

Appendix

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

Characteristic	Description
Study citation	Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). <i>Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools</i> (NCEE 2009-4052). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
Participants	<p>Schools were randomly assigned to one of four different curricula, using a stratified procedure that helped allocate similar numbers and types of schools to each curriculum. All first-grade classrooms in participating schools were included in the study. When compared to the national average, participating schools had a higher percentage of minority students and students eligible for free/reduced-price meals.</p> <p>The baseline sample consisted of four districts, 40 schools, 134 teachers, and 1,525 first-grade students. The analysis sample consisted of four districts, 39 schools, 131 teachers, and 1,309 first-grade students: 11 schools with 36 teachers and 359 students used <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> (the intervention), 10 schools with 33 teachers and 332 students used <i>Investigations in Number, Data, and Space</i> (comparison 1), 9 schools with 31 teachers and 314 students used <i>Math Expressions</i> (comparison 2), and 9 schools with 31 teachers and 304 students used <i>Saxon Math</i> (comparison 3).</p> <p>One school with 3 teachers and 32 students assigned to <i>Math Expressions</i> withdrew from the study and did not permit posttesting of students. Because this represents differential attrition of more than 5 percentage points for the comparison of <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> and <i>Math Expressions</i>, this particular comparison is rated as meeting evidence standards with reservations.</p> <p>The authors compared the baseline characteristics of the students in the analysis sample on seven characteristics, including the baseline assessment score. Statistical tests conducted on those characteristics for the analysis sample indicated no differences across the four groups. Subgroup findings based on school and classroom characteristics, including baseline fall math achievement and free/reduced-price meal eligibility, are provided in Appendix A4.</p>
Setting	The study included 39 schools in four districts located in Connecticut, Minnesota, Nevada, and New York. Two districts were in urban areas, one district was in a suburban area, and the other district was in a rural area.
Intervention	Students used the 2005 <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> curriculum as their core math curriculum during the 2006/07 school year. <i>Scott Foresman–Addison Wesley Mathematics</i> is published by Pearson Scott Foresman and is a basal curriculum that combines teacher-directed instruction with a variety of differentiated materials and instructional strategies. Teachers select the materials that seem most appropriate for their students. The curriculum is based on a consistent daily lesson structure, which includes direct instruction, hands-on exploration, the use of questioning, and practice of new skills. Some 87% of teachers reported completing at least 80% of the curriculum (not significantly different from the other three curricula, p -value = 0.24).

