WWC Intervention Report

A summary of findings from a systematic review of the evidence

ies INSTITUTE OF EDUCATION SCIENCES

Beginning Reading Updated March 2017

Success for All®

Report Contents

This intervention report presents findings from a systematic review of *Success for All®* conducted using the WWC Procedures and Standards Handbook, version 3.0, and the Beginning Reading review protocol, version 3.0.

Intervention Description[1]

Success for All (SFA®) is a whole-school reform model (that is, a model that integrates curriculum, school culture, family, and community supports) for students in prekindergarten through grade 8. *SFA®* includes a literacy program, quarterly assessments of student learning, a social-emotional development program, computer-assisted tutoring tools, family support teams for students' parents, a facilitator who works with school personnel, and extensive training for all intervention teachers. The literacy program emphasizes phonics for beginning readers and comprehension for all students. Teachers provide reading instruction to students grouped by reading ability for 90 minutes a day, 5 days a week. In addition, certified teachers or paraprofessionals provide daily tutoring to students who have difficulty reading at the same level as their classmates.

This review focuses on the literacy component of *SFA®*, which is implemented as part of the *SFA®* whole-school reform program. Ratings presented in this intervention report do not take into account the variations in implementation of the *SFA®* whole-school reform model. This review of the program for Beginning Reading focuses on students in grades K–4.

Research[2]

The What Works Clearinghouse (WWC) identified nine studies of *SFA®* that both fall within the scope of the Beginning Reading topic area and meet WWC group design standards. Two studies meet WWC group design standards without reservations, and seven studies meet WWC group design standards with reservations. Together, these studies included 10,908 beginning readers in grades K–4 in 155 schools in the United States and the United Kingdom.

According to the WWC review, the extent of evidence for *SFA®* on the reading achievement test scores of beginning readers was medium to large for all four outcome domains—alphabetics, reading fluency, comprehension, and general reading achievement.[3] (See the Effectiveness Summary on p. 7 for more details of effectiveness by domain.)

Effectiveness

SFA® had positive effects on alphabetics, potentially positive effects on reading fluency, and mixed effects on comprehension and general reading achievement for students in grades K–4.

Table 1. Summary of findings[4]

Outcome domain	Rating of effectiveness	Improvement index *(percentile points)*		Number of studies	Number of students	Extent of evidence
		Average	Range			
Alphabetics	Positive effects	+9	–2 to +22	8	7,957	Medium to large
Reading fluency	Potentially positive effects	+12	+5 to +18	2	1,186	Medium to large
Comprehension	Mixed effects	+3	–11 to +19	8	9,733	Medium to large
General reading achievement	Mixed effects	+1	–7 to +14	6	2,574	Medium to large

Intervention Information

Background

Developed by Robert Slavin and Nancy Madden in conjunction with Johns Hopkins University, *SFA*® is distributed by the Success for All Foundation, Inc. Address: 300 E. Joppa Road, Suite 500, Baltimore, MD 21286. Telephone: (410) 616–2300. Fax: (410) 324–4444. Web: http://www.successforall.org/.

Intervention details

SFA® is a comprehensive school-level intervention that aims to improve the reading skills of children. *SFA*® combines literacy instruction (reading, writing, and oral language development curricula), which is the focus of this review, with whole-school reform elements. *SFA*® whole-school reform elements include tutoring for students who have difficulty reading at the same level as their classmates, quarterly assessments of student learning, family support teams for students' parents, a facilitator who works with school personnel to ensure they implement and coordinate all programs elements, and extensive training for all intervention teachers. Because the literacy instruction takes place in the context of the *SFA*® whole-school reform program, most of the students who received the *SFA*® reading curriculum also received some or all of the other program components. Ratings presented in this intervention report do not take into account the various ways schools implement the *SFA*® whole-school reform model.

SFA® elementary school reading programs combine cooperative-learning strategies with detailed lessons, which incorporate multimedia, puppet skits, and videos to support students' engagement and classroom instruction. *SFA*® emphasizes sequenced literacy instruction that spans several years, focusing on phonemic awareness skills initially and broader reading skills later. Students in prekindergarten through first grade participate in *Reading Roots*, and students in second grade and above participate in *Reading Wings*, in which students also learn to write compositions in various genres. In both of these programs, students are grouped into reading classes of 15–20 students with others of similar reading ability (regardless of age or grade level) during the regular, daily 90-minute reading period. Regrouping students who have demonstrated improvement in reading skills enables teachers to teach the whole class without having to organize the class into multiple smaller reading groups.

Teachers begin the period by reading children's storybooks aloud, which they then discuss with students to enhance understanding of the story and its structure, and to increase listening and speaking vocabulary. In kindergarten and first grade, teachers emphasize developing language skills and use phonetic storybooks and instruction to focus on phonemic awareness, auditory discrimination, and sound blending. In the second through fifth grades, teachers use school- or district-provided reading materials in a structured set of interactive activities in which students read, discuss, and write about the books. At this stage, teachers emphasize cooperative learning activities built around partner reading. Students work on identifying characters, settings, and problem solutions in narratives. Students also receive direct instruction in reading comprehension skills that involves explicit teaching using lectures or demonstrations of the material to students.

Implementing the reading program is the crux of the *SFA*® whole-school reform staff development model. This model emphasizes a relatively brief initial training with extensive classroom follow-up, coaching, and group discussion. School staff in their first year of implementing *SFA*® receive a 3-day summer training and 12 additional on-site support days during the school year. Developer-provided trainers visit and observe teachers each month in the first year and less often thereafter. Trainers visit classrooms, meet with teachers, examine data on children's progress, and provide feedback to school staff on implementation quality and outcomes. Each school implementing *SFA*® also has a facilitator on staff, usually an experienced teacher. School facilitators and other *SFA*® program staff make additional in-service presentations throughout the year, covering topics such as classroom management, instructional pace, and cooperative learning. Facilitators structure the in-service presentations to allow teachers to share problems and solutions, suggest changes, and discuss individual children. Principals and facilitators receive 5

days of initial training in leadership, data collection and progress monitoring, classroom instructional practices, school climate, and intervention using *SFA*® strategies. Regular in-service training, an annual *SFA*® conference, and on-site implementation support visits for school principals and teachers reinforce *SFA*® implementation after the first year.

Cost

As of October 2016, the average cost of *SFA*® for a school is $104 per child, per year.

Research Summary

The WWC identified 49 eligible studies that investigated the effects of *SFA*® on the reading skills of beginning readers. An additional 145 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 194 studies are in the References section, which begins on p. 12.

Table 2. Scope of reviewed research

Grade	K–4
Delivery method	Whole school
Intervention type	Curriculum

The WWC reviewed 49 eligible studies against group design standards. Two studies (Borman et al., 2007; Quint, Zhu, Balu, Rappaport, & DeLaurentis, 2015) were randomized controlled trials that meet WWC group design standards without reservations, and seven studies (Madden et al., 1993; Ross, Albert, McNelis, & Rakow, 1998; Ross & Casey, 1998a; Ross & Casey, 1998b; Ross, Smith, & Casey, 1995; Skindrud & Gersten, 2006; Tracey et al., 2014) used quasi-experimental designs that meet WWC group design standards with reservations. This report summarizes those nine studies. The remaining 40 studies do not meet WWC group design standards.

Summary of studies meeting WWC group design standards without reservations

Borman et al. (2007) conducted a cluster, or group-based, randomized controlled trial that examined the effects of *SFA*® on schools and students in grades K–5 across 12 states. The study randomly assigned two cohorts of schools: six schools in fall 2001 and 35 schools in fall 2002. In fall 2001, the study randomly assigned schools to receive either *SFA*® or business-as-usual literacy instruction in kindergarten through grade 2. In fall 2002, the study randomly assigned schools to receive *SFA*® either in kindergarten through grade 2 or in grades 3 through 5, with students in comparison groups receiving business-as-usual literacy instruction. For analyses of students in grades K–2, the study authors combined the two cohorts of schools (assigned in 2001 and 2002). For analyses of students in grades 3–5 (reported in Hanselman & Borman, 2013), the authors used only the fall 2002 cohort of schools. In both sets of analyses, the study compared outcomes for students who received the *SFA*® program for up to 3 years with those for students who took part in their schools' typical reading programs. The analyses examining schoolwide impacts on student achievement met WWC group design standards.[5] The WWC based its effectiveness rating on findings from the third-year sample[6] of 1,936 second-grade students in 18 intervention and 17 comparison schools who began the study in kindergarten,[7] and the first-year sample of 2,420 students who began the study in third grade in the 17 intervention and 18 comparison schools. Rather than analyzing only students who were in schools when random assignment occurred, the analytic sample included students who enrolled in study schools after random assignment. Because *SFA*® may have influenced where students attended school, findings for this sample reflect both the effect of *SFA*® on the outcomes of students and the effect of changes in the composition of students within study schools.

Quint et al. (2015) conducted a cluster randomized controlled trial that examined the effects of *SFA*® on schools and students in grades K–2 across four states in the western, southern, and northeastern United States. The study randomly assigned 37 schools to *SFA*® and the comparison condition, and compared outcomes of students who had completed 1, 2, or 3 years of the program with outcomes of students who took part in their schools' typical reading programs. The analyses examining schoolwide impacts on student achievement met WWC group design standards.[8] The WWC based its effectiveness rating on findings from the third-year sample of 2,907 students who began the study in kindergarten in the 19 intervention and 18 comparison schools. Rather than analyzing only students who were in schools when random assignment occurred, the analytic sample included students who enrolled in study schools after random assignment. Because *SFA*® may have influenced where students attend school, findings for this sample reflect both the effect of *SFA*® on the outcomes of students and the effect of changes in the composition of students within study schools.

Summary of studies meeting WWC group design standards with reservations

Madden et al. (1993) conducted a quasi-experimental study that examined the effects of *SFA*® on students in Baltimore City elementary schools. The study matched each of the five schools implementing *SFA*® with a similar comparison school. The five comparison schools had comparable percentages of students receiving free lunch and similar prior achievement levels. Over the course of 5 years, the study tracked outcomes for students enrolled in grades prekindergarten–4. The intervention encompassed two versions of the *SFA*® program: full implementation (two schools) and dropout prevention (three schools).[9] Compared with the full implementation model, the *SFA*® dropout prevention schools had fewer tutors and family support staff but included other components of *SFA*®. Ratings presented in this intervention report do not take into account the variations in *SFA*® implementation. This report includes findings in the alphabetics domain for students who received 3 years of *SFA*® and in three other outcome domains for students who received up to 5 years of *SFA*®.[10] Across the four outcome domains reported in the study, the largest combined analytic sample that contributed findings to the WWC effectiveness rating included 671 students in five *SFA*® schools and 671 students in five comparison schools.

