



WWC Intervention Report

A summary of findings from a systematic review of the evidence



Primary Mathematics

May 2017*

Saxon Math

Intervention Description¹

Saxon Math is a curriculum for students in grades K–12. The amount of new math content students receive each day is limited and students practice concepts every day. New concepts are developed, reviewed, and practiced cumulatively rather than in discrete chapters or units. This review focuses on studies of *Saxon Math*'s primary courses, which include kindergarten through pre-algebra.

Research²

The What Works Clearinghouse (WWC) identified five studies of *Saxon Math* that both fall within the scope of the Primary Mathematics topic area and meet WWC group design standards.³ All five studies meet WWC group design standards with reservations. Together, these studies included 8,855 students in grades 1–3 and 6–8 in 149 schools across at least 18 states.⁴

According to the WWC review, the extent of evidence for *Saxon Math* on the mathematics test scores of students in primary courses was medium to large for the mathematics achievement domain, the only domain examined for studies reviewed under the Primary Mathematics topic area.⁵ (See the Effectiveness Summary on p. 5 for more details of effectiveness by domain.)

Effectiveness

Saxon Math had mixed effects on mathematics test scores of students in primary courses.

Table 1. Summary of findings⁶

Outcome domain	Rating of effectiveness	Improvement index (percentile points)		Number of studies	Number of students	Extent of evidence
		Average	Range			
Mathematics achievement	Mixed effects	+8	–1 to +16	5	8,855	Medium to large

Report Contents

Overview	p. 1
Intervention Information	p. 2
Research Summary	p. 3
Effectiveness Summary	p. 5
References	p. 7
Research Details for Each Study	p. 12
Outcome Measures for Each Domain	p. 24
Findings Included in the Rating for Each Outcome Domain	p. 25
Supplemental Findings for Each Outcome Domain	p. 27
Endnotes	p. 29
Rating Criteria	p. 32
Glossary of Terms	p. 33

This intervention report presents findings from a systematic review of *Saxon Math* conducted using the WWC Procedures and Standards Handbook (version 3.0) and the Primary Mathematics review protocol (version 3.1).

Intervention Information

Background

Saxon Math was originally developed by John Saxon. It is distributed by Houghton Mifflin Harcourt Supplemental Publishers. Address: Houghton Mifflin Harcourt Pre-K-12, 9205 Southpark Center Loop, Orlando, FL, 32819. Email: greatservice@hnhco.com. Website: www.hnhco.com. Telephone: (800) 225-5425. Fax: (800) 269-5232.

Intervention details

At each grade level, *Saxon Math* consists of at least 120 daily lessons and 12 investigation activities. Each lesson has three components:

- The teacher introduces one or more new math ideas daily, using examples and mathematical conversations, with a focus on integrating new ideas and concepts with ones previously introduced.
- The teacher guides students on practice problems relating to the new concepts.
- Students individually engage in written practice that aims to help them master new skills and maintain mastery of concepts previously taught.

Students complete written, cumulative assessments after every five lessons. The results of these assessments provide teachers with data for instructional decision making and provide feedback for students and parents. Students also have opportunities to demonstrate mastery of math content through in-depth investigations and performance tasks that require students to apply their mathematical knowledge and skills to real-world problems.

The primary curriculum includes *Saxon Math Primary K-3* and *Saxon Math Intermediate 3-5* for elementary grades, and *Saxon Math Courses 1, 2, and 3* for grades 6, 7, and 8, respectively. The publisher is currently selling the second edition and Common Core edition of *Saxon Math*. The publisher's website describes each of these current editions of the curriculum.

Cost

As of November 2016, the costs for curriculum materials were as follows:

- For *Saxon Math Primary K-3*, each set of teacher's materials costs \$276.20 to \$285.50, and student kits cost \$856.90 to \$995.05 for 24 students.
- For *Saxon Math Intermediate 3-5*, each set of teacher's materials costs \$278.90 for a hard-copy version. A teacher technology package is available for \$175.15 (for a 1-year subscription) and includes the Teacher's Manual eBook and various electronic teacher and planning resources. The student edition costs \$19.20 for the online version for a 1-year subscription, \$58.60 for a 6-year online subscription, or \$80.40 for the hard-copy version. A combined *Saxon Math Intermediate 3-5* online student/teacher edition costs \$1,168.85 for a 1-year subscription or \$5,849.40 for a 6-year subscription.
- For *Saxon Math Courses 1, 2, and 3* (grades 6–8), the teacher's manual costs \$131.75 for a hard copy, \$33.30 for a 1-year subscription to an online edition, or \$99.75 for a 6-year online subscription. The student edition for each course costs \$82.20 per student for a hard copy, \$65.75 for an eBook, \$20.60 for a 1-year subscription to an online edition, or \$61.65 for a 6-year online subscription. A combined *Saxon Math Courses 1-3* online student/teacher edition costs \$1,168.85 for a 1-year subscription or \$5,849.40 for a 6-year subscription.

Other materials, such as student workbooks, instructional presentations, and manipulative kits, are available and range in price. More detailed cost information is available from the publisher.

Research Summary

The WWC identified 26 eligible studies that investigated the effects of *Saxon Math* on the mathematics achievement of students in primary courses. An additional 33 studies were identified but do not meet WWC eligibility criteria (see the Glossary of Terms in this document for a definition of this term and other commonly used research terms) for review in this topic area. Citations for all 59 studies are in the References section, which begins on p. 7.

The WWC reviewed the 26 eligible studies against group design standards. None of the 26 studies is a randomized controlled trial that meets WWC group design standards without reservations. One study is a randomized controlled trial that meets WWC group design standards with reservations, and four studies use quasi-experimental designs that meet WWC group design standards with reservations. This report summarizes those five studies. The remaining 21 studies do not meet WWC group design standards.

Summary of studies meeting WWC group design standards without reservations

No studies of *Saxon Math* met WWC group design standards without reservations.

Summary of studies meeting WWC group design standards with reservations

Agodini, Harris, Seftor, Remillard, and Thomas (2013) conducted a cluster, or group-based, randomized controlled trial assigning one of four math curricula—*Saxon Math*; *Investigations in Number, Data, and Space (Investigations)*; *Math Expressions*; or *Scott-Foresman Addison Wesley Mathematics (SFAW)*—to 111 elementary schools in 12 school districts to use as their core math curriculum in first and second grades. The 111 schools enrolled in the study in either the 2006–07 or 2007–08 school year, and 58 of the schools participated in the study for 2 consecutive years. During the second year of the study, the publisher revised *SFAW* and renamed it *enVisionMATH*. The group of schools that used *enVisionMATH* in the second year is labeled *SFAW/enVisionMATH*. The study examined 1- and 2-year effects of the curricula on student math achievement using the Early Childhood Longitudinal Study–Kindergarten math assessment. The WWC based its effectiveness rating on a finding that compared 2-year outcomes in 12 schools that used *Saxon Math* to those in 46 schools that used other curricula. This sample included 2,045 students in the 58 schools that participated in the study for 2 years. Assumptions about equivalence in random assignment may not hold because schools were randomly assigned to curriculum before the student sample was identified. Families could know a school’s curriculum assignment and, in theory, could decide to move into or out of a school based on that knowledge. The study demonstrated equivalence on the analytic sample and therefore, meets WWC group design standards with reservations. Findings after 1 year based on all 111 schools are included as supplemental findings in Appendix D. The study did not specify the edition of *Saxon Math* it used, but indicated that the materials were copyrighted in 2005 and 2008.

Crawford and Raia (1986) used a quasi-experimental design to examine the effects of *Saxon Math* on eighth-grade students in four middle schools in one school district in the 1984–85 school year.⁷ Four eighth-grade teachers in the study schools taught at least one class using *Saxon Math* and at least one class using *Scott-Foresman Mathematics*. The authors grouped 78 students into 39 pairs. Each pair included two students with similar pretest math scores, one student in a *Saxon Math* class and one in a *Scott-Foresman Mathematics* class taught by the same teacher. The study used the California Achievement Test math assessment to measure eighth-grade student achievement. The study did not specify the edition of *Saxon Math* it used, but indicated that the materials were copyrighted in 1983.

Table 2. Scope of reviewed research

Grade	1, 2, 3, 6, 7, 8
Delivery method	Whole class
Program type	Curriculum

Good, Bickel, and Howley (2006) used a quasi-experimental design to compare 33 schools implementing *Saxon Math* to 24 schools using a variety of other math curricula in the 2005–06 school year. The study randomly selected intervention group schools, located in 16 states, from all schools in the United States implementing *Saxon Math*. The authors matched comparison schools to intervention schools based on school characteristics, including school size, percentage of students eligible for free or reduced-price meals, racial and ethnic makeup of the students, and school Title I status. Within each study school, students in K–3 classrooms participated in the study. The authors used the Math Problem Solving subtest of the Stanford Achievement Test, Ninth Edition, as the outcome measure, which they administered to students in grades 2 and 3. The sample used for study analysis included 745 second- and third-grade students (411 in 33 intervention schools and 334 in 24 comparison schools).⁸ The study did not specify the edition of *Saxon Math* it used.

Resendez, Fahmy, and Manley (2005, Sample 1) used a quasi-experimental design to compare schools implementing *Saxon Math* in grades 6–8 to schools using other math curricula with a chapter-based approach to math instruction.⁹ The study matched comparison schools to the intervention schools based on demographic characteristics including race, ethnicity, poverty, English language proficiency, and percentage of mobile students (that is, students who transfer frequently between schools during the school year). The analytic sample included 1,472 students from 12 intervention schools who used *Saxon Math* for 2 years in grades 6 and 7, and 1,582 students from 13 comparison schools in the same grades during the 1998–99 and 1999–2000 school years. The study measured student achievement in grades 6 and 7 using the Texas Assessment of Academic Skills Texas Learning Index. The study did not specify the edition of *Saxon Math* it used.