continued

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

Characteristic	Description
Comparisons	<p>Comparison 1 students used <i>Investigations in Number, Data, and Space</i> as their core math curriculum. The curriculum is published by Pearson Scott Foresman. It uses a student-centered approach that encourages reasoning and understanding and draws on constructivist learning theory. The lessons focus on understanding, rather than on “correct answers,” and build on students’ knowledge and understanding. Students are engaged in thematic units of three to eight weeks in which they first investigate and then discuss and reason about problems and strategies. Students frequently create their own representations. Some 80% of teachers reported completing at least 80% of the curriculum (not significantly different from the other three curricula, p-value = 0.24).</p> <p>Comparison 2 students used <i>Math Expressions</i> as their core math curriculum. <i>Math Expressions</i> is published by Houghton Mifflin Harcourt and uses a blend of student-centered and teacher-directed instructional approaches. Students using the curriculum question and discuss mathematics but are explicitly taught effective procedures. There is an emphasis on using multiple specified objects, drawings, and language to represent concepts, and an emphasis on learning through the use of real-world situations. Students are expected to explain and justify their solutions. Some 89% of teachers reported completing at least 80% of the curriculum (not significantly different from the other three curricula, p-value = 0.24).</p> <p>Comparison 3 students used <i>Saxon Math</i> as their core math curriculum. <i>Saxon Math</i> is published by Houghton Mifflin Harcourt and uses a teacher-directed approach that offers a script for teachers to follow in each lesson. The curriculum blends teacher-directed instruction of new material with daily distributed practice of previously learned concepts and procedures. The teacher introduces concepts or efficient strategies for solving problems. Students observe and then receive guided practice, followed by distributed practice. Students hear the correct answers and are explicitly taught procedures and strategies. Frequent monitoring of student achievement is built into the program. Daily routines are extensive and emphasize practice of number concepts and procedures and use of representations. Some 97% of teachers reported completing at least 80% of the curriculum (not significantly different from the other three curricula, p-value = 0.24).</p>
Primary outcomes and measurement	<p>Mathematics achievement was measured using the mathematics assessment developed for the Early Childhood Longitudinal Study–Kindergarten (ECLS-K) Class of 1998–99. The assessment is individually administered, nationally normed, and adaptive. According to the authors, the assessment meets accepted standards of validity and reliability. Scale scores from an item response theory (IRT) model were used in the analysis. For a more detailed description of the outcome measure, see Appendix A2.</p>
Staff/teacher training	<p>Teachers in all four groups were provided training by the curriculum publisher trainers.</p> <p>Intervention: All teachers were provided one day of initial training in the summer before the school year began. More than 90% of teachers reported feeling adequately or very well prepared to use the intervention after the initial training. Follow-up training was offered about every four to six weeks throughout the school year. Follow-up sessions were typically three to four hours long and held after school.</p> <p>Comparison 1: Teachers assigned to <i>Investigations in Number, Data, and Space</i> were provided one day of initial training in the summer before the school year began. More than 90% of teachers reported feeling adequately or very well prepared to use the curriculum after the initial training. Follow-up training was offered about every four to six weeks throughout the school year. Follow-up sessions were typically three to four hours long and held after school.</p> <p>Comparison 2: Teachers assigned to <i>Math Expressions</i> were provided two days of initial training in the summer before the school year began. Some 54% of teachers reported feeling adequately or very well prepared to use the curriculum after the initial training. Two follow-up trainings were offered during the school year. Follow-up sessions typically consisted of classroom observations followed by short feedback sessions with teachers.</p> <p>Comparison 3: Teachers assigned to <i>Saxon Math</i> were provided one day of initial training in the summer before the school year began. More than 90% of teachers reported feeling adequately or very well prepared to use the curriculum after the initial training. One follow-up training session was offered during the school year and tailored to meet each district’s needs.</p>