Ross et al. (1998) conducted a quasi-experimental study that examined the effects of "alternative" schoolwide programs on students in 19 elementary schools in Washington State, of which four received *SFA*®. The study categorized the 19 schools into four groups based on their similarity on several characteristics, including enrollment, percentage of minority students, percentage of students eligible for free or reduced-price lunch, and prior academic performance. The authors compared outcomes between *SFA*® and comparison schools within each group. The WWC based its effectiveness rating on findings from a group that contained neither schools with the most disadvantaged nor the most affluent students in the sample, the only subsample that meets WWC group design standards. This group included three *SFA*® schools and two schools that implemented the *Accelerated Schools* program. The analytic sample included 128 students at the end of the second grade who had received 2 years of either *SFA*® or the *Accelerated Schools* program.

Ross and Casey (1998a) conducted a quasi-experimental study that examined the effects of *SFA*® in two elementary schools in Fort Wayne, Indiana, by comparing them with five schools that implemented locally developed schoolwide programs. The comparison schools were comparable to *SFA*® schools on pretest reading measures, socioeconomic status, and ethnicity of students in the grades studied. The WWC focused on students who started *SFA*® in kindergarten. The WWC based its effectiveness rating on findings from 288 students at the end of first grade who received 2 years of either *SFA*® or locally developed schoolwide programs.

Ross and Casey (1998b) conducted a quasi-experimental study that examined the effects of *SFA*® on students in four elementary schools in the state of Oregon. The study compared students receiving *SFA*® instruction in two schools with students in two schools in the same district who never participated in *SFA*®. The study reported student outcomes for two cohorts of students who started the program in kindergarten and first grade, respectively. Because the first-grade sample did not meet WWC group design standards, the WWC based its effectiveness rating on 1-year findings from 265 kindergarten students: 156 students in the two *SFA*® schools and 109 students in the two comparison schools.

Ross et al. (1995) conducted a quasi-experimental study that evaluated the effectiveness of *SFA*® in two elementary schools in Fort Wayne, Indiana. The study focused on students who started the program in kindergarten (in 1991, Cohort 1, and 1992, Cohort 2) and first grade (1991, Cohort 3). The WWC based its effectiveness rating on findings from students in third and fourth grades who received 4 years of *SFA*® (Cohorts 1 and 2), and ethnic minority students in grade 2 (Cohort 3) who received 3 years of *SFA*®.[11] The combined analytic sample included 128 students in the two *SFA*® schools and 77 students in the two comparison schools.

Skindrud and Gersten (2006) conducted a quasi-experimental study that examined the effects of *SFA*® in 12 elementary schools in the Sacramento City Unified School District (California). The study focused on two cohorts

of students who started the program during the 1997–98 school year, one that began in the second grade and another that began in the third grade. The study matched four schools implementing *SFA*® to eight schools implementing *Open Court Reading*© by poverty level as measured by the percentage of students eligible for free or reduced-price meals and the percentage of students receiving Aid to Families with Dependent Children. The WWC based its effectiveness rating on findings from 142 students in third grade who received 2 years of *SFA*® and 36 third-grade students who received 1 year of *SFA*®.[12] The analytic sample across the two cohorts included 178 students in the *SFA*® group and 353 students in the comparison group.

Tracey et al. (2014) conducted a quasi-experimental study that examined the effects of *SFA*® on students in 35 schools in England during the 2008–09 through 2010–11 school years. The study matched 20 schools implementing *SFA*® to 20 comparison schools on prior student achievement and demographics. The study compared outcomes for students who had completed 3 years of *SFA*® with outcomes for students who took part in their schools' typical reading programs. The WWC based its effectiveness rating on findings from the sample of 886 first-grade students who began the study in prekindergarten in 17 intervention and 18 comparison schools; 415 students were in the *SFA*® group and 471 students were in the comparison group.

Effectiveness Summary

The WWC review of *SFA*® for the Beginning Reading topic area includes outcomes in four domains: alphabetics, reading fluency, comprehension, and general reading achievement. The nine studies of *SFA*® that meet WWC group design standards reported findings in the four domains. The following findings present the authors' estimates and WWC-calculated estimates of the size and statistical significance of the effects of *SFA*® on beginning readers. Within each study, the primary findings that the WWC considered for the effectiveness rating are those measured at the period closest to the end of the intervention and reflect the maximum exposure of students to the program. Additional comparisons are available as supplemental findings in Appendix D. These 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. 70.

Summary of effectiveness for the alphabetics domain

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

Rating of effectiveness	Criteria met
Positive effects *Strong evidence of a positive effect with no overriding contrary evidence.*	In the eight studies that reported findings, the estimated impact of the intervention on outcomes in the *alphabetics* domain was positive and statistically significant for four studies, two of which meet WWC group design standards without reservations. No studies showed statistically significant or substantively important negative effects.
Extent of evidence	**Criteria met**
Medium to large	Eight studies that included 7,957 students in 137 schools reported evidence of effectiveness in the *alphabetics* domain.

Eight studies that meet WWC group design standards with or without reservations reported findings in the alphabetics domain.

Borman et al. (2007) examined scores on the Woodcock Reading Mastery Test (WRMT) and reported statistically significant positive effects of *SFA*® on two phonics subtests, Word Identification and Word Attack, for students in grade 2 who began receiving the intervention in kindergarten. The WWC confirmed the statistically significant positive effect only on the WRMT Word Attack subtest. The average effect size across the two outcomes was large enough to be substantively important according to WWC criteria (that is, an effect size of at least 0.25). The WWC characterizes these study findings as a statistically significant positive effect.

Quint et al. (2015) examined scores on the two subtests of the Woodcock-Johnson III (WJ-III) Tests of Achievement—Letter-Word Identification and Word Attack—and the Test of Word Reading Efficiency, and reported a statistically significant positive effect of *SFA*® on the WJ-III Word Attack subtest for students in grade 2 after 3 years of program implementation. The WWC confirmed the statistical significance of this finding after adjusting for multiple comparisons (that is, changing significance levels to take into account several comparisons). The WWC characterizes these study findings as a statistically significant positive effect.

Madden et al. (1993) reported findings on the Woodcock Language Proficiency Battery (WLPB) Letter-Word Identification and Word Attack subtests for students in grades 1, 2, and 3 who received the program for 3 years. The authors analyzed each matched pair of schools separately and found statistically significant positive effects for pairwise (that is, matched) school comparisons on the WLPB Word Attack subtest for students in grade 1 and statistically significant positive effects on the WLPB Letter-Word Identification subtest for students in grade 2.[13] The WWC confirmed statistically significant positive effects only on the Word Attack subtest for students in grade 1 after adjusting for multiple comparisons across the six alphabetics outcomes.[14] The average effect size across the outcomes was substantively important. The WWC characterizes these study findings as a statistically significant positive effect.

Ross et al. (1998) reported, and the WWC confirmed, no statistically significant effects of *SFA*® on students in grade 2 who received the program for 2 years, based on the WRMT Word Identification and Word Attack subtests. The average effect size across the two outcomes was not large enough to be substantively important. The WWC characterizes these study findings as an indeterminate effect.

Ross and Casey (1998a) reported no statistically significant effect of *SFA*® on the WRMT Word Identification subtest for students in grade 1 who received the program for 2 years but found a statistically significant positive effect on the other phonics measure, the WRMT Word Attack subtest.[15] The WWC found that neither of the effects was statistically significant after adjusting for clustering of students within schools, and the average effect was not large enough to be substantively important. The WWC characterizes these study findings as an indeterminate effect.

Ross and Casey (1998b) reported, and the WWC confirmed, no statistically significant effects of *SFA*® on kindergarteners who received the program for 1 year, based on the WRMT Word Identification and Word Attack subtests. The average effect size across the two outcomes was not large enough to be substantively important. The WWC characterizes these study findings as an indeterminate effect.

Ross et al. (1995) reported, and the WWC confirmed, no statistically significant effects of *SFA*® on the WRMT Word Identification and Word Attack subtests for third- and fourth-grade students who received the program for 4 years. The authors also reported, and the WWC confirmed, no statistically significant effects of *SFA*® on the WRMT Word Identification and Word Attack subtests for second-grade minority students who received the program for 3 years. The average effect size across the six outcomes was not substantively important. The WWC characterizes these study findings as an indeterminate effect.

Tracey et al. (2014) examined scores on the WRMT and reported statistically significant positive effects for *SFA*® students who received the program for 3 years on the Word Identification and Word Attack subtests. The WWC confirmed these findings. The WWC characterizes these study findings as a statistically significant positive effect.

Thus, for the alphabetics domain, four studies, two of which meet WWC group design standards without reservations, showed a statistically significant positive effect, and four studies showed an indeterminate effect. This results in a rating of positive effects, with a medium to large extent of evidence.

Summary of effectiveness for the reading fluency domain

Table 4. Rating of effectiveness and extent of evidence for the reading fluency domain

Rating of effectiveness	Criteria met
Potentially positive effects *Evidence of a positive effect with no overriding contrary evidence.*	In the two studies that reported findings, the estimated impact of the intervention on outcomes in the *reading fluency* domain was positive and substantively important for one study, and one study showed indeterminate effects.
Extent of evidence	**Criteria met**
Medium to large	Two studies that included 1,186 students in 45 schools reported evidence of effectiveness in the *reading fluency* domain.

Two studies that meet WWC group design standards with reservations reported findings in the reading fluency domain.

Madden et al. (1993) found a statistically significant positive effect on the Passage subtest of the Gray Oral Reading Test (GORT) for students in grade 4 who began receiving the intervention in kindergarten. After adjusting for clustering of students within schools, the WWC found that this effect was not statistically significant but was large enough to be substantively important. The WWC characterizes this study finding as a substantively important positive effect.

Tracey et al. (2014) reported findings on the two subtests of the York Assessment of Reading Comprehension (YARC), Accuracy and Reading Rate, for students who received the program for 3 years. The authors reported, and the WWC confirmed, no statistically significant or substantively important differences between students in the *SFA*® group and students in the comparison group. The average effect size across the two outcomes was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Thus, for the reading fluency domain, one study reported a substantively important positive effect, and one study reported an indeterminate effect. This results in a rating of potentially positive effects, with a medium to large extent of evidence.