Resendez, Fahmy, and Manley (2005, Sample 3) used a quasi-experimental design to compare schools implementing *Saxon Math* in grades 6–8 to schools using other math curricula with a chapter-based approach to math instruction. The study matched comparison schools to the intervention schools based on demographic characteristics including race, ethnicity, poverty, English language proficiency, and percentage of mobile students. The analytic sample included 1,526 students from 10 intervention schools who used *Saxon Math* in grade 6, and 1,407 students from 10 comparison schools in the same grade during the 2003–04 school year. The authors measured student achievement in grade 6 using the Texas Assessment of Knowledge and Skills math scale score. The study did not specify the edition of *Saxon Math* it used.

Effectiveness Summary

The WWC review of *Saxon Math* for the Primary Mathematics topic area includes student outcomes in one domain: mathematics achievement. The following findings present the authors’ estimates and WWC-calculated estimates of the size and statistical significance of the effects of *Saxon Math* on mathematics achievement for students in primary courses. Additional comparisons are available as supplemental findings in Appendix D. The supplemental findings do not factor into the intervention’s rating of effectiveness. For a more detailed description of the rating of effectiveness and extent of evidence criteria, see the WWC Rating Criteria on p. 32.

Summary of effectiveness for the mathematics achievement domain

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

Rating of effectiveness	Criteria met
Mixed effects <i>Evidence of inconsistent effects.</i>	In the five studies that reported findings, the estimated impact of the intervention on outcomes in the <i>mathematics achievement</i> domain was positive and substantively important in two studies and indeterminate in three studies.
Extent of evidence	Criteria met
Medium to large	Five studies that included 8,855 students in 149 schools across at least 18 states reported evidence of effectiveness in the <i>mathematics achievement</i> domain.

Five studies that met WWC group design standards with reservations reported findings in the mathematics achievement domain.

Agodini et al. (2013) compared *Saxon Math* against each of the three other curricula. The authors reported, and the WWC confirmed, a positive and statistically significant difference between *Saxon Math* and *Investigations*, one of the three comparison group curricula, in the mathematics achievement domain. The authors reported, and the WWC confirmed, no statistically significant or substantively important differences between *Saxon Math* and the other two curricula (*Math Expressions* and *SFAW/enVisionMATH*). For the purposes of providing an overall rating of effectiveness, the WWC pooled the three comparison curricula groups and compared the pooled group to *Saxon Math*; the WWC found the difference was neither statistically significant nor substantively important. The WWC characterizes this study finding as an indeterminate effect.

Crawford and Raia (1986) reported a positive and statistically significant difference between *Saxon Math* and the comparison group in the mathematics achievement domain. However, after correcting for clustering, the WWC found that this difference was not statistically significant. The effect size is large enough to be considered substantively important according to WWC criteria. The WWC characterizes this study finding as a substantively important positive effect.

Good et al. (2006) reported a positive difference between *Saxon Math* and the comparison group in the mathematics achievement domain. The study did not report the statistical significance of this finding. After correcting for school clustering, the WWC found that the difference was neither statistically significant nor substantively important. The WWC characterizes this study finding as an indeterminate effect.

Resendez et al. (2005, Sample 1) reported a positive difference between *Saxon Math* and the comparison group in the mathematics achievement domain. The authors did not report the statistical significance of this finding. After correcting for school clustering, the WWC found that the difference was neither statistically significant nor substantively important. The WWC characterizes this study finding as an indeterminate effect.

Resendez et al. (2005, Sample 3) reported a positive and statistically significant difference between *Saxon Math* and the comparison group in the mathematics achievement domain. However, after correcting for school clustering, the WWC found that this difference was not statistically significant. The effect size is large enough to be considered substantively important according to WWC criteria. The WWC characterizes this study finding as a substantively important positive effect.

Thus, for the mathematics achievement domain, two studies show substantively important positive effects and three studies show indeterminate effects. This results in a rating of mixed effects, with a medium to large extent of evidence.

References

Studies that meet WWC group design standards without reservations

None.

Studies that meet WWC group design standards with reservations

Agodini, R., Harris, B., Seftor, N., Remillard, J., & Thomas, M. (2013). *After two years, three elementary math curricula outperform a fourth* (NCEE 2013-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED544185>

Additional sources:

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

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools* (NCEE 2009-4052). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED504418>

Agodini, R., Harris, B., Seftor, N., Remillard, J., & Thomas, M. (2013). *Technical appendix: After two years, three elementary math curricula outperform a fourth* (NCEE 2013-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED544187>

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders* (NCEE 2011-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED512551>

Clements, D. H., Agodini, R., & Harris, B. (2013). *Instructional practices and student math achievement: Correlations from a study of math curricula* (NCEE 2013-4020). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED544189>

Clements, D. H., Agodini, R., & Harris, B. (2013). *Technical appendix: Data and methodological approach* (NCEE 2013-4020). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED544192>

Remillard, J. T., Harris, B., & Agodini, R. (2014). The influence of curriculum material design on opportunities for student learning. *ZDM: The International Journal on Mathematics Education*, 46(5), 735–749.

Crawford, J., & Raia, F. (1986). *Analyses of eighth grade math texts and achievement*. Oklahoma City, OK: Oklahoma City Public Schools Planning, Research, and Evaluation Department.

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

Resendez, M., Fahmy, A., & Manley, M. A. (2005). *The relationship between using Saxon Middle School Math and student performance on Texas statewide assessments* [Sample 1]. Jackson, WY: PRES Associates, Inc.

Resendez, M., Fahmy, A., & Manley, M. A. (2005). *The relationship between using Saxon Middle School Math and student performance on Texas statewide assessments* [Sample 3]. Jackson, WY: PRES Associates, Inc.

Studies that do not meet WWC group design standards

Baldree, C. L. P. (2003). *The effectiveness of two mathematical instructional programs on the mathematics growth of eighth grade students* (Unpublished doctoral dissertation). University of Georgia, Athens. The study does not

meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

- Bell, G. (2012). The effects of *Saxon Math* instruction on middle school students' mathematics achievement. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 73(4-A), 1339. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.
- Bhatt, R., & Koedel, C. (2012). Large-scale evaluations of curricular effectiveness: The case of elementary mathematics in Indiana. *Educational Evaluation and Policy Analysis*, 34(4), 391–412. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.
- Calvery, R., Bell, D., & Wheeler, G. (1993). *A comparison of selected second and third graders' math achievement: Saxon vs. Holt*. New Orleans, LA: Mid-South Educational Research Association. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.
- Cummins-Colburn, B. J. L. (2007). Differences between state-adopted textbooks and student outcomes on the Texas Assessment of Knowledge and Skills examination. *Dissertation Abstracts International*, 68(06A), 168-2299. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.
- Educational Research Institute of America. (2009). *A longitudinal analysis of state mathematics scores for Indiana schools using Saxon Math* (Report No. 362). Bloomington, IN: Author. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the 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 does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.
- Hansen, E., & Greene, K. (2000). *A recipe for math. What's cooking in the classroom: Saxon or traditional?* The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.
- 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 group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.
- Lafferty, J. F. (1996). The links among mathematics text, students' achievement, and students' mathematics anxiety: A comparison of the incremental development and traditional texts. *Dissertation Abstracts International*, 56(08), 3014A. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.
- Additional source:**
- Lafferty, J. F. (1994). *The links among mathematics text, students' achievement, and students' mathematics anxiety: A comparison of the incremental development and traditional texts* (Unpublished doctoral dissertation). Widener University, Chester, PA.
- Nguyen, K., Elam, P., & Weeter, R. (1993). *The 1992-93 Saxon Mathematics Program evaluation report*. Oklahoma City, OK: Oklahoma City Public Schools. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.
- Rentschler, R. V. (1994). The effects of Saxon's incremental review on computational skills and problem-solving achievement of sixth-grade students. *Dissertation Abstracts International*, 56(2), 484A. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Resendez, M., & Azin, M. (2006). *Saxon Math randomized control trial: Final report*. Jackson, WY: PRES Associates, Inc. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

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, Inc. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

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

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, Inc. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Resendez, M., & Azin Manley, M. (2005). *The relationship between using Saxon Elementary and Middle School Math and student performance on Georgia statewide assessments*. Jackson, WY: PRES Associates, Inc. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

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, Inc. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Roan, C. (2012). *A comparison of elementary mathematics achievement in Everyday Math and Saxon Math schools in Illinois* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3507509) The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Roberts, F. H. (1994). The impact of Saxon Mathematics Program on group achievement test scores. *Dissertation Abstracts International*, 55(06), 1498A. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Severns, L. D. (2014). *Saxon Math and student achievement: A multiyear investigation* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1662809559) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Additional source:

Severns, L. D. (2014). Differences in grade 8 students' math achievement as a function of Saxon Math instruction. *The Online Journal of New Horizons in Education*, 4(4), 20–26.

Walsh, T. J. (2009). The effect of Saxon Math on student achievement of sixth-grade students. *Dissertation Abstracts International*, 70(06A), 135–1966. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

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

Abrams, B. J. (1989). *A comparison study of the Saxon Algebra 1 text* (Unpublished doctoral dissertation). University of Colorado at Boulder. The study is ineligible for review because it is out of scope of the protocol.

Andrus, H. A. (2005). *Metacognitive instruction in the realm of sixth grade Saxon Math* (Unpublished doctoral dissertation). Mount Mary College, Milwaukee, WI. The study is ineligible for review because it does not use an eligible design.