Appendix A1.2 Study characteristics: Resendez & Azin, 2006

Characteristic	Description
Study citation	Resendez, M., & Azin, M. (2006). <i>2005 Scott Foresman–Addison Wesley Elementary Math randomized control trial: Final report</i> . Jackson, WY: PRES Associates, Inc.
Participants¹	Third- and 5th-grade teachers were randomly assigned to the intervention or comparison condition. The baseline sample included 39 teachers (20 treatment and 19 comparison) and 915 students (468 treatment and 447 comparison). Twenty-three teachers taught 3rd grade (13 treatment and 10 comparison), and 16 taught 5th grade (7 treatment and 9 control). No teachers left the study, and student attrition was low. Between 837 and 863 students were posttested on the TerraNova Math Computation and Math Total assessments, respectively. ² In general, participating schools had a higher percentage of Asian students and students with higher ability levels than the national average. Participating schools had a lower percentage of Hispanic and African-American students, special education students, and students eligible for free/reduced-price meals than the national average.
Setting	Four schools (two in Ohio and two in New Jersey) participated in the study. Schools were in urban and suburban settings.
Intervention	Students used the 2005 <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> curriculum during the 2005/06 school year. The curriculum is a research-based program designed to make math simpler to teach, easier to learn, and more accessible to every student. The curriculum is a comprehensive, basal program that emphasizes independent learning, embedded assessment, and immediate and systematic remediation. The teachers covered 79% (SD = 18.1%) of the curriculum.
Comparison	Comparison students used three different math curricula. Students in two schools used a chapter-based, comprehensive basal program; students in a third school used a basal math program; and students in a fourth school used a school-created math program based on a number of different math materials from various resources. The comparison curricula generally covered the same content as <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> . Teachers covered 80% (SD = 9.5%) of the curricula.
Primary outcomes and measurement	The authors administered the TerraNova Basic Multiple Assessment with Plus test (Level 13 in 3rd grade and Level 15 in 5th grade). The math test provides two overall scores: the TerraNova Math Total and the TerraNova Math Computation Total. The Math Total score is based on multiple choice and constructed response items that are predominantly word problems that measure basic, applied, and higher-order thinking skills. The TerraNova Math Computation Total is based on the Plus test booklet, which contains only multiple-choice computational problems. Scale scores were used in the analysis. For a more detailed description of these outcome measures, see Appendix A2.
Staff/teacher training	<p>Teachers received three hours of initial training prior to implementing <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> in their classes. At the initial training session, the trainer described the key components of <i>Scott Foresman–Addison Wesley Elementary Mathematics</i>, reviewed the Teacher’s Edition and available ancillary resources, offered examples of when to use certain materials, provided an overview of the math technology available, and modeled a math lesson. The training focused on the components most vital to the program and those that were required for full implementation.</p> <p>Two follow-up sessions were offered during the school year. The first session was offered four to eight weeks into the school year and lasted two hours. The session was informal and allowed teachers to discuss and ask questions about issues encountered while implementing the program. A second follow-up session was provided to one school in March (the other three schools were offered the second follow-up session but chose not to receive it). The second follow-up addressed pacing issues and further covered the technology available with the program.</p>

1. The study presented results based on student-level analysis. However, the analysis included some students who did not take both the pre- and posttests. To make results comparable with other studies in this review, an author query was conducted to obtain results based on classroom-level means. The results in this review are based on the class means.
2. The exact number of students taking both the pretest and posttest is not available.

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

Characteristic	Description
Study citation	Resendez, M., & Manley, M. A. (2005). <i>Final report: A study on the effectiveness of the 2004 Scott Foresman–Addison Wesley Elementary Math program</i> . Jackson, WY: PRES Associates, Inc.
Participants	Second- and 4th-grade teachers were randomly assigned to the intervention using <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> . The baseline sample included 35 teachers (18 treatment and 17 comparison) and 742 students (389 treatment and 353 comparison). Of the 35 study teachers, 19 taught 2nd grade (10 treatment and 9 comparison) and 16 taught 4th grade (8 treatment and 8 comparison). The analysis sample included 35 teachers (18 treatment and 17 comparison) and 533 (290 treatment and 243 comparison) to 645 (352 treatment and 293 comparison) students in the TerraNova Math Computation and Math Total analyses, respectively. ¹ For both assessments, the differential attrition exceeded 5 percentage points; therefore, this study is rated as meeting evidence standards with reservations. Some 37% of participating students were minorities. At two of the six participating schools, more than 90% of students were eligible for free/reduced-price meals; the percentage of students eligible for free/reduced-price meals at the other four schools was similar to the national average of 37%.
Setting	This study took place in six elementary schools in urban, suburban, and rural communities in Washington (one urban school), Wyoming (one rural and one suburban school), Virginia (one urban school), and Kentucky (two suburban schools).
Intervention	Students used the 2004 <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> curriculum during the 2004/05 school year. The curriculum is a comprehensive basal program that uses several research-based strategies to promote student success. The curriculum's goal is to help students both do and understand math. The study teachers were implementing the intervention curriculum for the first time and covered 70% (SD = 15.3%) of the curriculum.
Comparison	Students used five different comprehensive math curricula that used basal or investigative approaches. The comparison curricula covered the same content as <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> . Teachers covered 75% (SD = 18.2%) of the curricula.
Primary outcomes and measurement	The primary outcome measure was the TerraNova CTBS, Basic Multiple Assessment with Plus test (Level 12 for 2nd grade and Level 14 for 4th grade). As noted by the authors, the TerraNova CTBS is a reliable and standardized test consisting of multiple-choice, constructed response, and computational problems. According to the authors, it offers broad coverage of mathematics content in most textbooks and reflects the National Council of Teachers of Mathematics (NCTM) standards. The assessment provides two overall scores: the TerraNova Math Total and TerraNova Math Computation Total. Normal curve equivalent (NCE) scores were used in the analysis. For a more detailed description of these outcome measures, see Appendix A2.
Staff/teacher training	Teachers in the intervention classrooms met with a <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> professional trainer for approximately four hours prior to implementing the curriculum in their classes. In the initial training session, the trainer described the key components of the curriculum, reviewed the materials provided, and offered examples of when to use certain materials. Two follow-up sessions, approximately two hours each, were offered. The first session occurred 4 to 8 weeks after teachers began implementation. A second session occurred 10 to 18 weeks after implementation and was provided by the <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> trainer and one of the curriculum authors. This second session focused on the curriculum's philosophy, lesson modeling, and how teachers could use <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> to help students understand mathematics. The second session was provided to five of the six schools.