Summary of effectiveness for the comprehension domain

Table 5. Rating of effectiveness and extent of evidence for the comprehension domain

Rating of effectiveness	Criteria met
Mixed effects *Evidence of inconsistent effects.*	In the eight studies that reported findings, the estimated impact of the intervention on outcomes in the *comprehension* domain was positive and statistically significant for one study, negative and substantively important for one study, and indeterminate for six studies.
Extent of evidence	**Criteria met**
Medium to large	Eight studies that included 9,733 students in 143 schools reported evidence of effectiveness in the *comprehension* domain.

Eight studies that meet WWC group design standards with or without reservations reported findings in the comprehension domain.

Borman et al. (2007) reported a statistically significant positive effect of *SFA*® on the WRMT Passage Comprehension subtest for students in grade 2 who began receiving the intervention in kindergarten. The WWC applied a clustering correction to unadjusted results for students in grade 2 and determined that the finding was not statistically significant. The study also reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the Gates-MacGinitie Reading Test for third-grade students after 1 year of program implementation. The average

effect size across the two outcomes was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Quint et al. (2015) reported, and the WWC confirmed, no statistically significant difference between second-grade *SFA*® students and comparison students on the WJ-III Passage Comprehension subtest after 3 years of program implementation. The effect size was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Madden et al. (1993) reported statistically significant positive effects of *SFA*® on the GORT Comprehension subtest for students in grade 4 who received the program for 5 years. After adjusting for clustering of students within schools, the WWC determined that this effect was not statistically significant. The authors also reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the Comprehensive Tests of Basic Skills (CTBS) Total Reading subtest for students in grade 4 who received the program for 5 years. The authors reported, and the WWC confirmed, a statistically significant positive effect of SFA® on the WRMT Passage Comprehension subtest for second-grade students who scored in the lowest quartile at baseline who received the program for 4 years. The authors reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the Durrell Analysis of Reading Difficulty (DARD) Silent Reading test for students in grade 2 who received the program for 3 years. The average effect size across these outcomes was substantively important. The WWC characterizes these study findings as a statistically significant positive effect.

Ross et al. (1998) reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the WRMT Passage Comprehension subtest for second-grade students who began receiving the intervention in first grade. The effect size was negative and substantively important. The WWC characterizes this study finding as a substantively important negative effect.

Ross and Casey (1998a) reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the WRMT Passage Comprehension subtest for first-grade students who began receiving the intervention in kindergarten. The effect size was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Ross and Casey (1998b) reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the Passage Comprehension subtest of the WRMT for kindergarteners who received the program for 1 year. The effect size was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Ross et al. (1995) reported, and the WWC confirmed, no statistically significant effects of *SFA*® on the WRMT Passage Comprehension subtest for third- and fourth-grade students who received the program for 4 years. The authors also reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the WRMT Passage Comprehension subtest for second-grade minority students who received the program for 3 years. The average effect size across the three grades was not large enough to be substantively important. The WWC characterizes these study findings as an indeterminate effect.

Tracey et al. (2014) reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the YARC Comprehension subtest for students who received the program for 3 years. The effect size was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Thus, for the comprehension domain, six studies showed an indeterminate effect, one study showed a statistically significant positive effect, and one study showed a substantively important negative effect. This results in a rating of mixed effects, with a medium to large extent of evidence.

Summary of effectiveness for the general reading achievement domain

Table 6. Rating of effectiveness and extent of evidence for the general reading achievement domain

Rating of effectiveness	Criteria met
Mixed effects *Evidence of inconsistent effects.*	In the six studies that reported findings, the estimated impact of the intervention on outcomes in the *general reading achievement* domain was positive and substantively important for one study and indeterminate for five studies.
Extent of evidence	**Criteria met**
Medium to large	Six studies that included 2,574 students in 42 schools reported evidence of effectiveness in the *general reading achievement* domain.

Six studies that meet WWC group design standards with or without reservations reported findings in the general reading achievement domain.

Madden et al. (1993) reported a statistically significant positive effect of *SFA*® on the CTBS Total Language scores for students in grade 4 who received the program for 5 years. After adjusting for clustering of students within schools, the WWC did not find the result to be statistically significant. The authors also reported findings on the DARD Oral Reading subtest for students in grades 1 and 3 who received the program for 3 years. The authors reported statistically significant positive effects of *SFA*® on the Durrell Oral Reading subtest for students in grade 3 for each pair of matched schools,[16] but the WWC found that the average effect size across schools was not statistically significant after adjusting for clustering of students within schools. The average effect size across the three outcomes was large enough to be substantively important. The WWC characterizes this study finding as a substantively important positive effect.

Ross et al. (1998) reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the Durrell Oral Reading subtest for second-grade students who began receiving the intervention in first grade. The effect size was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Ross and Casey (1998a) reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the DARD Oral Reading subtest for first-grade students who began receiving the intervention in kindergarten. The effect size was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Ross and Casey (1998b) found, and the WWC confirmed, no statistically significant effect of *SFA*® on the Oral Reading subtest of the DARD test for kindergarteners who received the program for 1 year. The effect size was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Ross et al. (1995) reported, and the WWC confirmed, no statistically significant effects of *SFA*® on the Durrell Oral Reading subtest for third-grade students and on the Gray Oral Reading Test for fourth-grade students who received the program for 4 years. The authors also reported, and the WWC confirmed, no statistically significant effects of *SFA*® on the Durrell Oral Reading subtest for second-grade students who received the program for 3 years. The average effect size across the three outcomes was not large enough to be substantively important. The WWC characterizes these study findings as an indeterminate effect.

Skindrud and Gersten (2006) found a statistically significant negative effect of *SFA*® on the reading subtest of the Stanford Achievement Test, 9th Edition (SAT-9) for students in grade 3 who received the program for 2 years. However, after adjusting for clustering of students within schools, the WWC does not find the result to be statistically significant. The authors also reported, and the WWC confirmed, no statistically significant effect of *SFA*® on the SAT-9 Language subtest for third-grade students who scored in the lowest quartile on reading achievement at baseline after receiving the program for 1 year. The average effect size across the two outcomes was not large enough to be substantively important. The WWC characterizes this study finding as an indeterminate effect.

Thus, for the general reading achievement domain, one study showed a substantively important positive effect, and five studies showed an indeterminate effect. 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

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal, 44*(3), 701–731.

Additional sources:

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., & Madden, N. A. (2004). *Success for All: Preliminary first-year results from the national randomized field trial*. Baltimore, MD: Success for All Foundation.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005a). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42*(4), 673–696. Retrieved from ERIC: https://eric.ed.gov/?id=ED485351

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005b). Success for All: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27*(1), 1–22.

Hanselman, P., & Borman, G. D. (2013). The impacts of Success for All on reading achievement in grades 3–5: Does intervening during the later elementary grades produce the same benefits as intervening early? *Educational Evaluation and Policy Analysis, 35*(2), 237–251.

Slavin, R. E., Madden, N. A., Cheung, A., Chamberlain, A., Chambers, B., & Borman, G. (2005). *A randomized evaluation of Success for All: Second-year outcomes*. Baltimore, MD: Success for All Foundation.

Quint, J. C., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015). *Scaling up the Success for All model of school reform: Final report from the Investing in Innovation (i3) Scale-Up*. New York, NY: MDRC.

Additional sources:

Quint, J., Zhu, P., Doolittle, F., & Society for Research on Educational Effectiveness. (2012). *Understanding variation in implementation of SFA in the i3 Scale-Up project.* Washington, DC: Society for Research on Educational Effectiveness. Retrieved from ERIC: https://eric.ed.gov/?id=ED530361

Quint, J. C., Balu, R., DeLaurentis, M., Rappaport S., Smith, T. J., & Zhu, P. (2013). *The Success for All model of school reform: Early findings from the Investing in Innovation (i3) Scale-Up*. New York, NY: MDRC. Retrieved from ERIC: https://eric.ed.gov/?id=ED545452

Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T. J., & Zhu, P. (2014). *The Success for All model of school reform: Interim findings from the Investing in Innovation (i3) Scale-Up.* New York, NY: MDRC. Retrieved from ERIC: https://eric.ed.gov/?id=ED546642

Studies that meet WWC group design standards with reservations

Madden, N. A., Slavin, R. E., Karweit, N., Dolan, L., & Wasik, B. A. (1993). Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal, 30*(1), 123–148.

Additional sources:

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24*(4), 243–266.

Madden, N. A., Slavin, R. E., Karweit, N., Dolan, L., & Wasik, B. A. (1991). *Success for All: Multi-year effects of a schoolwide elementary restructuring program* [Baltimore, MD]. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students. Retrieved from ERIC: https://eric.ed.gov/?id=ED336492

Slavin, R. E., Madden, N. A., Dolan, L. J., & Wasik, B. A. (1993). *Success for All in the Baltimore City Public Schools: Year 6 report*. Baltimore, MD: Johns Hopkins University, Center for Research in Effective Schooling for Disadvantaged Students.

Slavin, R. E., Madden, N. A., Karweit, N. L., Dolan, L., & Wasik, B. A. (1990). *Success for All: Second year report*. Baltimore, MD: Baltimore Public Education Institute and Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Slavin, R. E., Madden, N. A., Karweit, N. L., Dolan, L., & Wasik, B. A. (1993). *Success for All in the Baltimore City Public Schools: Year 5 report*. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Ross, S. M., Alberg, M., McNelis, M., & Rakow, J. (1998). *Evaluation of elementary school school-wide programs: Clover Park School District year 2: 1997–98*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Additional source:

Ross, S. M., Alberg, M., & McNelis, M. (1997). *Evaluation of elementary school school-wide programs: Clover Park School District, year 1: 1996–97*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., & Casey, J. (1998a). *Longitudinal study of student literacy achievement in different Title I school-wide programs in Fort Wayne Community Schools Year 2: First grade results.* Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., & Casey, J. (1998b). *Success for All evaluation: 1997–1998 Tigard-Tualatin School District*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., & Casey, J. (1995). *Final report: 1994–1995 Success for All program in Fort Wayne, Indiana*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Additional sources:

Casey, J., Smith, L. J., & Ross, S. M. (1994). Final report: 1993–1994 *Success for All program in Fort Wayne, Indiana*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., & Casey, J. (1997). Preventing early school failure: Impacts of Success for All on standardized test outcomes, minority group performance, and school effectiveness. *Journal of Education for Students Placed at Risk, 2*(1), 29–53.

Ross, S. M., Smith, L. J., & Casey, J. (1999). "Bridging the gap": The effects of the Success for All program on elementary school reading achievement as a function of student ethnicity and ability level. *School Effectiveness and School Improvement, 10*(2), 129–150.