- Aquino, A., & Zoet, C. (1985). Reinforcement in Algebra I: A study in the use of the Saxon Algebra I textbook. *Mathematics in Michigan*, 23–28. The study is ineligible for review because it is out of scope of the protocol.
- Baroody, A. J., Purpura, D. J., Eiland, M. D., & Reid, E. E. (2014). Fostering first graders' fluency with basic subtraction and larger addition combinations via computer-assisted instruction. *Cognition and Instruction*, 32(2), 159–197. The study is ineligible for review because it is out of scope of the protocol.
- Bolser, S., & Gilman, D. A. (2003). *Saxon Math, Southeast Fountain Elementary School: Effective or ineffective?* Retrieved from <https://eric.ed.gov/?id=ED474537> The study is ineligible for review because it does not use an eligible design.
- Christofori, P. (2005). *The effect of direct instruction math curriculum on higher-order problem solving* (Unpublished doctoral dissertation). University of South Florida. The study is ineligible for review because it is out of scope of the protocol.
- Clay, D. W. (1998). *A study to determine the effects of a non-traditional approach to algebra instruction on student achievement* (Unpublished master's thesis). Salem-Teikyo University, Salem, WV. Retrieved from <https://eric.ed.gov/?id=ED428963> The study is ineligible for review because it is out of scope of the protocol.
- Clewell, B. C., Cosentino de Cohen, C., Campbell, P. B., & Perlman, L. (2005). *Review of evaluation studies of mathematics and science curricula and professional development models*. Washington, DC: The Urban Institute. Retrieved from <http://www.urban.org/> This study is ineligible for review because it does not use an eligible design.
- Denson, P. S. (1989). A comparison of the effectiveness of the Saxon and Dolciani texts and theories about the teaching of high school algebra. *Dissertation Abstracts International*, 50, 10A. The study is ineligible for review because it is out of scope of the protocol.
- Doabler, C. T., Clarke, B., Fien, H., Baker, S. K., Kosty, D. B., & Cary, M. S. (2015). The science behind curriculum development and evaluation: Taking a design science approach in the production of a Tier 2 mathematics curriculum. *Learning Disability Quarterly*, 38(2), 97–111. The study is ineligible for review because it is out of scope of the protocol.
- Educational Research Institute of America. (2009). *A longitudinal analysis of state mathematics scores for Florida schools using Saxon Math* (Report No. 365). Washington, DC: Author. The study is ineligible for review because it does not use an eligible design.
- Educational Research Institute of America. (2009). *A longitudinal analysis of state mathematics scores for Oklahoma schools using Saxon Math* (Report No. 363). Washington, DC: Author. The study is ineligible for review because it does not use an eligible design.
- Fitzpatrick, S. B. (2001). An exploratory study of the implementation of an educational technology in two eighth grade mathematics classes. *Dissertation Abstracts International*, 62(06), 2082A. The study is ineligible for review because it is out of scope of the protocol.
- Harcourt Achieve, Inc. (2005). *Case study research summaries of Saxon Math*. Retrieved from <http://saxonpublishers.hmhco.com/> The study is ineligible for review because it does not use an eligible design.
- Harris, K. L. (2008). *Saxon Math: An analysis for middle school students at-risk of low performance* (Unpublished doctoral dissertation). Capella University, Minneapolis, MN. The study is ineligible for review because it does not use an eligible design.
- Imrisek, J. P. (1989). *Incremental development: A more effective means of mathematics instruction?* (Unpublished master's thesis). Bloomsburg University, PA. The study is ineligible for review because it is out of scope of the protocol.
- Johnson, D. M., & Smith, B. (1987). An evaluation of Saxon's Algebra text. *Journal of Educational Research*, 81(2), 97–102. The study is ineligible for review because it is out of scope of the protocol.
- 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 an eligible design.
- Lawrence, L. K. (1992). *The long-term effects of an incremental development model of instruction upon student achievement and student attitude toward mathematics* (Unpublished doctoral dissertation). University of Tulsa, OK. The study is ineligible for review because it is out of scope of the protocol.

- Mayers, K. S. (1995). *The effect of using the Saxon Algebra I textbook on the achievement of ninth-grade Algebra I students from 1989–1993* (Unpublished doctoral dissertation). Delta State University, Cleveland, MS. The study is ineligible for review because it is out of scope of the protocol.
- McBee, M. (1982). *Dolciani versus Saxon: A comparison of two algebra I textbooks with high school students*. Oklahoma City, OK: Oklahoma City Public Schools. The study is ineligible for review because it is out of scope of the protocol.
- McNeil, N., Grandau, L., Knuth, E., Alibali, M., Stephens, A., Hattikudur, S., & Krill, D. (2006). Middle-school students' understanding of the equal sign: The books they read can't help. *Cognition and Instruction*, 24(3), 367. The study is ineligible for review because it does not use an eligible design.

Additional source:

- McNeil, N. M., Grandau, L., Stephens, A. C., Krill, D. E., Alibali, M. W., & Knuth, E. J. (2004). Middle-school students' experience with the equal sign: Saxon Math \neq Connected Mathematics. In D. McDougall (Ed.), *Proceedings of the XXVI Annual Conference of the North American Chapter of the International Group for the Psychology of Mathematics Education (PME-NA), Toronto, Canada* (Vol. 1, pp. 271–276). Columbus, OH: ERIC.
- Peters, K. G. (1992). *Skill performance comparability of two algebra programs on an eighth-grade population* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9314428) The study is ineligible for review because it is out of scope of the protocol.
- Pierce, R. D. (1984). *A quasi-experimental study of Saxon's Incremental Development Model and its effects on student achievement in first-year algebra* (Unpublished doctoral dissertation). University of Tulsa, OK. The study is ineligible for review because it is out of scope of the protocol.
- Plato, J. (1998). *An evaluation of Saxon Math at Blessed Sacrament School*. Retrieved from <http://lrs.ed.uiuc.edu/>. The study is ineligible for review because it does not use an eligible design.
- Resendez, M., Fahmy, A., & Manley, M. A. (2005). *The relationship between using Saxon Middle School Math and student performance on Texas statewide assessments*. [Sample 2] Jackson, WY: PRES Associates, Inc. The study is ineligible for review because it does not use an eligible design.
- Sanders, B. B. (1997). *The effects of using the Saxon Mathematics method of instruction vs. a traditional method of mathematical instruction on the achievement of high school juniors*. Americus: Georgia Southwestern State University. The study is ineligible for review because it is out of scope of the protocol.
- Saxon Publishers. (2004). *Scientific research base for Saxon Math K-12: Foundational research and program efficacy studies*. Norman, OK: Author. Retrieved from <http://www.saxonpublishers.com/> The study is ineligible for review because it does not use an eligible design.
- Saxon, J. (1982). Incremental development: A breakthrough in mathematics. *Phi Delta Kappan*, 63(4), 482–484. Retrieved from <https://eric.ed.gov/?&id=EJ259471> The study is ineligible for review because it is out of scope of the protocol.
- Silvious, N. B. (2008). *Effects of Saxon Math program of instruction on the mathematics achievement of students with learning disabilities in Grades 2 through 8*. Chester, PA: Widener University. The study is ineligible for review because it is out of scope of the protocol.
- Slavin, R. E., & Lake, C. (2007). Effective programs in elementary mathematics: A best-evidence synthesis. *The Best Evidence Encyclopedia*, 1(2). Retrieved from <http://www.bestevidence.org/> The study is ineligible for review because it does not use an eligible design.
- Additional source:**
- Slavin, R. E., & Lake, C. (2009). *Effective programs for elementary mathematics: A best evidence synthesis. Educator's summary*. Retrieved from <http://www.bestevidence.org/>
- 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 use an eligible design.
- Williams, D. D. (1986). *The incremental method of teaching Algebra I*. Kansas City: University of Missouri. The study is ineligible for review because it is out of scope of the protocol.

Appendix A.1: Research details for Agodini et al. (2013)

Agodini, R., Harris, B., Seftor, N., Remillard, J., & Thomas, M. (2013). *After two years, three elementary math curricula outperform a fourth* (NCEE 2013-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED544185>

*Additional sources:*¹⁰

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. Retrieved from <https://eric.ed.gov/?&id=ED504418>

Agodini, R., Harris, B., Seftor, N., Remillard, J., & Thomas, M. (2013). *Technical appendix: After two years, three elementary math curricula outperform a fourth* (NCEE 2013-4019). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED544187>

Agodini, R., Harris, B., Thomas, M., Murphy, R., & Gallagher, L. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders* (NCEE 2011-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://eric.ed.gov/?&id=ED512551>

Table A1. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
Mathematics achievement	58 schools/2,045 students	+3	No

Setting The study was conducted in 111 schools in 12 districts in 10 states (Connecticut, Florida, Kentucky, Minnesota, Mississippi, Missouri, Nevada, New York, South Carolina, and Texas). Of the 12 districts, three were in urban areas, five were in suburban areas, and four were in rural areas. The study data were collected during the 2006–07, 2007–08, and 2008–09 school years. Of the 111 schools, 58 participated in the study for 2 years. These 58 schools were located in seven districts in up to seven states (the authors did not specify which states).

Study sample

The study authors randomly assigned 111 schools within 12 school districts to one of four math curricula (*Saxon Math*, *Investigations*, *Math Expressions*, or *SFAW*). Random assignment was conducted within district and stratified on characteristics such as school size, free or reduced-price meal eligibility, math proficiency, and race/ethnicity. Random assignment was conducted before the school year began in the first year of the study, and the student sample was defined immediately prior to the pretest. Thus, the study may have included students in the analytic sample that enrolled in schools after random assignment. Within study schools, all students at the target grade levels (first and/or second grades) used their school's assigned curriculum. Approximately 30 students per grade level were randomly sampled for testing by the study team.

The primary findings that contribute to the rating of effectiveness are based on students from 58 of the 111 schools (within seven of the 12 districts) who participated in the study and used their assigned curriculum for 2 consecutive years.¹¹ This analytic sample was comprised of 2,045 students in 222 classrooms who experienced their assigned curriculum in first and second grades. Students were pretested at the beginning of first grade and posttested at the end of second grade. Of the 58 study schools, 12 used *Saxon Math*, 14 used *Investigations*, 14 used *Math Expressions*, and 18 used *SFAW/enVisionMATH*. In the analytic sample of students, 49% were female, 40% were non-Hispanic Black, 32% were other non-Hispanic, and 28% were Hispanic. Students with limited English proficiency or classified as English language learners were 17% of the sample, and 9% of students had individualized education plans (IEPs) or were receiving special education services.