1. Number of students indicates the number posttested.

Appendix A2 Outcome measures for the mathematics achievement domain

Outcome measure	Description
ECLS-K Math Assessment	The ECLS-K Math Assessment was developed for the National Center for Education Statistics' Early Childhood Longitudinal Study–Kindergarten (ECLS-K) Class of 1998–99. The assessment is individually administered, nationally normed, and adaptive. The authors indicate that they selected the test because it met accepted standards of validity and reliability, because it measured achievement gains over the study's grade range, and because of the test's accuracy in measuring achievement of students from a wide range of backgrounds and ability levels. The assessment measures the following content areas: (1) Number Sense, Properties, and Operations; (2) Measurement; (3) Geometry and Spatial Sense; (4) Data Analysis, Statistics, and Probability; and (5) Patterns, Algebra, and Functions. The student tests were scored by the Educational Testing Service using a three-parameter item response theory (IRT) model. Scale scores from the IRT scoring were used in the analysis.
TerraNova CTBS Basic Multiple Assessment	The TerraNova CTBS Basic Multiple Assessment is a standardized test that provides an overall score for mathematics (the Math Total score). Level 12 was administered to 2nd-grade (34 questions), Level 13 to 3rd-grade (38 questions), Level 14 to 4th-grade (43 questions), and Level 15 to 5th-grade (43 questions) students. The test is administered during two class sessions and takes 75 to 90 minutes to complete. The majority of items are word problems measuring basic, applied, and higher-order thinking skills, and the test also contains a few computational problems, as well as multiple choice and constructed response questions. The authors state that they selected the test because of its validity, reliability, and sensitivity; because it assesses content presented in the latest textbook series available from multiple publishers; and because it reflects NCTM standards. The test is scored by CTB-McGraw Hill, which provides a normal curve equivalent (NCE) score and scale score. Scorers demonstrated inter-rater reliability on the constructed response items of 0.86 to 0.98 in Resendez and Manley (2005) and 0.81 to 0.90 in Resendez and Azin (2006).
TerraNova CTBS Basic Multiple Assessment with Plus	The TerraNova CTBS Basic Multiple Assessment with Plus test is a supplemental test that can be administered with the TerraNova CTBS Basic Multiple Assessment. It provides a separate overall score (the Math Computation score). The test contains 20 multiple-choice items measuring basic and advanced computational skills. The test takes 20 minutes to complete. It is scored by CTB-McGraw Hill, which provides a normal curve equivalent (NCE) score and scale score.