Ross, S. M., Smith, L. J., Casey, J., & Johnson, B. (1993). *Final report: 1992–93 Success for All program in Ft. Wayne, Indiana*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., Casey, J., Johnson, B., & Bond, C. (1994, April). *Using "Success For All" to restructure elementary schools: A tale of four cities* [Ft. Wayne, IN]. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Smith, L. J., Ross, S. M., & Casey, J. (1996) Multi-site comparison of the effects of Success for All on reading achievement [Ft. Wayne, IN]. *Journal of Literacy Research, 28*(3), 329–353.

Smith, L. J., Ross, S. M., Faulks, A., Casey, J., Shapiro, M., & Johnson, B. (1993). *1991–1992 Ft. Wayne, Indiana Success for All results*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Skindrud, K., & Gersten, R. (2006). An evaluation of two contrasting approaches for improving reading achievement in a large urban district. *Elementary School Journal, 106*(5), 389–407.

Tracey, L., Chambers, B., Slavin, R. E., Madden, N. A., Cheung, A., & Hanley, P. (2014). Success for All in England: Results from the third year of a national evaluation. *SAGE Open, 4*(3), 1–10.

Studies that do not meet WWC group design standards

Atkinson, C. L. H. (1998). *An analysis of the impact of Success for All on reading, attendance, and academic self-efficacy with at-risk elementary school students* (Doctoral dissertation). Available from ProQuest Dissertations

and Theses database. (UMI No. 9905180) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Bifulco, R. (2001). *Do whole-school reform models boost student performance: Evidence from New York City* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3019134) The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Chambers, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2005). *Enhancing Success for All for Hispanic students: Effects on beginning reading achievement*. Baltimore, MD: Success for All Foundation. Retrieved from ERIC: https://eric.ed.gov/?id=ED485350 The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Chambers, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2004). *Effects of Success for All with embedded video on the beginning reading achievement of Hispanic children*. Baltimore, MD: Center for Research on the Education of Students Placed at Risk.

Cheung, A., & Slavin, R. (2014). *Effects of Success for All on reading achievement: A secondary analysis using data from the Study of Instructional Improvement (SII)*. Baltimore, MD: Center for Research and Reform in Education. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Correnti, R. (2009, March). *Examining CSR program effects on student achievement: Causal explanation through examination of implementation rates and student mobility.* Paper presented at the second annual conference of the Society for Research on Educational Effectiveness, Washington, DC. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Rowan, B., Correnti, R., Miller, R., & Camburn, E. (2009). *School improvement by design: Lessons from a study of comprehensive school reform programs.* Philadelphia, PA: Consortium for Policy Research in Education. Retrieved from ERIC: https://eric.ed.gov/?id=ED507546

Datnow, A., Borman, G. D., Stringfield, S., Overman, L. T., & Castellano, M. (2003). Comprehensive school reform in culturally and linguistically diverse contexts: Implementation and outcomes from a four-year study. *Educational Evaluation and Policy Analysis, 25*(2), 143–170. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Dianda, M., & Flaherty, J. (1995, April). *Effects of Success for All on the reading achievement of first graders in California bilingual programs*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional sources:

Dianda, M. R., Flaherty, J. F., & Southwest, R. L. (1995). *Report on workstation uses: Effects of Success for All on the reading achievement of first graders in California bilingual programs.* Washington, DC: Office of Educational Research and Improvement. Retrieved from ERIC: https://eric.ed.gov/?id=ED394327

Livingston, M., & Flaherty, J. (1997). *Effects of Success for All on reading achievement in California schools.* Los Alamitos, CA: WestEd.

Slavin, R. E., & Madden, N. (1999). Effects of bilingual and English as a second language adaptations of Success for All on the reading achievement of students acquiring English [California]. *Journal of Education for Students Placed at Risk, 4*(4), 393–416. Retrieved from ERIC: https://eric.ed.gov/?id=ED432927

Fuller, E. A. (2013). *The impact of selected initiatives on the reading criterion referenced competency test scores of African-American and disadvantaged students in grades 3, 5, and 8* (Doctoral dissertation). Available from Pro-

Quest Dissertations and Theses database. (UMI No. 3593031) The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Gander, B. D. (2008). *A comparison of early reading outcomes and program costs in four primary reading programs for improved decision-making* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3293995) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Grehan, A. W. (2001). *The effects of the Success for All program on improving reading readiness skills for at-risk students in kindergarten* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3029892) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Jackson, W. D. (2008). *An investigation of the impact of the Success for All whole-school reform model on the Elementary School Proficiency Assessment and the New Jersey Assessment of Skills and Knowledge in an urban district* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3301057) The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Jones, E., Gottfredson, G., & Gottfredson, D. (1997). Success for some: An evaluation of a Success for All program. *Evaluation Review, 21*(6), 643–670. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

McCollum-Rogers, S. A. (2004). *Comparing direct instruction and Success for All with a basal reading program in relation to student achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3149920) The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

McCollum, S., McNeese, M. N., Styron, R., & Lee, D. E. (2007). A school district comparison of reading achievement based on three reading programs. *Journal of At-Risk Issues, 13*(1), 1–6. Retrieved from ERIC: https://eric.ed.gov/?id=EJ853383

McCoy, E. B. (2005). *The performance of student participants in externally developed Title I schoolwide reading interventions in the Little Rock School District as measured by national, state and local standardized tests* (Unpublished doctoral dissertation). University of Arkansas at Little Rock. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Muñoz, M. A., & Dossett, D. H. (2004). Educating students placed at risk: Evaluating the impact of Success for All in urban settings. *Journal of Education for Students Placed at Risk, 9*(3), 261–277. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Munoz, M. A., Dossett, D., & Judy-Gullans, K. (2003). *Educating students placed at risk: Evaluating the impact of Success for All in urban settings*. Louisville, KY: Jefferson County Public Schools. Retrieved from ERIC: https://eric.ed.gov/?id=ED480178

Nunnery, J. A., Slavin, R. E., Madden, N. A., Ross, S. M., Smith, L. J., Hunter, P., & Stubbs, J. (1997, March). *Effects of full and partial implementations of Success for All on student reading achievement in English and Spanish.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional sources:

Nunnery, J. A. (1995). *An assessment of Success for All program component effects on the reading achievement of at-risk first-grade students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9615378)

Nunnery, J. A., Slavin, R. E., Ross, S. M., Smith, L. J., Hunter, P., & Stubbs, J. (1996, April). *An assessment of Success for All program component configuration effects on the reading achievement of at-risk first grade students.* Paper presented at the meeting of the American Educational Research Association, New York, NY.

Ross, S. M., Fleischman, S. W., & Hornbeck, M. (2003). *Progress and options regarding the implementation of Direct instruction and Success for All in Toledo public schools.* Memphis, TN: University of Memphis, Center for Research in Educational Policy. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Ross, S. M., McNelis, M., Lewis, T., & Loomis, S. (1998). *Evaluation of Success for All programs: Little Rock School District year 1: 1997–1998*. Memphis, TN: University of Memphis, Center for Research in Educational Policy. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Wang, L. W., & Ross, S. M. (1999). *Evaluation of Success for All program: Little Rock School District year 2: 1998–1999.* Memphis, TN: University of Memphis, Center for Research in Education Policy.

Ross, S. M., Nunnery, J. A., Goldfeder, E., McDonald, A., Rachor, R., Hornbeck, M., & Fleischman, S. (2004). Using school reform models to improve reading achievement: A longitudinal study of direct instruction and Success for All in an urban district. *Journal of Education for Students Placed at Risk, 9*(4), 357–388. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Ross, S. M., Nunnery, J. A., & Smith, L. J. (1996). Evaluation of Title I reading programs: *Amphitheater Public Schools—year 1: 1995–1996.* Memphis, TN: University of Memphis, Center for Research in Educational Policy. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Ross, S. M., & Smith, L. J. (1994). Effects of the Success for All model on kindergarten through second-grade reading achievement, teachers' adjustment, and classroom-school climate at an inner-city school. *The Elementary School Journal, 95*(2), 121–138. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Additional sources:

Ross, S. M., & Smith, L. J. (1992). *Final report: 1991–92 Success for All program in Memphis, Tennessee*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., Casey, J., Johnson, B., & Bond, C. (1994, April). *Using Success for All to restructure elementary schools: A tale of four cities* [Memphis, TN]. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Ross, S. M., Smith, L. J., Crawford, A., Eck, L., Lohr, L., & Faulks, A. (1991). *Final report: Success for All 1990–91 Memphis program*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Smith, L. J., Ross, S. M., & Casey, J. (1994). *Final report: 1993–1994 Success for All program in Memphis, Tennessee.* Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., & Casey, J. (1997). *Final report: 1996–97 Success for All program in Clark County, Georgia.* Memphis, TN: University of Memphis, Center for Research in Education Policy. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Ross, S. M., Smith, L. J., & Nunnery, J. A. (1998, April). *The relationship of program implementation quality and student achievement.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Schneider, F. H. (1999). *Impact of the Success for All program in the teaching of reading for third grade students in selected elementary schools in the Pasadena Independent School District* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9934489) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Simpson, S. H. (1997). *A principal's perspective of the implementation of Reading Recovery in six metropolitan Nashville elementary schools* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9806596) The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Slavin, R. E., Madden, N. A., Karweit, N., Livermon, B. J., & Dolan, L. (1990). *Success for All: First-year outcomes of a comprehensive plan for reforming urban education. American Educational Research Journal, 27*(2), 255–278. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Slavin, R. E., & Madden, N. (1999). Effects of bilingual and English as a second language adaptations of Success for All on the reading achievement of students acquiring English [Philadelphia, PA]. *Journal of Education for Students Placed at Risk, 4*(4), 393–416. Retrieved from ERIC: https://eric.ed.gov/?id=ED432927 The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Additional sources:

Madden, N. A., Slavin, R. E., Karweit, N. L., Dolan, L., & Wasik, B. (1991). *Success for All: Multi-year effects of a schoolwide elementary restructuring program* [Philadelphia, PA]. Baltimore, MD: The Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students. Retrieved from ERIC: https://eric.ed.gov/?id=ED336492

Slavin, R. E., Madden, N. A., Dolan, L. J., & Wasik, B. A. (1993). *Success for All: Evaluations of national replications* [Philadelphia, PA]. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students. Retrieved from ERIC: https://eric.ed.gov/?id=ED362606