This review also includes supplemental findings from two study reports. The first set of supplemental findings is based on the 2009 report that presented findings for 1,309 first-grade students in 39 schools that participated in the study during the 2006–07 school year. The second set of findings is from the 2010 report that presented finding for 8,060 first- or second-grade students in 110 schools that used the study curricula for 1 year and participated in the study in either the 2006–07 or 2007–08 school years. In the 2009 and 2010 reports, the findings are 1-year effects where students were pre- and posttested in first or second grade, and findings are reported separately by grade. The supplemental findings are presented in Appendix D and do not factor into the intervention's rating of effectiveness.

Intervention group

Students in the intervention group used *Saxon Math* as their core curriculum in the first and second grades in the 2006–07, 2007–08, or 2008–09 school years. The study did not specify which edition of *Saxon Math* was used but indicated that the 2005 and 2008 copyright years were used. All teachers in the intervention group reported using *Saxon Math* as their core math curriculum in first and second grades and provided, on average, 6 hours of math instruction per week in each grade. Additionally, most teachers (87% in first grade and 71% in second grade) reported completing at least 80% of the lessons from *Saxon Math*.

Comparison group

Students in the comparison group used *Investigations*, *Math Expressions*, or *SFAW/enVision-MATH* as their core math curriculum.

Investigations in Number, Data, and Space is a K–5, student-centered curriculum that emphasizes reasoning and communicating about math concepts, using multiple approaches to problem solving, through in-depth investigations of problems. Study schools used the first or second editions of the curriculum. All teachers in the first grade and 96% in the second grade reported using *Investigations* as their core math curriculum and provided, on average, 5 hours of math instruction per week in each grade. Additionally, most teachers (86% in first grade and 56% in second grade) reported completing at least 80% of the lessons from *Investigations*.

Math Expressions is a K–5 curriculum that uses a combination of teacher-directed and student-centered approaches within consistent daily routines. The study schools used the 2005 or 2008 copyright years of the curriculum. Approximately 98% of teachers in the first grade and all teachers in the second grade reported using *Math Expressions* as their core math curriculum and provided, on average, 5 hours of math instruction per week in each grade. Additionally, most teachers (82% in first grade and 79% in second grade) reported completing at least 80% of the lessons from *Math Expressions*.

SFAW is a K–6, teacher-directed curriculum that uses a consistent daily structure, involving a brief review of previous material, exploration of a new concept, explicit instruction on the new concept, and a closure activity to check student understanding. During the course of the study, the publisher revised the *SFAW* curriculum and renamed it *enVisionMATH*. This change affected four of the seven study districts. Results for this comparison group should be interpreted as a mix of students just receiving *SFAW* curriculum in grades 1–2, and some that received *SFAW* in grade 1 and *enVisionMATH* in grade 2. Outcomes that include grade 2 in the 2008–09 school year are referred to as *SFAW/enVisionMATH*. All teachers in the first grade and 99% in the second grade reported using *SFAW* or *enVisionMATH* as their core math curriculum and provided, on average, 5 hours of math instruction per week in each grade. Additionally, most teachers (87% in first grade and 71% in second grade) reported completing at least 80% of the lessons from *SFAW* or *enVisionMATH*.

Outcomes and measurement

The outcome measure used by the study was the mathematics assessment developed for the Early Child Longitudinal Study–Kindergarten (ECLS-K) class of 1998–99. The ECLS-K was administered in the fall of first grade (within 4 weeks of the first day of classes) and in the spring of second grade (from 1–6 weeks before the end of the school year). Student tests were sent to the test publisher (Educational Testing Service) for scoring. For a more detailed description of the outcome measure, see Appendix B.

Support for implementation

Training was provided by Houghton Mifflin Harcourt trainers to *Saxon Math* teachers. *Saxon* teachers were provided 1 day of initial training in the summer before the school year began and one follow-up training session in the fall, tailored to meet each district's needs. In the follow-up training, some teachers watched demonstration lessons or participated in a math workshop. In other cases, trainers observed teachers conduct lessons and provided feedback afterwards. Approximately 90% and 81% of first- and second-grade *Saxon* teachers, respectively, attended an initial or refresher curriculum training; 76% and 17% of first- and second-grade teachers, respectively, attended a follow-up training.

The publishers of the comparison curricula provided between 1–2 days of initial training and varying levels of follow-up support. *Investigations* and *SFAW/enVisionMATH* trainers offered afterschool sessions, every 4–6 weeks for about 3–4 hours each. *Math Expressions* trainers conducted classroom observations and provided feedback once or twice a year. For *Investigations*, all first-grade teachers and 83% of second-grade teachers attended an initial or refresher training; 88% and 69% of first- and second-grade teachers, respectively, attended a follow-up training. For *Math Expressions*, 92% and 80% of first- and second-grade teachers, respectively, attended an initial or refresher training; 88% and 76% attended a follow-up training. For *SFAW* and *enVisionMATH*, 85% and 91% of first- and second-grade teachers, respectively, attended an initial or refresher training; 94% and 79% attended a follow-up training.

In addition to formal trainings, teachers received ongoing support from the publishers of each curriculum in person, by phone, and through published materials, as well as instructional support provided by coaches and math specialists in their school or district. Take-up rates for these types of supports are not provided for the analytic sample.

Appendix A.2: Research details for Crawford and Raia (1986)

Crawford, J., & Raia, F. (1986). *Analyses of eighth grade math texts and achievement*. Oklahoma City, OK: Oklahoma City Public Schools Planning, Research, and Evaluation Department.

Table A2. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
Mathematics achievement	4 schools/78 students	+16	No

Setting The study was conducted in four middle schools in eighth-grade classrooms in Oklahoma City Public Schools (OCPS). No other information was provided about the study setting. The study data were collected during the 1984–85 school year.

Study sample The study included 78 eighth-grade students (39 intervention and 39 comparison) taught by four teachers in four middle schools.¹² Each teacher taught at least one intervention class and one comparison class. To create similar intervention and comparison groups of students based on math ability, the researchers matched each intervention group student to a comparison group student with the same teacher based on their total math scores at baseline on the California Achievement Test (CAT). When more than one student from the comparison group matched a student in the intervention group, the comparison student match was selected at random. When no student from the comparison group matched a student in the intervention group, the student in the intervention group was excluded from the sample. The study authors do not provide demographic information on the sample.

Intervention group Students in the intervention group used *Saxon Algebra 1/2*, a pre-algebra math course, as their core math curriculum during the 1984–85 school year. The authors did not specify which edition of *Saxon Algebra 1/2* was used but indicate that the 1983 copyright year was used. Further information about the level of implementation in study schools was not provided.

Comparison group Students in the comparison group were taught using the district’s usual math curriculum, *Scott-Foresman Mathematics* (1980 copyright year). The authors do not provide details about how the comparison curriculum was implemented in study schools.

Outcomes and measurement

The primary outcome measure was the CAT total math score. The seventh-grade score (from spring 1984) was used as a pretest, and the spring 1985 eighth-grade score was used as the posttest. The scores are calculated as normal curve equivalent (NCE) scores, where scores range from 0 to 100, reflecting each student's percentile as compared to all others taking the exam. For a more detailed description of this outcome measure, see Appendix B.

The study also presents analyses based on two subtests of CAT: Math Computation score and Math Concepts score. The WWC includes these as supplemental findings in Appendix D that do not factor into the intervention's rating of effectiveness. The authors do not describe the content of these two subtests.

The study also examined three outcomes that do not meet WWC standards. These outcomes were author-created subscales of the CAT that included either items from the CAT that reflected math content in both the *Saxon* and *Scott-Foresman* texts, or that only reflected content in the *Scott-Foresman* text and not *Saxon*. These measures do not meet WWC standards because they are author-created measures without information about their reliability.

Support for implementation

The study does not provide information on the support for implementation.

Appendix A.3: Research details for Good et al. (2006)

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

Table A3. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
Mathematics achievement	57 schools/745 students	+4	No

Setting The study was conducted in 57 schools across 16 states (Alabama, Arizona, California, Georgia, Indiana, Nebraska, Nevada, New York, North Carolina, Oklahoma, Oregon, Tennessee, Texas, Utah, Virginia, and Washington) in kindergarten through third-grade classrooms in the 2005–06 school year. No further information was provided about the study setting.

Study sample Forty schools were randomly selected from a list of schools in the United States implementing *Saxon Math* and invited to participate in the study; 33 schools agreed to participate. In addition, 24 comparison schools agreed to participate in the study. The comparison schools were selected based on their similarity to the intervention schools, including school size, grade-level configuration, students eligible for free and reduced-price meals, racial/ethnic make-up, whether they were charter schools, Title I status, geographic location, and setting (for example, urban or rural). Within each study school, one classroom in each grade from K–3 participated in the study. This review focuses on the analytic sample of students who took the Math Problem Solving subtest; this is the only sample that demonstrates baseline equivalence. This analytic sample includes a total of 745 students in grades 2 and 3 comprised of 411 intervention students in 33 schools and 334 comparison students in 24 schools. The study authors do not provide demographic information on the analytic sample, but they do provide information on all students in their study (in grades K–3). In the full sample of students, about 65% were Caucasian, about 10% were English language learners, about 5% were in special education, about 50% were male, and about 45% were eligible for free or reduced-price meals.

Intervention group Students in the intervention group used *Saxon Math* as their core math curriculum in grades K–3 during the 2005–06 school year. The authors did not specify the edition of *Saxon Math* used. The study assessed implementation fidelity and found that, in general, the *Saxon* curriculum was implemented as intended, with 70% of teachers routinely using *Saxon Math*. In the analytic sample examined in this review, most teachers implemented the majority of the lesson components as intended in grades 2 and 3. On average, teachers in second and third grade expected to complete over 95% of *Saxon Math* lessons by the end of the school year (actual curriculum completion was not assessed). Teachers supplemented *Saxon Math* with additional materials to reinforce concepts, match state standards, or provide learning extensions.