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

Outcome measure	Study sample	Sample size (teachers/ students)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ³ (Scott Foresman–Addison Wesley Elementary Mathematics – comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
			Scott Foresman–Addison Wesley Elementary Mathematics group	Comparison group				
Agodini et al., 2009⁷								
Comparison 1: <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> compared with <i>Investigations in Number, Data, and Space</i>								
ECLS-K	Grade 1	69/691	45.43 ⁸ (8.27)	44.87 ⁹ (8.64)	0.56	0.07	ns	+3
Comparison 2: <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> compared with <i>Math Expressions</i>								
ECLS-K	Grade 1	67/673	43.34 ⁸ (8.27)	45.45 ⁹ (8.97)	–2.11	–0.24	Statistically significant	–10
Comparison 3: <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> compared with <i>Saxon Math</i>								
ECLS-K	Grade 1	67/663	44.54 ⁸ (8.27)	46.47 ⁹ (7.62)	–1.93	–0.24	Statistically significant	–10
Average for mathematics achievement (Agodini et al., 2009)¹⁰						–0.14	nr	–6
Resendez & Azin, 2006⁷								
TerraNova Math Total	Grades 3 and 5	39/863 ¹¹	654.71 ¹² (42.40)	656.00 (47.81)	–1.29	–0.03 ¹³	ns	–1
TerraNova Math Computation	Grades 3 and 5	39/838 ¹¹	633.28 ¹² (52.03)	624.83 (52.58)	8.45	0.16 ¹³	ns	+6
Average for mathematics achievement (Resendez & Azin, 2006)¹⁰						0.07	ns	+3
Resendez & Manley, 2005⁷								
TerraNova Math Total	Grades 2 and 4	35/645 ¹⁴	55.59 (18.49)	54.14 (19.78)	1.45	0.08	ns	+3
TerraNova Math Computation	Grades 2 and 4	35/533 ¹⁴	53.89 (21.35)	57.49 (20.46)	–3.60	–0.17	ns	–7
Average for mathematics achievement (Resendez & Manley, 2005)¹⁰						–0.05	ns	–2
Domain average for mathematics achievement across all studies¹⁰						–0.04	na	–2

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

ECLS-K = Math assessment developed for the Early Childhood Longitudinal Study–Kindergarten Class of 1998–99

continued

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

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the mathematics achievement domain. Subgroup findings from the same studies are not included in these ratings but are reported in Appendix A4.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between -50 and +50, with positive numbers denoting favorable results for the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Agodini et al. (2009), no corrections for clustering or multiple comparisons were needed. In the cases of Resendez and Azin (2006) and Resendez and Manley (2005), corrections for multiple comparisons were needed, so the significance levels may differ from those reported in the original studies.
8. The intervention group mean is the unadjusted control mean plus the program coefficients from the hierarchical linear modeling (HLM) analysis.
9. The control group mean is the unadjusted control group mean.
10. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.
11. Number of students indicates the number posttested. The exact number of students taking both the pretest and posttest is not available.
12. The intervention group values are the comparison group means plus the difference in mean gains between the intervention and comparison groups. The outcome means are classroom/teacher-level means obtained through an author query. The reported standard deviation is the student-level unadjusted posttest standard deviation obtained from the report.
13. The effect size reported here differs from the report. The effect size was calculated by the WWC using classroom-level means and student-level standard deviations.
14. Number of students indicates the number posttested.