Slavin, R. E., Madden, N. A., Dolan, L. J., & Wasik, B. A. (1994). *Implementing Success for All in the Philadelphia public schools: Final report to the Pew Foundation.* Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Slavin, R. E., & Madden, N. A. (1991). *Success for All at Buckingham Elementary: Second year evaluation.* Baltimore, MD: The Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Slavin, R. E., & Yampolsky, R. (1991). *Effects of Success for All on students with limited English proficiency: A three-year evaluation.* Baltimore, MD: Center for Research on Effective Schooling for Disadvantaged Students. Retrieved from ERIC: https://eric.ed.gov/?id=ED346199 The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Additional sources:

Slavin, R. E., Leighton, M., & Yampolsky, R. (1990). *Success for All: Effects on the achievement of limited English proficient children (Report No. 5)*. Baltimore, MD: The Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students. Retrieved from ERIC: https://eric.ed.gov/?id=ED331585

Slavin, R. E., & Yampolsky, R. (1991). *Success for All: Effects on language minority students.* Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students. Retrieved from ERIC: https://eric.ed.gov/?id=ED331294

Smith, L. J., Ross, S. M., & Casey, J. (1996). Multi-site comparison of the effects of Success for All on reading achievement [Caldwell, ID]. *Journal of Literacy Research, 28*(3), 329–353. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional sources:

Ross, S. M., Smith, L. J., & Casey, J. (1992). *Final report: 1991–92 Success for All program in Caldwell, Idaho.* Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., Casey, J., Johnson, B., & Bond, C. (1994, April). *Using Success for All to restructure elementary schools: A tale of four cities* [Caldwell, ID]. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. Retrieved from ERIC: https://eric.ed.gov/?id=ED373456

Ross, S. M., Smith, L. J., Casey, J., & Slavin, R. E. (1995). Increasing the academic success of disadvantaged children: An examination of alternative early intervention programs. *American Educational Research Journal, 32*(4), 773–800.

Smith, L. J., Ross, S. M., Johnson, B., & Casey, J. (1993). *Final report: 1992–1993 Memphis, Tennessee Success for All results.* Memphis, TN: University of Memphis, Center for Research in Educational Policy. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Smith-Davis, S. L. (2007). *Does Success for All impact reading achievement of students with learning disabilities* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3276390) The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Sterbinsky, A., Ross, S. M., & Redfield, D. (2006). Effects of comprehensive school reform on student achievement and school change: A longitudinal multi-site study. *School Effectiveness and School Improvement, 17*(3), 367–397. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Stringfield, S., Millsap, M. A., & Herman, R. (1997). *Urban and suburban/rural special strategies for educating disadvantaged children: Findings and policy implications of a longitudinal study*. Baltimore, MD: Johns Hopkins University, Center for Social Organization of Schools. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Tivnan, T., & Hemphill, L. (2005). Comparing four literacy reform models in high-poverty schools: Patterns of first-grade achievement. *The Elementary School Journal, 105*(5), 419–441. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Urdegar, S. M. (2001). *Evaluation of the Success for All program, 1999–2000*. Miami, FL: Miami-Dade Public Schools, Office of Evaluation Research. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional sources:

Urdegar, S. M. (1998). *Evaluation of the Success for All program 1997–1998.* Miami, FL: Miami-Dade County Public Schools, Office of Evaluation and Research.

Urdegar, S. M. (2000). *Evaluation of the Success for All program 1998–99*. Miami, FL: Miami-Dade County Public Schools, Office of Evaluation and Research.

Veals, C. J. (2003). *The impact of the Success for All reading program on the reading performance of third grade students in two southwest Mississippi schools* (Doctoral dissertation). Available from ProQuest Dissertations

and Theses database. (UMI No. 3049586) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Venezky, R. L. (1998). An alternative perspective on Success for All. *Advances in Educational Policy, 4*, 145–165. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Wang, L. W., & Ross, S. M. (1999). *Results for Success for All program: Alhambra School District*. Memphis, TN: University of Memphis, Center for Research in Educational Policy. The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention.

Wang, L. W., & Ross, S. M. (2003). *Comparisons between elementary school programs on reading performance: Albuquerque Public Schools.* Memphis, TN: University of Memphis, Center for Research in Educational Policy. The study does not meet WWC group design standards because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Studies that are ineligible for review using the Beginning Reading Evidence Review Protocol

Ahearn, E. M. (1994). *Involvement of students with disabilities in the New American Schools Development Corporation projects*. Alexandria, VA: National Association of State Directors of Special Education. Retrieved from ERIC: https://eric.ed.gov/?id=ED371513 The study is ineligible for review because it is out of the scope of the protocol.

Ashe, D. L. (2014). *The impact of comprehensive school reform programs on teacher self-efficacy* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3636180) The study is ineligible for review because it is out of the scope of the protocol.

Barnes, C., Camburn, E., Kim, J. S., & Rowan, B. (2005, April). *School leadership and instructional improvement in CSR schools*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA. The study is ineligible for review because it is out of the scope of the protocol.

Berends, M., Chun, J., Schuyler, G., Stockly, S., & Briggs, R. J. (2002). *Challenges of conflicting school reforms: New American schools in a high-poverty district*. Santa Monica, CA: RAND Cooperation. The study is ineligible for review because it does not use an eligible design.

Berends, M., Kirby, S. N., Naftel, S., & McKelvey, C. (2000). *Implementation in a longitudinal sample of new American schools: Three years into scale-up*. Santa Monica, CA: RAND Cooperation. The study is ineligible for review because it does not use an eligible design.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73(2), 125–230. The study is ineligible for review because it does not use an eligible design.

Chamberlain, A., Daniels, C., Madden, N. A., & Slavin, R. E. (2009). A randomized evaluation of the Success for All middle school reading program. In D. L Hough (Ed.), *Middle grades research: Exemplary studies linking theory to practice* (pp. 21–41). Charlotte, NC: Information Age Publishing. The study is ineligible for review because it does not use a sample aligned with the protocol.

Chambers, B., Abrami, P. C., & Morrison, S. (2001). Can Success for All succeed in Canada? In R. E. Slavin & N. A. Madden (Eds.), *Success for All: Research and reform in elementary education* (pp. 93–109). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. The study is ineligible for review because it does not use a sample aligned with the protocol.

Chambers, B., Slavin, R. E., Madden, N. A., Abrami, P. C., Tucker, B. J., Cheung, A., & Gifford, R. (2008). Technology infusion in Success for All: Reading outcomes for first graders. *The Elementary School Journal, 109*(1), 1–15. The study is ineligible for review because it does not use an eligible design.

Cheung, A., Ledesma, J. A., & Fung, A. (2009). The effectiveness of the Success for All reading programme on primary ELA pupils in Hong Kong. *Effective Education, 1*(2), 123–134. The study is ineligible for review because it does not use a sample aligned with the protocol.

Cheung, A. C. K., & Slavin, R. E. (2012). Effective reading programs for Spanish-dominant English language learners (ELLs) in the elementary grades: A synthesis of research. *Review of Educational Research, 82*(4), 351–395. The study is ineligible for review because it does not use a sample aligned with the protocol.

Cooper, R., Slavin, R. E., & Madden, N. A. (1997). *Success for All: Exploring the technical, normative, political, and socio-cultural dimensions of scaling up*. Baltimore, MD: Center for Research on the Education of Students Placed at Risk. Retrieved from ERIC: https://eric.ed.gov/?id=ED412324 The study is ineligible for review because it does not use an eligible design.

Crowe, E. C., Connor, C. M., & Petscher, Y. (2009). Examining the core: Relations among reading curricula, poverty, and first through third grade reading achievement. *Journal of School Psychology, 47*(3), 187–214. The study is ineligible for review because it does not use an eligible design.

Datnow, A., & Castellano, M. (2000). Teachers' responses to Success for All: How believes, experiences, and adaptations shape implementation. *American Educational Research Journal, 37*, 775–799. The study is ineligible for review because it is out of the scope of the protocol.

Dere, M. (2006). *Success for All and America's Choice: A comparative evaluation of two alternative instructional programs for elementary school students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3203148) The study is ineligible for review because it does not use a sample aligned with the protocol.

Dolge, A. (2012). Cooperative learning helps turn around struggling school. *Education Daily, 45*(86), 1–2. The study is ineligible for review because it does not use a sample aligned with the protocol.

Ford, T. G. (2014). Trust, control, and comprehensive school reform: Investigating growth in teacher-teacher relational trust in Success for All schools. In D. M. Maele, P. B. Forsyth, and M. V. Houtte (Eds.), *Trust and school life* (pp. 229–258). Netherlands: Springer Netherlands. The study is ineligible for review because it is out of the scope of the protocol.

Goldenberg, C., & Coleman, R. (2010). *Promoting academic achievement among English learners: A guide to the research*. Thousand Oaks, CA: Corwin Press. The study is ineligible for review because it does not use an eligible design.

Good, T. L., & McCaslin, M. (2008). What we learned about research on school reform: Considerations for practice and policy. *Teachers College Record, 110*(11), 2475–2495. The study is ineligible for review because it is out of the scope of the protocol.

Greenlaw, M. J. (2004). *A case study examining the relationships among teachers' perceptions of the Success for All reading program, teachers' sense of efficacy, students' attitudes toward reading and students' reading achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3139431) The study is ineligible for review because it does not use an eligible design.

Hankerson, K. M. (2004). *A cross-case study of the practices of the Success for All (SFA) facilitator* (Doctoral dissertation). Available from ProQuest Dissertation and Theses database. (UMI No. 3126318) The study is ineligible for review because it is out of the scope of the protocol.

Harris, A., Hopkins, D., & Wordsworth, J. (2001). The implementation and impact of Success for All in English schools. In R. E. Slavin & N. A. Madden (Eds.), *Success for All: Research and reform in elementary education* (pp. 81–92). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. The study is ineligible for review because it is out of the scope of the protocol.

Hertz-Lazarowitz, R. (2001). Success for All: A community model for advancing Arabs and Jews in Israel. In R. E. Slavin & N. A. Madden (Eds.), *Success for All: Research and reform in elementary education* (pp. 149–177). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. The study is ineligible for review because it is out of the scope of the protocol.

Hess, P. M. (2004). *A study of teachers' selection and implementation of meta-cognitive reading strategies for fourth/fifth grade reading comprehension from a Success for All reading program perspective: Moving beyond*

the fundamentals (Doctoral dissertation). Available from ProQuest Dissertation and Theses database. (UMI No. 3140930) The study is ineligible for review because it does not use a sample aligned with the protocol.