Comparison group Students in the comparison group used a variety of math curricula including *Harcourt Brace*, *Houghton Mifflin*, *Silver Burdett Ginn*, *McGraw-Hill*, and *Scott-Foresman*. Specific details about how these curricula were implemented are not provided by the authors. As in the *Saxon* group, comparison group teachers supplemented their core curriculum with additional materials.

Outcomes and measurement

The study included one eligible outcome that met standards: scores on the Stanford Achievement Test, Ninth Edition (SAT-9) Math Problem Solving subtest. This test has grade-level versions that were administered to grades 2 and 3 (abbreviated Primary 2 and abbreviated Primary 3, respectively). Scores are vertically equated allowing for aggregation across grade-level versions of the test. The pretest was administered in fall of 2005 and the posttest in spring of 2006 (at the start and end of the 2005–06 school year). For a more detailed description of this outcome measure, see Appendix B.

Other outcome measures were collected but were ineligible for review or do not meet WWC standards. The SAT-9 test, of which the SAT-9 Math Problem Solving subtest is part, is not used by an analytic sample that meets WWC standards. In addition, students in the intervention group completed summative assessments as part of the *Saxon Math* curriculum. These assessments were ineligible for review since they were not used by students in the comparison group.

Support for implementation

The study does not provide information on the support for implementation. However, the authors note that intervention schools were using *Saxon Math* prior to the study.

Appendix A.4: Research details for Resendez et al. (2005), Sample 1

Resendez, M., Fahmy, A., & Manley, M. A. (2005). *The relationship between using Saxon middle school math and student performance on Texas statewide assessments.* [Sample 1] Jackson, WY: PRES Associates, Inc.

Table A4. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
Mathematics achievement	25 schools/3,054 students	+7	No

Setting The study took place in 25 Texas schools located in rural, suburban, and urban districts. Students in Cohort A (the analytic sample in this review) were in the sixth, seventh, and eighth grades in the 1998–99 through 2000–01 school years.

Study sample Data were collected from 15 intervention schools in Texas districts that used *Saxon Math* in the sixth, seventh, and eighth grades between 1993 and 2004. The Texas Education Agency identified 40 potential comparison schools that were similar to the intervention schools based on demographic characteristics including race, ethnicity, poverty, English language proficiency, and percentage of mobile students. Fifteen of the 40 potential schools were randomly selected for the comparison group. Within this group of 30 schools, the author selected three distinct samples of students and examined outcomes for multiple cohorts in each sample.¹³ This review focuses on Sample 1, which included Cohorts A, B, and C. Cohorts B and C were ineligible for review because they fall within the Secondary Mathematics topic area; therefore, this review focuses on the analytic sample in Cohort A.

Cohort A included data for students in 25 of the 30 schools, including a total of 3,054 students. The intervention group contained 1,472 students in 12 schools, and the comparison group contained 1,582 students in 13 schools. The study did not report the characteristics of the analytic sample of students in this review, but they did provide information for all students in the study: about 45% were Caucasian, about 40% were Hispanic, about 10% were African American, about 5% were limited English proficient, about 15% were special education status, about 50% were female, and about 45% were economically disadvantaged.

This intervention report considers the outcome in the seventh grade, after the intervention was implemented for 2 consecutive years, as the primary finding for the evidence rating of effectiveness because it is the highest grade in the study that met standards. The outcome in sixth grade is considered a supplemental finding that does not factor into the intervention’s rating of effectiveness. Because some students in the grade 8 analytic sample used *Saxon Algebra I*, the outcome measure using this sample is ineligible for review under the Primary Mathematics topic area; therefore, only outcomes in grades 6 and 7 are eligible for this review.

Intervention group

Students in the intervention group used *Saxon Math* as their core math curriculum in grades 6 and 7 during the 1998–99 and 1999–2000 school years. In the sixth grade, at least 80% of students used *Saxon Math 7/6* as their core math curriculum; in the seventh grade, at least 80% used *Saxon Math 8/7*. The remaining students used the *Saxon* curriculum at the next grade level. The study did not specify which editions of *Saxon Math* were used. Further information about the level of implementation in study schools was not provided.

Comparison group

Students in the comparison schools used core basal math curricula, which typically consist of a chapter-based approach to math instruction. Specific details about how these curricula were implemented in comparison schools are not provided by the authors.

Outcomes and measurement

The outcome measure was the Texas Assessment of Academic Skills (TAAS) Texas Learning Index for math. This was measured in the spring of 1998 (when students were in the fifth grade) and again in the spring of 1999 and spring of 2000 (for grades 6 and 7, respectively). The sixth-grade outcome is considered a supplemental finding that does not factor into the intervention’s rating of effectiveness. For a more detailed description of this outcome measure, see Appendix B.

The study also examined the number of TAAS math objectives mastered and the percentage of students meeting the TAAS math standards. However, the analytic samples examined for these outcomes do not meet WWC standards because the authors did not demonstrate their equivalence at baseline. In addition, the study presented eighth grade outcomes, which are based on a sample that is ineligible for review under the Primary Mathematics topic area because some students may have used *Saxon Algebra I*.

Support for implementation

The study does not provide information on the support for implementation. However, intervention schools were already using the *Saxon Math* curriculum prior to the study.

Appendix A.5: Research details for Resendez et al. (2005), Sample 3

Resendez, M., Fahmy, A., & Manley, M. A. (2005). *The relationship between using Saxon middle school math and student performance on Texas statewide assessments.* [Sample 3] Jackson, WY: PRES Associates, Inc.

Table A5. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index (percentile points)	Statistically significant
Mathematics achievement	20 schools/2,933 students	+10	No

Setting The study took place in 20 Texas schools located in rural, suburban, and urban districts. Students in Cohort F (the study sample in this review) were in the sixth grade in the 2003–04 school year.

Study sample Data were collected from 15 intervention schools in Texas districts that used *Saxon Math* in the sixth, seventh, and eighth grades between 1993 and 2004. The Texas Education Agency identified 40 potential comparison schools that were similar to the intervention schools based on demographic characteristics including race, ethnicity, poverty, English language proficiency, and percentage of mobile students. Fifteen of the 40 potential schools were randomly selected for the comparison group. Within this group of 30 schools, the author selected three samples of students with multiple cohorts in each sample.¹⁴ This review focuses on Sample 3, which included Cohorts F, G, and H. Cohorts G and H do not meet WWC group design standards; therefore, this review focuses on the analytic sample in Cohort F.

Cohort F included data in 20 of the 30 schools, including a total of 2,933 students. The intervention group contained 1,526 students in 10 schools, and the comparison group contained 1,407 students in 10 schools. The study did not report the characteristics of the analytic sample of students in this review, but they did provide information for all students in the study: about 45% were Hispanic, about 40% were Caucasian, about 13% were African American, about 5% were limited English proficient, about 15% were special education status, 49% were female, and about 50% were economically disadvantaged.

Intervention group Students in the intervention group used *Saxon Math* as their core math curriculum in grade 6 during the 2003–04 school year. At least 80% of students used *Saxon 7/6*, and the remainder used *Saxon 8/7*. The study did not specify which editions of *Saxon Math* were used. Further information about the level of implementation in study schools was not provided.

Comparison group The comparison students used core basal math curricula, which typically consist of a chapter-based approach to math instruction. Specific details about how these curricula were implemented in comparison schools are not provided by the authors.

Outcomes and measurement

The outcome measure was the Texas Assessment of Knowledge and Skills (TAKS) math score. This was measured in the spring of 2003 (when students were in the fifth grade) and again in the spring of 2004 (in the sixth grade). For a more detailed description of this outcome measure, see Appendix B.

The study also examined the number of TAKS math objectives mastered and the percentage of students meeting the TAKS math standards. However, the analytic samples examined for these outcomes do not meet WWC standards because the authors did not demonstrate their equivalence at baseline.

Support for implementation

The study does not provide information on the support for implementation. However, intervention schools were already using the *Saxon Math* curriculum prior to the study.

Appendix B: Outcome measures for the mathematics achievement domain

Mathematics achievement	
<i>California Achievement Test (CAT) Total Math Score</i>	The CAT is a nationally normed standardized test published by Seton Testing Services. The authors calculated the math scores as Normal Curve Equivalent scores, which range from 1 to 100 and reflect the student's score as a percentile of all students taking the exam. The CAT math test includes subtest scores for Math Computation and Math Concepts (as cited in Crawford & Raia, 1986).
<i>Early Child Longitudinal Study–Kindergarten (ECLS-K) Math Assessment</i>	The ECLS-K is a nationally normed, individually administered, adaptive math assessment developed by Educational Testing Service (ETS). The main finding reported in Appendix C is the second-grade test. The first-grade outcome, reported as a supplemental finding in Appendix D, is the scaled score of the math assessment developed by ETS. Originally, the ECLS-K was not administered or developed for use in second grade. Therefore, ETS worked with the study authors to develop an assessment for the second grade based on items existing in the ECLS-K math assessments (including the K–1, grade 3, and grade 5 tests). Cronbach's alphas for the first-grade sample on the fall (pretest) and spring (posttest) tests were .91 and .93, respectively. For the second grade sample, they were .88 and .91, respectively (as cited in Agodini et al., 2013).
<i>Stanford Achievement Test, Ninth Edition (SAT-9) Math Problem Solving subtest</i>	The SAT-9 is a nationally normed test of math achievement. The Math Problem Solving subtest was administered in grades 2 and 3 using versions abbreviated Primary 2 and abbreviated Primary 3, respectively. Scores are vertically equated allowing for aggregation across grade-level versions of the test (as cited in Good et al., 2006).
<i>Texas Assessment of Academic Skills (TAAS) Texas Learning Index</i>	The TAAS is a criterion-referenced state test that measures problem-solving and critical-thinking skills. The Texas Learning Index is an outcome metric, based on student performance on the TAAS, allowing for comparisons between administrations and between grades. The Index ranges from 0 to approximately 90, with a score of 70 representing a passing standard across grades. The reliability estimates based on internal consistency for the TAAS range from .92 to .93 for grades 6 to 8. The TAAS was used in Texas from 1990–2002 (as cited in Resendez et al., 2005, Sample 1).
<i>Texas Assessment of Knowledge and Skills (TAKS) math scale score</i>	The TAKS is a Texas statewide assessment administered to students at the end of each school year since spring of 2003. The math 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 with scores in the range of 1000–3200. The reliability estimates based on internal consistency for the TAKS range from .89 to .90 for grades 6 to 8 (as cited in Resendez et al., 2005, Sample 3).