Appendix A4 Summary of subgroup findings for the mathematics achievement domain¹

Outcome measure	Study sample ⁴	Sample size (students) ⁵	Authors' findings from the study ²		WWC calculations			
			Mean outcome (standard deviation) ³		Mean difference ⁷ (Scott Foresman–Addison Wesley Elementary Mathematics – comparison)	Effect size ⁸	Statistical significance ⁹ (at $\alpha = 0.05$)	Improvement index ¹⁰
			Scott Foresman–Addison Wesley Elementary Mathematics group ⁶	Comparison group ⁶				
Agodini et al., 2009¹¹								
Comparison 1: <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> compared with <i>Investigations in Number, Data, and Space</i>								
ECLS-K	Lowest third	172	nr	nr	nr	0.15	ns	+6
ECLS-K	Middle third	206	nr	nr	nr	0.18	ns	+7
ECLS-K	Highest third	313	nr	nr	nr	–0.03	ns	–1
ECLS-K	Up to 40% FRP	396	nr	nr	nr	0.02	ns	+1
ECLS-K	Greater than 40% FRP	295	nr	nr	nr	0.16	ns	+6
Comparison 2: <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> compared with <i>Math Expressions</i>								
ECLS-K	Lowest third	199	nr	nr	nr	–0.21	ns	–8
ECLS-K	Middle third	252	nr	nr	nr	–0.18	ns	–7
ECLS-K	Highest third	222	nr	nr	nr	–0.25	ns	–10
ECLS-K	Up to 40% FRP	334	nr	nr	nr	–0.29	ns	–11
ECLS-K	Greater than 40% FRP	339	nr	nr	nr	–0.21	ns	–8
Comparison 3: <i>Scott Foresman–Addison Wesley Elementary Mathematics</i> compared with <i>Saxon Math</i>								
ECLS-K	Lowest third	201	nr	nr	nr	–0.56	Statistically significant	–21
ECLS-K	Middle third	195	nr	nr	nr	0.01	ns	0
ECLS-K	Highest third	267	nr	nr	nr	–0.18	ns	–7
ECLS-K	Up to 40% FRP	346	nr	nr	nr	–0.30	ns	–12

continued

Appendix A4 Summary of subgroup findings for the mathematics achievement domain¹ (continued)

Outcome measure	Study sample ⁴	Sample size (students) ⁵	Authors' findings from the study ²		WWC calculations			
			Mean outcome (standard deviation) ³		Mean difference ⁷ (Scott Foresman–Addison Wesley Elementary Mathematics – comparison)	Effect size ⁸	Statistical significance ⁹ (at $\alpha = 0.05$)	Improvement index ¹⁰
			Scott Foresman–Addison Wesley Elementary Mathematics group ⁶	Comparison group ⁶				
ECLS-K	Greater than 40% FRP	317	nr	nr	nr	–0.20	ns	–10

ns = not statistically significant

nr = not reported

ECLS-K = Math assessment developed for the Early Childhood Longitudinal Study–Kindergarten Class of 1998–99

FRP = Free/reduced-price meal eligibility

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

Appendix A5 *Scott Foresman–Addison Wesley Elementary Mathematics* rating for the mathematics achievement domain

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

For the outcome domain of mathematics achievement, the WWC rated *Scott Foresman–Addison Wesley Elementary Mathematics* as having mixed effects for elementary students. The remaining ratings (no discernible effects, potentially negative effects, and negative effects) were not considered, as *Scott Foresman–Addison Wesley Elementary Mathematics* was assigned the highest applicable rating.

Rating received

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

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, and at least one study showing a statistically significant or substantively important *negative* effect, but no more such studies than the number showing a statistically significant or substantively important *positive* effect.

Not met. No studies showed a statistically significant or substantively important positive effect.

OR

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

Met. One study showed statistically significant negative effects, and two studies showed indeterminate effects.

Other ratings considered

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

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

Not met. No studies showed a statistically significant or substantively important positive effect.

AND

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

Not met. One study showed statistically significant negative effects.

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

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

Not met. No studies showed a statistically significant or substantively important positive effect.

AND

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

Not met. One study showed statistically significant or substantively important negative effects, and two studies showed indeterminate effects.

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

Appendix A6 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence ¹
		Schools	Students	
Mathematics achievement	3	49	2,817 ²	Medium to large

1. A rating of “medium to large” requires at least two studies and two schools across studies in one domain and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is “small.” For more details on the extent of evidence categorization, see the WWC Procedures and Standards Handbook, Appendix G.
2. This number is an estimate based on students with available posttest scores across the three studies. The exact number of students in the analytical sample is not available for Resendez and Azin (2006).