Hurley, E. A., Chamberlain, A., Slavin, R. E., & Madden, N. A. (2001). Effects of Success for All on TAAS reading scores. *Phi Delta Kappan, 82*(10), 750. The study is ineligible for review because it is out of the scope of the protocol.

James, D. W., Jurich, S., & Estes, S. (2001). *Raising minority academic achievement: A compendium of education programs and practices*. Washington, DC: American Youth Policy Forum. Retrieved from ERIC: https://eric.ed.gov/?id=ED473901 The study is ineligible for review because it does not use an eligible design.

James, L. R. D. (2003). *The effect of the Success for All reading approach on fourth- and fifth-grade students' standardized reading assessment scores* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3072259) The study is ineligible for review because it does not use a sample aligned with the protocol.

Kapushion, B. M. (2003). *A qualitative study of "Success for All—Roots and Wings" on four Jefferson County schools* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3078189) The study is ineligible for review because it does not use an eligible design.

Koh, M. S., & Robertson, J. S. (2003). School reform models and special education. *Education and Urban Society, 35*(4), 421–442. The study is ineligible for review because it does not use an eligible design.

Korelich, K. A. (2015). *The Success for All reading program and effect on student achievement in a south central Texas major suburban school district* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3722594) The study is ineligible for review because it does not use a sample aligned with the protocol.

Lewis, J. L., & Bartz, M. (1999). *New American Schools designs: An analysis of program results in district schools—Cincinnati Public Schools.* Cincinnati, OH: Cincinnati Public Schools, Research and Evaluation Office. The study is ineligible for review because it does not use an eligible design.

Lucius, L. B. (2000). *A comparison of three kindergarten curricula on language and literacy performance* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3003007) The study is ineligible because it is out of the scope of the protocol.

Madden, N. A. (2006). Reducing the gap: Success for All and the achievement of African American students. *Journal of Negro Education, 75*(3), 389–400. The study is ineligible for review because it is out of the scope of the protocol.

Madden, N. A., Slavin, R. E., & Simons, K. (1999). *MathWings: Effects on student mathematics performance* (Report No. 39). Baltimore, MD: Center for Research on the Education of Students Placed at Risk. Retrieved from ERIC: https://eric.ed.gov/?id=ED431631 The study is ineligible for review because it does not use an eligible design.

Manset, G., St. John, E. P., Simmons, A., Michael, R., Bardzell, J., Hodges, D., … Gordon, D. (1999). *Indiana's early literacy intervention grant program impact study for 1997–98*. Retrieved from ERIC: https://eric.ed.gov/?id=ED439392 The study is ineligible for review because it is out of the scope of the protocol.

Mason, B. (2005). *Achievement effects of five comprehensive school reform designs implemented in Los Angeles Unified School District* (Doctoral dissertation). Pardee RAND Graduate School, Santa Monica, CA. The study is ineligible for review because it does not use a sample aligned with the protocol.

Pogrow, S. (2002). Success for All is a failure. *Phi Delta Kappan, 83*(6), 463–468. The study is ineligible for review because it does not use an eligible design.

Reis, S. M., McCoach, D. B., Coyne, M., Schreiber, F., Eckert, R. D., & Gubbins, E. J. (2007). Using planned enrichment strategies with direct instruction to improve reading fluency, comprehension, and attitude toward reading: An evidence-based study. *The Elementary School Journal, 108*(1), 3–25. The study is ineligible for review because it is out of the scope of the protocol.

Ross, S. M., Sanders, W. L., & Wright, S. P. (2000). *Fourth-year achievement results on the Tennessee Value-Added Assessment System for restructuring schools in Memphis*. Memphis, TN: University of Memphis, Center for Research in Educational Policy. This study is ineligible because it does not use a sample aligned with the protocol.

Ross, S. M., Smith, L. J., & Nunnery, J. A. (1998, April). *Title I as a catalyst for school improvement: Impact of alternative school-wide models on the reading achievement of students at risk*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA. The study is ineligible for review because it is out of the scope of the protocol.

Ross, S. M., Tabachnick, S., & Sterbinsky, A. (2002). *Using comprehensive school reform models to raise student achievement: Factors associated with success in Memphis.* Memphis TN: University of Memphis. The study is ineligible for review because it is out of the scope of the protocol.

Ross, S. M., Wang, L. W., Sanders, W. L., & Wright, S. P. (1999). *Teacher mobility and effectiveness in restructuring and non-restructuring schools in an inner-city district*. Memphis, TN: University of Memphis. Retrieved from SAS Institute website: https://www.sas.com/. The study is ineligible for review because it is out of the scope of the protocol.

Ross, S. M., Wang, L. W., Sanders, W. L., Wright, S. P., & Stringfield, S. (1999). *Two and three year achievement results on the Tennessee Value-Added Assessment System for restructuring schools in Memphis*. Memphis, TN: University of Memphis, Center for Research in Educational Policy. The study is ineligible for review because it is out of the scope of the protocol.

Sanders, W. L., Wright, S. P., Ross, S. M., & Wang, L. W. (2000). *Value-added achievement results for three cohorts of Roots and Wings schools in Memphis: 1995–1999 outcomes.* Memphis, TN: University of Memphis, Center for Research in Educational Policy. Retrieved from http://citeseerx.ist.psu.edu/. The study is ineligible for review because it does not use a sample aligned with the protocol.

Simon, J. (2011). *A cost-effectiveness analysis of early literacy interventions* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3450763) The study is ineligible for review because it does not use an eligible design.

Slavin, R. E. (2006). *Translating research into widespread practice: The case of Success for All*. Baltimore, MD: Johns Hopkins University. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., Davis, S., & Center for Data-Driven Reform in Education. (2009). *Effective beginning reading programs: A best-evidence synthesis*. Baltimore, MD: Center for Data-Driven Reform in Education. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., & Madden, N. A. (1994, April). *Lee Conmigo: Effects of Success for All in bilingual first grades*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. The study is ineligible for review because it is out of the scope of the protocol.

Slavin, R. E., & Madden, N. A. (1998). *Success for All/Éxito Para Todos—Effects on the reading achievement of students acquiring English (Report No. 19)*. Baltimore, MD: Center for Research on the Education of Students Placed at Risk, Johns Hopkins University. Retrieved from ERIC: https://eric.ed.gov/?id=ED423327 The study is ineligible for review because it is out of the scope of the protocol.

Slavin, R. E., & Madden, N. A. (1999). Roots & Wings: A comprehensive approach to elementary school reform. In J. Block, S. Everson, & T. Guskey (Eds.), *Comprehensive school reform: A program perspective*. Dubuque, IA: Kendall/Hunt. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., & Madden, N. A. (2000). Roots & Wings: Effects of whole school reform on student achievement. *Journal of Education for Students Placed at Risk, 5*(1 & 2), 109–136. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., & Madden, N. A. (2004). Scaling up Success for All: Lessons for policy and practice. In T. K. Glennan, Jr., S. J. Bodilly, J. R. Galegher, & K. A. Kerr (Eds.), *Expanding the reach of education reforms: Perspectives from leaders in the scale-up of educational interventions* (pp. 135–174). Arlington, VA: RAND. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., & Madden, N. A. (2006). *Success for All/Roots & Wings: 2006 summary of research on achievement outcomes*. Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., & Madden, N. A. (2007). Scaling up Success for All: The first sixteen years. In B. Schneider & S. McDonald (Eds.), *Scale-up in education* (pp. 201–228). Lanham, MD: Rowman & Littlefield. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., & Madden, N. A. (2010). Success for All: Prevention and early intervention in school-wide reform. In J. Meece & J. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 434–445). New York, NY: Routledge. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., & Madden, N. A. (2012). *Success for All: Summary of research on achievement outcomes (revised)*. Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., Madden, N. A., & Chambers, B. (2008). *Success for All, embedded multimedia, and the teaching-learning orchestra*. Baltimore, MD: Johns Hopkins University. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., Madden, N. A., Chambers, B., & Haxby, B. (2009). *Two million children: Success for All.* Thousand Oaks, CA: Corwin Press. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., Madden, N. A., Cheung, A., & Liang, C. (2002). *Success for All in California: Gains on SAT-9 reading and the academic performance index.* Baltimore, MD: Success for All Foundation. The study is ineligible for review because it does not use a sample aligned with the protocol.

Slavin, R. E., Madden, N. A., Dolan, L. J., & Wasik, B. A. (1994). Roots and Wings: Inspiring academic excellence. *Strategies for Success, 52*(3), 10–13. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S. M., & Smith, L. J. (1994). Whenever and wherever we choose: The replication of Success for All. *Phi Delta Kappan, 75*(8), 639–647. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., Madden, N. A., Dolan, L. J., Wasik, B. A., Ross, S. M., Smith, L. J., & Dianda, M. (1996). Success for All: A summary of research. *Journal of Education for Students Placed at Risk, 1*(1), 41–76. The study is ineligible for review because it does not use an eligible design.

Slavin, R. E., Madden, N. A., Shaw, A. H., Mainzer, K. L., & Donnelly, M. C. (1993). Success for All: Three case studies of comprehensive restructuring of Urban elementary schools. In J. Murphy & P. Hallinger (Eds.), *Restructuring schools: Learning from ongoing efforts* (pp. 84–112). Newbury Park, CA: Corwin. The study is ineligible for review because it does not use an eligible design.

Smith, K. (2013). *Educators' perceptions of reading first and Success for All for mobile and nonmobile first through third grade students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3553382) The study is ineligible for review because it does not use an eligible design.

Sparks, S. D. (2013). Success for All yields early gains in first i3 evaluation. *Education Week, 33*(11), 8. The study is ineligible for review because it does not use an eligible design.

St. John, E. P., Manset, G., Chung, C., Simmons, A. B., Musoba, G. D., & Indiana University, Bloomington Education Policy Center. (2000). *Research-based reading interventions: The impact of Indiana's early literacy grant program*. Bloomington, IN: Indiana Education Policy Center, Smith Center for Research in Education. Retrieved from ERIC: https://eric.ed.gov/?id=ED447466 The study is ineligible for review because it does not use an eligible design.

St. John, E. P., Manset, G., Chung, C., Worthington, K., & Indiana University, Bloomington Education Policy Center. (2001). *Assessing the rationales for educational reforms: A test of the professional development, comprehensive reform, and direct instruction hypotheses*. Bloomington, IN: Indiana Education Policy Center, Smith Center for Research in Education. Retrieved from ERIC: https://eric.ed.gov/?id=ED458641 The study is ineligible for review because it is out of the scope of the protocol.