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

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			p-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Agodini et al. (2013)^a								
<i>ECLS-K Math Assessment</i>	Grade 2 (vs. <i>Investigations</i>)	26 schools/ 882 students	71.72 (16.75)	67.31 (18.47)	4.41	0.25	+10	nr
<i>ECLS-K Math Assessment</i>	Grade 2 (vs. <i>Math Expressions</i>)	26 schools/ 931 students	67.56 (16.75)	67.99 (18.92)	-0.43	-0.02	-1	nr
<i>ECLS-K Math Assessment</i>	Grade 2 (vs. <i>SFAW/enVisionMATH</i>)	30 schools/ 1,136 students	68.90 (16.75)	68.87 (17.42)	0.03	0	0	nr
Domain average for mathematics achievement (Agodini et al., 2013)						0.07	+3	Not statistically significant
Crawford & Raia (1986)^b								
<i>CAT Total Math Score</i>	Grade 8	4 teachers/ 78 students	55.56 (11.86)	50.72 (11.75)	4.84	0.41	+16	< .01
Domain average for mathematics achievement (Crawford & Raia, 1986)						0.41	+16	Not statistically significant
Good et al. (2006)^c								
<i>SAT-9 Math Problem Solving subtest</i>	Grades 2 and 3	57 schools/ 745 students	632.89 (47.89)	627.63 (45.50)	5.26	0.11	+4	nr
Domain average for mathematics achievement (Good et al., 2006)						0.11	+4	Not statistically significant
Resendez et al. (2005), Sample 1^d								
<i>TAAS Texas Learning Index</i>	Grade 7	25 schools/ 3,054 students	83.78 (8.19)	82.27 (9.47)	1.51	0.17	+7	nr
Domain average for mathematics achievement (Resendez et al., 2005, Sample 1)						0.17	+7	Not statistically significant
Resendez et al. (2005), Sample 3^e								
<i>TAKS math scale score</i>	Grade 6	20 schools/ 2,933 students	2,229.02 (225.89)	2,174.49 (205.10)	54.53	0.25	+10	< .01
Domain average for mathematics achievement (Resendez et al., 2005, Sample 3)						0.25	+10	Not statistically significant
Domain average for mathematics achievement across all studies						0.20	+8	na

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. The statistical significance of each study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. nr = not reported. na = not applicable. CAT = California Achievement Test. ECLS-K = Early Childhood Longitudinal Study-Kindergarten. SAT-9 = Stanford Achievement Test, Ninth Edition. TAAS = Texas Assessment of Academic Skills. TAKS = Texas Assessment of Knowledge and Skills.

^a Agodini et al. (2013) includes four groups – the *Saxon Math* intervention group and three comparison curricula groups. Each pairwise comparison of *Saxon* with each comparison group is of interest to this review. Outcomes are from grade 2, after 2 years of the intervention. The authors reported p-values for some results, but not for the analyses that met standards. The WWC applied a correction for multiple comparisons and calculated a p-value of < .01 for grade 2 vs. *Investigations*, .71 for grade 2 vs. *Math Expressions*, and .98 for

grade 2 vs. *SFAW/enVisionMATH*. The comparison group mean is the unadjusted posttest comparison group mean. The intervention group means are obtained from WWC calculations and are the unadjusted comparison group means plus the coefficient from the hierarchical linear model (HLM) with the comparison curriculum as the reference category. These means were obtained from the HLM model controlling for pretest scores and randomization block only. The study authors also reported results that used multiple imputation for missing data (that is, replacing missing data with substituted values); however, the multiple imputation procedure was not carried out separately for the intervention and comparison groups, as required by the WWC Procedures and Standards Handbook (version 3.0, p. 18). The results presented in this report are based on analyses that did not include imputed data. The WWC computed a p -value of .20 for an effect size of 0.07 after pooling across the three comparison groups to form a single comparison group. Based on this, the study is characterized as having an indeterminate effect because the WWC-calculated pooled effect size is neither statistically significant nor substantively important (0.25 standard deviations or larger). For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

^b For Crawford and Raia (1986), the p -values presented here were reported in the original study. A correction for clustering was needed, and the WWC determined that the p -value could be no smaller than .18; therefore, the WWC does not find the result to be statistically significant. The WWC was unable to perform an exact adjustment for clustering because the study did not report the number of classrooms. However, even when using the most generous assumption that the study included 17 classrooms (information in the study indicates that there were at most 17 classrooms), the WWC-computed p -value would be .18, and therefore not statistically significant. The means reported in the table are ANCOVA-adjusted means, controlling for the pretest. This study is characterized as having a substantively important positive effect because the effect is positive and not statistically significant but is substantively important (0.25 standard deviations or larger). For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

^c For Good et al. (2006), the authors did not report a p -value for the SAT-9 Math Problem Solving subtest. The WWC calculated a p -value of .27 for this contrast using a cluster correction based on an intraclass correlation of 0.06, which was reported by the authors; after the correction was applied, the WWC does not find the result to be statistically significant. The study authors provided the WWC with unadjusted means, standard deviations, and sample sizes for the *Saxon Math* and comparison groups in response to an author query. The WWC calculated the program group mean using a difference-in-differences approach by adding the impact of the program (i.e., difference in mean gains between the intervention and comparison groups) to the unadjusted comparison group posttest means. The mean difference in the table may not exactly equal the difference between the reported intervention and comparison group means because the WWC's calculation accounts for changes in the standard deviation of the baseline and outcome measures. Please see the WWC Procedures and Standards Handbook (version 3.0) for more information. This study is characterized as having an indeterminate effect that is neither statistically significant nor substantively important (0.25 standard deviations or larger). For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

^d For Resendez et al. (2005), Sample 1, the authors did not report a p -value for grade 7. The WWC calculated a p -value of .35 for this contrast using a cluster correction; therefore, the WWC does not find the result to be statistically significant. The means for the *Saxon Math* group and comparison group are repeated measures ANCOVA-adjusted means, controlling for the pretest. The standard deviations are the unadjusted standard deviations provided to the WWC by the study authors in response to an author query. The findings in this table differ from the prior February 2013 intervention report, which presented grade 8 data instead of grade 7. Because a portion of the analytic sample in grade 8 included students in Algebra I, this grade sample is ineligible for review under the Primary Mathematics review protocol. This study is characterized as having an indeterminate effect that is neither statistically significant nor substantively important (0.25 standard deviations or larger). For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

^e For Resendez et al. (2005), Sample 3, the p -values presented here were reported in the original study. A correction for clustering was needed and resulted in a WWC-computed p -value of .22; therefore, the WWC does not find the result to be statistically significant. The means for the *Saxon Math* group and comparison group are ANCOVA-adjusted means, controlling for the pretest. The study authors provided the WWC with unadjusted standard deviations for the *Saxon Math* and comparison groups in response to an author query. This study is characterized as having a substantively important positive effect because the effect is positive and not statistically significant but is substantively important (0.25 standard deviations or larger). For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

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

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			p-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Agodini et al. (2013)^a								
<i>ECLS-K Math Assessment</i>	Grade 1, first cohort (vs. <i>Investigations</i> , 1 year effect)	19 schools/ 636 students	47.56 (7.62)	44.87 (8.64)	2.69	0.33	+13	nr
<i>ECLS-K Math Assessment</i>	Grade 1, first cohort (vs. <i>Math Expressions</i> , 1 year effect)	18 schools/ 618 students	45.40 (7.62)	45.45 (8.97)	-0.05	-0.01	0	nr
<i>ECLS-K Math Assessment</i>	Grade 1, first cohort (vs. <i>SFAW</i> , 1 year effect)	20 schools/ 663 students	46.17 (7.62)	44.28 (8.27)	1.89	0.24	+9	nr
<i>ECLS-K Math Assessment</i>	Grade 1, both cohorts (vs. <i>Investigations</i> , 1 year effect)	54 schools/ 2,235 students	45.13 (7.32)	44.51 (8.04)	0.62	0.08	+3	nr
<i>ECLS-K Math Assessment</i>	Grade 1, both cohorts (vs. <i>Math Expressions</i> , 1 year effect)	52 schools/ 2,320 students	44.52 (7.32)	44.74 (8.52)	-0.22	-0.03	-1	nr
<i>ECLS-K Math Assessment</i>	Grade 1, both cohorts (vs. <i>SFAW</i> , 1 year effect)	55 schools/ 2,377 students	45.08 (7.32)	44.43 (8.15)	0.65	0.08	+3	nr
<i>ECLS-K Math Assessment</i>	Grade 2, both cohorts (vs. <i>Investigations</i> , 1 year effect)	36 schools/ 1,711 students	71.88 (16.16)	69.85 (15.75)	2.03	0.13	+5	nr
<i>ECLS-K Math Assessment</i>	Grade 2, both cohorts (vs. <i>Math Expressions</i> , 1 year effect)	35 schools/ 1,721 students	73.09 (16.16)	71.38 (16.70)	1.71	0.10	+4	nr
<i>ECLS-K Math Assessment</i>	Grade 2, both cohorts (vs. <i>SFAW</i> , 1 year effect)	36 schools/ 1,706 students	72.98 (16.16)	70.31 (15.74)	2.67	0.17	+7	nr
Crawford & Raia (1986)^b								
<i>CAT Math Computation subtest</i>	Grade 8	4 teachers/ 78 students	57.66 (13.35)	51.44 (14.14)	6.22	0.45	+17	.01
<i>CAT Math Concepts subtest</i>	Grade 8	4 teachers/ 78 students	53.18 (12.44)	50.00 (12.40)	3.18	0.25	+10	.10
Resendez et al. (2005) Sample 1^c								
<i>TAAS Texas Learning Index</i>	Grade 6	25 schools/ 3,054 students	83.66 (7.72)	82.50 (9.42)	1.16	0.13	+5	nr

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported. CAT = California Achievement Test. ECLS-K = Early Childhood Longitudinal Study–Kindergarten. TAAS = Texas Assessment of Academic Skills.

