Two studies showed substantively important positive effects, and three 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. No studies showed statistically significant positive effects.

AND

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

Met. No studies showed statistically significant or substantively important 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. Two studies showed substantively important positive effects.

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. Three studies showed indeterminate effects, and two studies showed substantively important effects.

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	
Math achievement	5	52	>6500	Medium to large

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.