Stevens, R. J., Madden, N. A., Slavin, R. E., & Farnish, A. M. (1987). Cooperative integrated reading and composition: Two field experiments. *Reading Research Quarterly, 22*(4), 433–454. Retrieved from ERIC: https://eric.ed.gov/?id=ED291075 The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *A Lackland, Texas, elementary school thrives amid change*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Alliance School District in Ohio boost student achievement*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Bessemer City, Alabama, skips the frills, gets results*. Baltimore, MD: Author. Retrieved from http://www.successforall.org. The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Blanchester School District in Ohio sees decade of improvement*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Carson City, Nevada elementary schools continue to improve in just about everything that the state measures*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Choices in learning elementary charter school is a top performer in Seminole County*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Clove Campbell Senior Elementary School in Phoenix, Arizona, goes from good to great*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *David Crockett Elementary in San Antonio boosts results with Success for All*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Detroit Edison Public School Academy in Detroit, Michigan, is a beacon of light*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Edison Elementary School in Centralia, Washington, wins Title I academic achievement award.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Edward Hynes Charter School in New Orleans, Louisiana, hits new highs*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Edward Hynes takes Katrina devastation in stride*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Grasonville Elementary School in Maryland gained 59 points on Maryland state assessment in reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Great Bend, Kansas, continues to thrive.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Hawaii Success for All facilitator sees how Success for All fosters reading across cultural lines in Alaska.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Kentucky's McCracken County Public School District makes strides in reading scores.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.) *Knox County Public Schools in Barbourville, Kentucky, is a high-performing rural district.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Langley Park-McCormick Elementary sees big gains in first year of implementation*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Liberty Elementary in Tucson, Arizona, takes on i3 leadership role*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Liza Jackson Preparatory School is spearheading student achievement in Okaloosa County.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Makawao Elementary in Maui, Hawaii: Students engaged in the process leads to results in reading and beyond*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *McDermott Elementary School in Liberal, Kansas, wins state awards for effective Success for All implementation*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Oakman and Thomas Edison Elementary Schools thriving in Detroit Public Schools system*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Ohio's Princeton City School District gets beyond the classroom.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Reading scores at Woodlawn Hills Elementary, San Antonio, increase 23 percent.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Richards R-5 Elementary School in West Plains, Missouri, rallies to improve reading instruction.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Alabama gain on ARMT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Alaska gain on Alaska standards-based assessment reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Arkansas gain on Arkansas benchmark exams reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in California gain on CST reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Colorado gain on TCSAP reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Florida gain on FCAT-2 reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Georgia gain on CRCT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Hawaii gain on HSA reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Idaho gain on ISAT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Illinois gain on ISAT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Kansas gain on KSA reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Kentucky gain on KCCT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Louisiana gain on iLEAP 21 ELA.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Minnesota gain on Minnesota comprehensive assessments (MCA-II) reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Missouri gain on MAP reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Montana gain on CRT reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Nevada gain on CRT reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in New Mexico gain on standards-based assessment reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in North Carolina gain on end-of-grade reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in North Dakota gain on state reading assessment.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Ohio gain on Ohio achievement test reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Oklahoma gain on OCCT.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Oregon gain on OAKS reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Pennsylvania gain on PSSA reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Texas gain on TAKS reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Utah gain on CRT - English language arts.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Washington state gain on WASL/MSP reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All elementary schools in Wyoming gain on PAWS reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Alaska gain on Alaska standards-based assessment reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Arizona gain on AIMS reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Colorado gain on CSAP reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Florida gain on FCAT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Hawaii gain on HSA reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Illinois gain on ISAT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Louisiana gain on LEAP ELA.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Massachusetts gain on MCAS.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Minnesota gain on MCA-II reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Montana gain on CRT reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All middle schools in Pennsylvania gain on PSSA reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in Indiana gain on ISTEP+ English language arts reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in Kansas City gain on MAP - communication arts.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in Lawrence gain on MCAS*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in Michigan gain on MEAP - reading*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in Nebraska gain on assessment of state reading standards*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in New Jersey gain on New Jersey ASK - literacy language arts.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in Virginia gain on SOL - reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Success for All schools in Wisconsin gain on WKCE.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *The Alhambra Elementary School District in Phoenix, Arizona, boasts some of state's top performers*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All. (n.d.). *Wells Academy in Steubenville, Ohio, maintains years of 100% proficiency in reading.* Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Success for All Foundation. (2007). *Independent reviews of Success for All*. Baltimore, MD: Author. Retrieved from http://www.successforall.org/ The study is ineligible for review because it does not use an eligible design.

Wells, L. R. (2000, November). *An investigation of the Success for All reading program at two Mississippi elementary schools*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Bowling Green, KY. The study is ineligible for review because it does not use an eligible design.

Appendix A.1: Research details for Borman et al. (2007)

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal, 44*(3), 701–731.

Additional sources:

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., & Madden, N. A. (2004). *Success for All: Preliminary first-year results from the national randomized field trial.* Baltimore, MD: Success for All Foundation.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005a). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42*(4), 673–696. Retrieved from ERIC: https://eric.ed.gov/?id=ED485351

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2005b). Success for All: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27*(1), 1–22.

Hanselman, P., & Borman, G. D. (2013). The impacts of Success for All on reading achievement in grades 3–5: Does intervening during the later elementary grades produce the same benefits as intervening early? *Educational Evaluation and Policy Analysis, 35*(2), 237–251.

Slavin, R. E., Madden, N. A., Cheung, A., Chamberlain, A., Chambers, B., & Borman, G. (2005). A *randomized evaluation of Success for All: Second-year outcomes.* Baltimore, MD: Success for All Foundation.

Table A1. Summary of findings **Meets WWC group design standards without reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Alphabetics	35 schools/1,936 students	+13	Yes
Comprehension	41 schools/4,355 students	+5	No

Setting

The analysis sample included 41 elementary schools across 12 states located in rural and small towns in the South and urban areas of the Midwest.

Study sample

The study used a cluster randomized controlled trial design. The study piloted the *SFA*® program in fall 2001, when three schools were randomly assigned to *SFA*® and three schools were randomly assigned to the comparison condition. In fall 2002, 35 new schools were recruited, with 18 schools randomly assigned to implement *SFA*® in grades K–2 and 17 schools randomly assigned to implement *SFA*® in grades 3–5.

In Borman et al. (2007), the K–2 group had been the focus, with the 3–5 group providing the comparison. For the K–2 analyses, the study combined the two cohorts of schools and presented findings after the intervention students completed 1, 2, and 3 years of *SFA*®.

The authors used two samples to evaluate the effectiveness of the *SFA*® program: a sample that focused on students who were present in schools at the time of baseline and outcome assessments (referred to as the "longitudinal sample" in the study), and a sample that included all students who were given the outcome measure (referred to as the "combined longitudinal and in-mover sample" in the study). Both samples may include students who moved into the study schools after random assignment.

For the effectiveness rating, the WWC focused on third-year findings from the larger (combined) sample of students. Six schools were lost to attrition and reduced the third-year analytic sample to 35 schools. The third-year analyses focused on second-grade students who were in kindergarten when implementation began, and consisted of 1,011 students in 18 *SFA*® schools and 925 students in 17 comparison schools.

The 18 intervention schools were comprised of 61% minority students, and in the 17 comparison schools, 73% of students were minorities. The percentage of students eligible for free or reduced-price lunch was 66% in intervention schools, and 77% in comparison schools.

For the grade 3–5 analyses (Hanselman & Borman, 2013), the authors only used the fall 2002 cohort of schools, but flipped the comparison, using the K–2 group as an experimental control to estimate the effect of the *SFA*® literacy instruction in grades 3–5.

For the grade 3–5 analyses, the study included two cohorts of students, referred to as "primary" and "secondary" in the study. Students in the primary cohort began using the *SFA*® reading programs in grade 3, while students in the secondary cohort began using the *SFA*® reading programs in grade 4.

This report focuses on the primary cohort of students who were in third grade in 2002–03 and experienced the program over 1 year of the study. Their reading achievement outcomes were measured in the spring of the third grade. The analytic sample included 1,197 students in 17 *SFA*® schools and 1,223 students in 18 comparison schools. Some students in the analytic sample moved into the schools between random assignment and the posttest.

At baseline in the fall of 2002, the percentage of minority students in 17 intervention schools was 83%, while the percentage of minority students in the 18 comparison schools was 75%. The percentage of students eligible for free or reduced-price lunch was 86% in intervention schools and 75% in comparison schools.

Intervention group

Students in intervention group received the *SFA*® whole-school reform program, including the *SFA*® reading curriculum, tutoring for students' quarterly assessments, family support teams for students' parents, a facilitator who worked with school personnel, and training for all intervention teachers. Students were regrouped from across grade levels into reading classes based on their reading level. Classroom instruction was structured around direct instruction, cooperative work in small groups, and regular individual assessments. Some schools took a year to fully implement the program.

For intervention schools that implemented *SFA*® in grades 3–5, students received *Reading Wings*, the *SFA*® reading curriculum for elementary students at the second-grade level and above. The curricular focus throughout lessons was on comprehension of complex text. No intervention students had prior exposure to the K–2 *SFA*® curriculum.

Comparison group

For the grades K–2 analyses, comparison schools continued using their regular curriculum for grades K–2. In the second cohort of schools recruited in 2002, *SFA*® was implemented in grades 3–5 in comparison schools (in comparison schools recruited in 2001, no grade levels implemented *SFA*®). Authors conducted observations at all schools and indicated that students in grades K–2 were not exposed to classroom-level components of *SFA*® in schools that implemented the intervention in grades 3–5. However, K–2 students in these comparison schools may have had access to some schoolwide components of the grades 3–5 *SFA*® intervention, such as family support. If comparison students in grades K–2 used these services, the study's estimate of the effectiveness of *SFA*® may not reflect the full impact of the schoolwide components of *SFA*® on outcomes.

For the grades 3–5 analyses, no information is provided on the instruction in grades 3–5 used in the comparison condition. No comparison students had prior exposure to the K–2 *SFA*® curriculum. While the *SFA*® school reform program was concurrently implemented in grades K–2 in the comparison schools, *SFA*® monitored the intervention and comparison classrooms during quarterly visits and found no evidence that the comparison classrooms in grades 3–5 had adopted any of the *SFA*® components.

Outcomes and measurement

For the grades K–2 analyses, outcomes were measured at the conclusion of kindergarten, grade 1, and grade 2, respectively; a pretest was administered in the fall of kindergarten. The findings that contribute to the effectiveness rating are based on those measured at the end of second grade and reflect 3 years of exposure to the *SFA*® intervention for the majority of students. Some students in the analytic sample who moved into study schools after implementation began received less than the full 3 years of exposure.