^a Agodini et al. (2013), includes four groups – the *Saxon Math* intervention group and three comparison curricula groups. Each pairwise comparison of *Saxon* with each comparison group in grade 1 and grade 2 is of interest to this review. Outcomes reported are 1 year impacts from the fall to the spring in each grade. Agodini et al. (2009) presented results for grade 1 in the first cohort of schools. The authors reported p -values for some results, but not for the analyses that met standards. The WWC applied a correction for multiple comparisons and computed a p -value of $< .01$ for grade 1, first cohort vs. *Investigations*; .94 for grade 1, first cohort vs. *Math Expressions*; and $< .01$ for grade 1, first cohort vs. *SFAW*. Agodini et al. (2010) presented results for grade 1 and grade 2 across both cohorts. The authors did not report p -values for the results that meet standards and these are not provided in this table. The WWC applied a correction for multiple comparisons and computed a p -value of .06 for grade 1, both cohorts vs. *Investigations*; .51 for grade 1, both cohorts vs. *Math Expressions*; .04 for grade 1, both cohorts vs. *SFAW*; $< .01$ for grade 2, both cohorts vs. *Investigations*; .03 for grade 2, both cohorts vs. *Math Expressions*; and $< .01$ for grade 2, both cohorts vs. *SFAW*. The comparison group mean is the unadjusted posttest comparison group mean. The intervention group means are obtained from WWC calculations and are the unadjusted comparison group means plus the coefficient from the hierarchical linear model (HLM) with the comparison curriculum as the reference category. These means were obtained from the HLM model controlling for pretest scores and randomization block only. These results differ from those presented in two earlier WWC reports in the Elementary School Mathematics topic area in May 2013 and the Middle School Mathematics topic area in February 2013, which were based on 1-year findings (from Agodini et al. [2009, 2010]). The difference between the data used in this report and the data used in the previous reports is due to changes in the standards pertaining to imputation procedures for missing data (that is, replacing missing data with substituted values). The findings reported in the prior reports were based on a model that controlled for student-level characteristics and included imputed data. The study authors did not carry out the multiple imputation procedure separately for the intervention and comparison groups, as required by the WWC Procedures and Standards Handbook (version 3.0, p. 18). Therefore, the results presented in this report are based on analyses that did not include imputed data.

^b For Crawford and Raia (1986), the p -values presented here were reported in the original study. A correction for clustering was needed and resulted in a WWC-computed p -value of .25 for the CAT Math Computation subtest and .51 for the CAT Math Concepts subtest; therefore, the WWC does not find the result to be statistically significant.

^c For Resendez et al. (2005), Sample 1, the authors did not report a p -value for grade 6 on the TAAS. A correction for clustering was needed and resulted in a WWC-computed p -value of .47 for grade 6 on the TAAS; therefore, the WWC does not find the result to be statistically significant. The means for the *Saxon Math* group and comparison group are repeated measures ANCOVA-adjusted means, controlling for the pretest. The standard deviations are the unadjusted standard deviations provided to the WWC in response to an author query.

Endnotes

^{*} Due to the 2015 restructuring of the Mathematics topic area from three areas (Elementary, Middle, and High School) to two areas (Primary and Secondary Mathematics), this is considered a new report rather than an updated report. The information in this report includes reviews of some, but not all, of the studies in the prior Elementary and Middle School Mathematics reports, as not all studies in the prior reports are eligible for review under the Primary Mathematics review protocol. Endnote 2 explains which studies from the prior reports are treated differently in this report.

¹ The descriptive information for this program comes from a publicly available source: the publisher's website (www.hmhco.com, downloaded July 2016). The What Works Clearinghouse (WWC) requests publishers review the program description sections for accuracy from their perspective. The WWC provided the developer with the program description in July 2016; however, the WWC did not receive a response. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review.

² The WWC previously released reports on *Saxon Math* under the Elementary School Mathematics (ESM) topic area in May 2013 and the Middle School Mathematics (MSM) topic area in February 2013; the WWC prepared the reports using the WWC Procedures and Standards Handbook (version 2.1) and the Elementary and Middle School Mathematics review protocols (version 2.0). In June 2015, the WWC restructured the reviews of research on math interventions into two areas instead of three. These two review areas are Primary Mathematics (which includes interventions in which math is presented through multi-topic materials and curricula, typically used in grades K–8), and Secondary Mathematics (which includes interventions organized by math content area [e.g., algebra, geometry, and calculus], typically taught in grades 9–12). These two areas are replacing the prior ESM, MSM, and High School Mathematics areas, which were organized by student grade level. The WWC is updating and replacing intervention reports written under the prior topic areas.

The literature search for the current report reflects documents publicly available by July 2016. This updated report includes reviews of 17 studies that the previous intervention reports did not include. Of the additional studies, 16 were not within the scope of the review protocol for the Primary Mathematics topic area, and one was within the scope of the review protocol for Primary Mathematics but did not meet WWC group design standards. A complete list and disposition of all studies reviewed is available in the references.

The current report, which includes reviews of all previous studies that met WWC group design standards with or without reservations, resulted in a revised disposition for five studies.

Agodini et al. (2013) received a rating of meets WWC group design standards with reservations in this report, whereas it had previously received a rating of meets WWC group design standards without reservations in the ESM intervention report. The citation and rating has changed for two reasons. First, since the prior WWC review, the authors released a report (in 2013) with longitudinal 2-year findings, which are now the focus of the current review. The earlier 1-year findings (from Agodini et al. [2009, 2010]) are now considered supplemental findings that do not contribute to the study rating. Second, the current rating is based on version 3.0 of the WWC Procedures and Standards Handbook, which provides new guidance on rating cluster, or group-based, randomized controlled trials. Because the study is a cluster randomized controlled trial that might have analyzed outcomes for students who were not present at the time of school random assignment, the integrity of the study's random assignment was jeopardized. The study now meets WWC group design standards with reservations, which is the highest rating a cluster randomized controlled trial with joiners can receive when the authors discuss the effects of the intervention on students. In addition to the changes to the citation and rating, different findings contribute to the effectiveness rating in this intervention report. The supplementary results presented in Appendix D of this report for both cohorts were presented in the prior ESM report as findings that contribute to the effectiveness rating in Appendix C.

Good et al. (2006) received a rating of meets WWC group design standards with reservations, whereas it previously received a rating of does not meet WWC group design standards in the ESM intervention report. The prior rating, based on version 2.1 of the WWC Procedures and Standards Handbook, was based on the analytic sample (that is, the sample used for study analysis) used to examine outcomes on the overall Stanford Achievement Test, Ninth Edition (SAT-9) math score. That analytic sample was not similar prior to the study. This review, based on version 3.0 standards, allows equivalence to be demonstrated on a subtest outcome measure, even when equivalence is not demonstrated on the overall outcome measure. Because equivalence was demonstrated on the SAT-9 Math Problem Solving subtest, the study is rated as meets WWC group design standards with reservations.

Peters (1992) is rated ineligible for review, whereas previously it received a rating of meets WWC group design standards with reservations in the MSM intervention report. The change in rating is due to the restructuring of the Math topic area into Primary and Secondary Mathematics. The study is ineligible for review under the Primary Mathematics review protocol because the study examines the effect of *Saxon Algebra I*, which is eligible for review under the Secondary Mathematics topic area.

Resendez and Azin (2006) received a rating of does not meet WWC group design standards, whereas it previously received a rating of meets WWC group design standards without reservations in the MSM intervention report. The WWC changed the rating for two reasons. First, a portion of the analytic sample in grade 8 included students in Algebra I, which is ineligible for review under the

Primary Mathematics review protocol. Therefore, only the sample excluding grade 8 was eligible for this review. Second, the new guidance in version 3.0 of the WWC Procedures and Standards Handbook on cluster randomized controlled trials requires an assessment of student-level attrition because the authors discuss the effects of the intervention on students. Because student-level attrition is unknown (that is, the outcome variable is not available for all participants initially assigned to the intervention and comparison groups, but the study does not report the exact sample sizes), the study must demonstrate baseline equivalence. The study does not demonstrate baseline equivalence on student-level data; therefore, the WWC changed the study rating from the prior review that was based on version 2.1 of the WWC Procedures and Standards Handbook.

Resendez and Manley (2005) received a rating of does not meet WWC group design standards, whereas it previously received a rating of meets WWC group design standards with reservations in the ESM intervention report. The revised rating is due to a clarification in version 3.0 of the WWC Procedures and Standards Handbook on cluster design studies that requires the study to demonstrate baseline equivalence on student-level data because the authors discuss the effects of the intervention on students and not only effects on schools. The study does not demonstrate baseline equivalence on student-level data, and therefore, the WWC changed the study rating from the prior review that was based on version 2.1 of the WWC Procedures and Standards Handbook.

Reviews of the studies in this report used the standards from the WWC Procedures and Standards Handbook (version 3.0) and the Primary Mathematics review protocol (version 3.1). The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

³ Absence of conflict of interest: This intervention report includes a study conducted by staff from Mathematica Policy Research, Inc. Because Mathematica is one of the contractors that administers the WWC, staff members from a different organization reviewed the study. The lead methodologist, a WWC quality assurance reviewer, and an external peer reviewer reviewed this report.

⁴ As few as 18 and as many as 23 states formed the analytic sample across the five studies. The lower bound of 18 states assumes the greatest amount of overlap between the analytic sample of states in Agodini et al. (2013) and those in the other studies. The authors did not identify the states participating in the study for a second year.

⁵ Please see the Primary Mathematics review protocol (version 3.1) for more information about the outcome domain.

⁶ For criteria used to determine the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 32. These improvement index numbers show the average and range of individual-level improvement indices for all findings across the studies.

⁷ The authors presented three analytic samples, only one of which meets WWC group design standards. The other analytic samples do not meet WWC group design standards because baseline equivalence is required, but not demonstrated.

⁸ Good et al. (2006) presented findings for three analytic samples: (1) the SAT-9 math score for kindergarten through third-grade students, (2) the SAT-9 Math Procedures subtest for second- and third-grade students, and (3) the SAT-9 Math Problem Solving subtest for second- and third-grade students. Only the analytic sample used to present outcomes on the SAT-9 Math Problem Solving subtest meets WWC group design standards. The other two analytic samples do not meet WWC group design standards because baseline equivalence is required but not demonstrated.

⁹ Resendez et al. (2005) presented results for eight cohorts of students (A through H) who came from three different samples: Sample 1 contained Cohorts A, B, and C; Sample 2 contained Cohorts D and E; and Sample 3 contained Cohorts F, G, and H. According to the WWC Procedures and Standards Handbook (version 3.0, p. 7), quasi-experimental designs that involve independent samples are considered separate studies even if the findings appear in the same research article. As such, the WWC treated the three samples as separate studies in this review. Samples 1 and 3 contain analytic samples that meet WWC group design standards. In Sample 1, Cohort A meets WWC group design standards. Cohorts B and C were ineligible for review because the analysis examined outcomes in tenth grade, when students may have taken Algebra I or other secondary math courses at the time outcomes were measured; therefore, the analytic sample is ineligible for review in the Primary Mathematics topic area. In Sample 3, Cohort F meets WWC group design standards. Cohorts G and H are eligible for review but do not meet WWC group design standards because the analytic samples were based on a quasi-experimental design in which baseline equivalence is required, but not demonstrated. Sample 2 (Cohorts D and E) is ineligible for review because it did not use an eligible design. Sample 2 is based on eight schools that used *Saxon Math* between 1994 and 2001. The findings in this report are based on Cohort A in Sample 1 and Cohort F in Sample 3, two independent analytic samples that meet WWC group design standards.

¹⁰ The WWC identified four other additional sources related to Agodini (2013). These studies do not contribute unique information to Appendix A.1 and are not listed here.

¹¹ In two earlier publications (Agodini et al., 2009; Agodini et al., 2010), the study authors reported results after study schools had implemented the assigned curriculum for 1 year (in first and second grades). The primary findings (those used to determine the rating of effectiveness) included in this report are from a subsample of classrooms (reported in Agodini et al., 2013) followed longitudinally for

2 years that used the assigned curriculum over both years. The WWC selected the longitudinal sample as the primary outcome, rather than the 1-year result, as the longitudinal student sample received greater exposure to the intervention.

¹² The authors present three analytic samples; however, only one meets WWC group design standards. Two other samples did not meet WWC group design standards because they use quasi-experimental designs, and baseline equivalence is not demonstrated as required.

¹³ As explained in Endnote 9, the authors present findings on three distinct samples that are considered separate studies by the WWC. Within Sample 1, Cohort A was the only cohort in this sample to meet WWC group design standards. The remaining cohorts in this sample (Cohorts B and C) were ineligible for review in the Primary Mathematics topic area because the analysis examined outcomes in tenth grade, when students may have taken Algebra I or other secondary math courses at the time outcomes were measured.

¹⁴ As explained in Endnote 9, authors present findings on three distinct samples of students that are considered separate studies by the WWC. Within Sample 3, Cohort F was the only cohort in this sample to meet WWC group design standards. The remaining cohorts in this sample (Cohorts G and H) do not meet WWC group design standards because the analytic samples are based on a quasi-experimental design where baseline equivalence is required, but not demonstrated.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2017, May).
Primary Mathematics intervention report: Saxon Math. Retrieved from <https://whatworks.ed.gov>

WWC Rating Criteria

Criteria used to determine the rating of a study

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

Criteria used to determine the rating of effectiveness for an intervention

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

Criteria used to determine the extent of evidence for an intervention

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

Glossary of Terms

Attrition Attrition occurs when an outcome variable is not available for all subjects initially assigned to the intervention and comparison groups. If a randomized controlled trial (RCT) or regression discontinuity design (RDD) study has high levels of attrition, the validity of the study results can be called into question. An RCT with high attrition cannot receive the highest rating of *Meets WWC Group Design Standards without Reservations*, but can receive a rating of *Meets WWC Group Design Standards with Reservations* if it establishes baseline equivalence of the analytic sample. Similarly, the highest rating an RDD with high attrition can receive is *Meets WWC RDD Standards with Reservations*.

For single-case design research, attrition occurs when an individual fails to complete all required phases or data points in an experiment, or when the case is a group and individuals leave the group. If a single-case design does not meet minimum requirements for phases and data points within phases, the study cannot receive the highest rating of *Meets WWC Pilot Single-Case Design Standards without Reservations*.

Baseline A point in time before the intervention was implemented in group design research and in regression discontinuity design studies. When a study is required to satisfy the baseline equivalence requirement, it must be done with characteristics of the analytic sample at baseline. In a single-case design experiment, the baseline condition is a period during which participants are not receiving the intervention.

Clustering adjustment An adjustment to the statistical significance of a finding when the units of assignment and analysis differ. When random assignment is carried out at the cluster level, outcomes for individual units within the same clusters may be correlated. When the analysis is conducted at the individual level rather than the cluster level, there is a mismatch between the unit of assignment and the unit of analysis, and this correlation must be accounted for when assessing the statistical significance of an impact estimate. If the correlation is not accounted for in a mismatched analysis, the study may be too likely to report statistically significant findings. To fairly assess an intervention's effects, in cases where study authors have not corrected for the clustering, the WWC applies an adjustment for clustering when reporting statistical significance.

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

Design The method by which intervention and comparison groups are assigned (group design and regression discontinuity design) or the method by which an outcome measure is assessed repeatedly within and across different phases that are defined by the presence or absence of an intervention (single-case design). Designs eligible for WWC review are randomized controlled trials, quasi-experimental designs, regression discontinuity designs, and single-case designs.

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

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

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

Glossary of Terms (continued)

Extent of evidence An indication of how much evidence from group design studies supports the findings in an intervention report. The extent of evidence categorization for intervention reports focuses on the number and sizes of studies of the intervention in order to give an indication of how broadly findings may be applied to different settings. There are two extent of evidence categories: small and medium to large.

- **small:** includes only one study, or one school, or findings based on a total sample size of less than 350 students and 14 classrooms (assuming 25 students in a class)
- **medium to large:** includes more than one study, more than one school, and findings based on a total sample of at least 350 students or 14 classrooms

Gain scores The result of subtracting the pretest from the posttest for each individual in the sample. Some studies analyze gain scores instead of the unadjusted outcome measure as a method of accounting for the baseline measure when estimating the effect of an intervention. The WWC reviews and reports findings from analyses of gain scores, but gain scores do not satisfy the WWC's requirement for a statistical adjustment under the baseline equivalence requirement. This means that a study that must satisfy the baseline equivalence requirement and has baseline differences between 0.05 and 0.25 standard deviations *Does Not Meet WWC Group Design Standards* if the study's only adjustment for the baseline measure was in the construction of the gain score.

Group design A study design in which outcomes for a group receiving an intervention are compared to those for a group not receiving the intervention. Comparison group designs eligible for WWC review are randomized controlled trials and quasi-experimental designs.

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

Intervention An educational program, product, practice, or policy aimed at improving student outcomes.

Intervention report A summary of the findings of the highest-quality research on a given program, product, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against design standards, and summarizes the findings of those that meet WWC design standards.

Multiple comparison adjustment An adjustment to the statistical significance of results to account for multiple comparisons in a group design study. The WWC uses the Benjamini-Hochberg (BH) correction to adjust the statistical significance of results within an outcome domain when study authors perform multiple hypothesis tests without adjusting the p -value. The BH correction is used in three types of situations: studies that tested multiple outcome measures in the same outcome domain with a single comparison group; studies that tested a given outcome measure with multiple comparison groups; and studies that tested multiple outcome measures in the same outcome domain with multiple comparison groups. Because repeated tests of highly correlated constructs will lead to a greater likelihood of mistakenly concluding that the impact was different from zero, in all three situations, the WWC uses the BH correction to reduce the possibility of making this error. The WWC makes separate adjustments for primary and secondary findings.

Glossary of Terms (continued)

- Outcome domain** A group of closely-related outcomes. A domain is the organizing construct for a set of related outcomes through which studies claim effectiveness.
- Quasi-experimental design (QED)** A quasi-experimental design (QED) is a research design in which study participants are assigned to intervention and comparison groups through a process that is not random.
- Randomized controlled trial (RCT)** A randomized controlled trial (RCT) is an experiment in which eligible study participants are randomly assigned to intervention and comparison groups.
- Rating of effectiveness** For group design research, the WWC rates the effectiveness of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. For single-case design research, the WWC rates the effectiveness of an intervention in each domain based on the quality of the research design and the consistency of demonstrated effects. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 32.
- Regression discontinuity design (RDD)** A design in which groups are created using a continuous scoring rule. For example, students may be assigned to a summer school program if they score below a preset point on a standardized test, or schools may be awarded a grant based on their score on an application. A regression line or curve is estimated for the intervention group and similarly for the comparison group, and an effect occurs if there is a discontinuity in the two regression lines at the cutoff.

Please see the WWC Procedures and Standards Handbook (version 3.0) for additional details.



An **intervention report** summarizes the findings of high-quality research on a given program, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against evidence standards, and summarizes the findings of those that meet standards.

This intervention report was prepared for the WWC by Mathematica Policy Research under contract ED-IES-13-C-0010.