Three subtests of the WRMT—Word Identification, Word Attack, and Passage Comprehension—were administered at the end of each school year. The WRMT Letter Identification subtest was administered in the spring of grade 1. The WWC reviewed WRMT Word Identification, Word Attack, and Letter Identification under the alphabetics domain. The WRMT Passage Comprehension subtest falls under the comprehension domain.

For the grades 3–5 analyses, the study measured outcomes using the Gates-MacGinitie Reading Test (GMRT, 4th Edition, Levels 3–4, Form S) in spring 2003. The GMRT outcome was reviewed in the comprehension domain. For a more detailed description of these outcome measures, see Appendix B.

For the grades K–2 analyses, supplemental schoolwide findings are presented for first graders after 2 years of exposure to *SFA*® and for subgroups of third graders by reading level after 1 year of exposure to *SFA*®. For the grades 3–5 analyses, supplemental schoolwide findings are presented for subgroups of third graders by reading level (at grade level or below grade level as measured on the GRMT pretest assessment in the fall of 2002) after 1 year of exposure to *SFA*®. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

The grades 3–5 analyses of the reading outcomes from the study's secondary cohort at Years 1 and 2, as well as from the study primary cohort at Years 2 and 3, are not eligible for review because they do not use a sample aligned with the Beginning Reading review protocol, version 3.0.

Support for implementation

SFA® teachers received 3 days of training during the summer and approximately 8 days of on-site follow-up during the first implementation year. Success for All Foundation trainers visited classrooms, met with groups of teachers, looked at data on children's progress, and provided feedback to school staff on implementation quality and outcomes.

Appendix A.2: Research details for Quint et al. (2015)

Quint, J. C., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015). *Scaling up the Success for All model of school reform: Final report from the Investing in Innovation (i3) Scale-Up.* New York, NY: MDRC.

Additional sources:

Quint, J., Zhu, P., Doolittle, F., & Society for Research on Educational Effectiveness. (2012). *Understanding variation in implementation of SFA in the i3 Scale-Up project*. Washington, DC: Society for Research on Educational Effectiveness. Retrieved from ERIC: https://eric.ed.gov/?id=ED530361

Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, J. T., & Zhu, P. (2013). *The Success for All model of school reform: Early findings from the Investing in Innovation (i3) Scale-Up.* New York, NY: MDRC. Retrieved from ERIC: https://eric.ed.gov/?id=ED545452

Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, J. T., & Zhu, P. (2014). *The Success for All model of school reform: Interim findings from the Investing in Innovation (i3) Scale-Up.* New York, NY: MDRC. Retrieved from ERIC: https://eric.ed.gov/?id=ED546642

Table A2. Summary of findings **Meets WWC group design standards without reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Alphabetics	37 schools/ 2,907 students	+4	Yes
Comprehension	37 schools/ 2,894 students	+1	No

Setting

The study took place in five districts in four states in the western, southern, and northeastern United States. Most districts were located in mid-size to large cities.

Study sample

The study used a cluster randomized controlled trial design. Thirty-seven schools that met the study eligibility criteria were randomly assigned to intervention or comparison groups in spring 2011 after blocking by school district. To be eligible to participate in the study, schools were required to serve grades K–5, have at least 40% of their students eligible for free or reduced-price lunch, and be willing to participate in the study and support program implementation. The program was implemented for all students in the schools starting in fall 2011.

The authors used three samples to evaluate the effectiveness of *SFA*®, which they refer to as the main sample, the spring sample, and the auxiliary sample. The main sample focused on students who were present in schools at the time of baseline and outcome assessments. The spring sample included all students who had at least one valid score on the end-of-year outcomes. The auxiliary sample consisted of students who were present in grades 3, 4, or 5

in the study schools during program implementation years. All three samples may include students who moved into the study schools after random assignment.

For the effectiveness ratings, the WWC focused on third-year findings from the sample of students who had at least one valid score on the end-of-year outcomes (referred to as the spring sample in the study). The third-year analyses focused on second-grade students who were in kindergarten when implementation began. This cohort included 1,557 students in 19 *SFA*® schools and 1,350 students in 18 comparison schools.

Across all study schools, 57% of the population received free or reduced-price lunch, 62% of students were Hispanic, 23% were Black, and 14% were White. Males made up 52% of the overall school sample.

Intervention group

Intervention students received features of the full *SFA*® program, including the *SFA*® reading curriculum that is the focus of this intervention report, tutoring for students in grades 1–3, a facilitator who worked with school personnel, and training for all intervention teachers. Some other features of the full *SFA*® program, such as regular tutoring for struggling students, periodic testing and regrouping, and support for families, were not provided to all students in all schools. The study relied on local district coaches rather than coaches employed by *SFA*®. The *SFA*® model calls for a 90-minute reading block each day, and most schools adhered to this. Schools began using the program for the first time at the beginning of the first study year, and in general improved their implementation over the course of the study based on the authors' monitoring.

Comparison group

The comparison condition included schools that implemented standard reading programs from publishers such as *Macmillan/McGraw-Hill, Houghton Mifflin Harcourt*, and *Scott Foresman*. During the 3-year study period, most comparison schools continued to use the same curriculum, while others switched from one common program to another.

Outcomes and measurement

Outcomes were measured at three points in time: in spring 2012, spring 2013, and spring 2014. Findings collected in spring 2014 reflect 3 years of exposure to the *SFA*® intervention for the majority of students at the end of second grade. Because the analytic sample includes students who moved into the study schools after random assignment, second graders had received varying amounts of the *SFA*® intervention, ranging from less than 1 year to 3 years. However, the majority of the students (about 63%) were in the study for all 3 years. Students were assessed using the Letter-Word Identification, Word-Attack, and Passage Comprehension subtests of the Woodcock-Johnson (WJ) Reading Test and the Test of Word Reading Efficiency (TOWRE). The WWC reviewed WJ Word Identification, WJ Word Attack and TOWRE under the alphabetics domain. WJ Passage Comprehension was reviewed under the comprehension domain. For a more detailed description of these outcome measures, see Appendix B.

Supplemental findings are presented for first graders after 2 years of exposure to the *SFA*® intervention, and for kindergarteners after 1 year of exposure to *SFA*®. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Results for the reading outcomes from the other study samples (referred to as the main and auxiliary samples in the study), as well as subgroup analyses, do not meet WWC group design standards. These samples were not shown to be equivalent at baseline across the intervention and comparison groups and, therefore, are not included in this review.[17]

Support for implementation

Each school implementing *SFA*® appointed a facilitator who oversaw the implementation of the program. Principals and other school leaders attended a week-long conference the summer before implementation, in which they were introduced to the various parts of the programs. *SFA*® coordinators visited the schools for 4 days before the beginning of the school year. One day of programming focused on principals and school leaders, the second day on all teachers, and the third and fourth days on reading teachers. During the school year, *SFA*® coaches visited the schools implementing the program to provide additional support. This was focused primarily on assisting principals and other leaders in implementing program features, but also included classroom visits and feedback on lessons.

Appendix A.3: Research details for Madden et al. (1993)

Madden, N. A., Slavin, R. E., Karweit, N., Dolan, L., & Wasik, B. A. (1993). Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal, 30*(1), 123–148.

Additional sources:

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24*(4), 243–266.

Madden, N. A., Slavin, R. E., Karweit, N., Dolan, L., & Wasik, B. A. (1991). *Success for All: Multi-year effects of a schoolwide elementary restructuring program*. Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Slavin, R. E., Madden, N. A., Dolan, L. J., & Wasik, B. A. (1993). *Success for All in the Baltimore City Public Schools: Year 6 report*. Baltimore, MD: Johns Hopkins University, Center for Research in Effective Schooling for Disadvantaged Students.

Slavin, R. E., Madden, N. A., Karweit, N. L., Dolan, L., & Wasik, B. A. (1990). *Success for All: Second year report.* Baltimore, MD: Baltimore Public Education Institute and Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Slavin, R. E., Madden, N. A., Karweit, N., Dolan, L., & Wasik, B. A. (1993). *Success for All in the Baltimore City Public Schools: Year 5 report.* Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

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
Alphabetics	10 schools/1,342 students	+22	Yes
Reading fluency	10 schools/306 students	+18	No
Comprehension	10 schools/730 students	+19	Yes
General reading achievement	10 schools/1,157 students	+14	No

Setting

The analysis sample included 10 elementary schools in Baltimore, Maryland.

Study sample

This study examined the effects of *SFA*® in the Baltimore City public elementary schools by contrasting eight intervention schools with six comparison schools. Each comparison school was matched with an intervention school based on the percentage of students receiving free or reduced-price lunch and prior achievement level. Students were then individually matched based on a standardized test administered by the school district. The study investigated the effects of three versions of the *SFA*® program: full implementation, dropout prevention, and curriculum only.

SFA® schools introduced the reading program during the 1988–89 school year. Over the course of 5 years, the study tracked outcomes for students enrolled in grades pre-K–4. This report emphasizes findings from three cohorts of students who started *SFA*® in prekindergarten (Cohort 1), kindergarten (Cohort 2), and first grade (Cohort 3). To determine the effectiveness ratings, the WWC focused on results measured after the highest exposure to *SFA*® among the analytic samples that were found to be equivalent at baseline and met WWC group design standards. In particular, this report includes findings for students after 3 years of exposure to *SFA*® in the alphabetics domain, and up to 5 years of exposure in other outcome domains. The number of students included in the analytic samples that contribute to the effectiveness rating varied by cohort, outcome domain, and period of exposure to the intervention:

Cohort 1: 246 students in *SFA*® schools and 246 students in comparison schools were followed from prekindergarten to first grade in the alphabetics and general reading achievement domains, and 48 *SFA*® and 56 comparison students were followed to second grade in the comprehension domain;

Cohort 2: 220 students in *SFA*® schools and 220 students in comparison schools were followed from kindergarten to second grade in the alphabetics domain, and 151 *SFA*® and 156 comparison students were followed to fourth grade in the reading fluency, comprehension, and general reading achievement domains; and

Cohort 3: 205 students in *SFA*® schools and 205 students in comparison schools were followed from first grade to third grade in the alphabetics and general reading achievement domains, and 160 *SFA*® and 160 comparison students were followed to second grade in the comprehension domain.

The largest combined analytic sample across cohorts that contributed findings to the effectiveness rating in an outcome domain included 671 students in five *SFA*® schools and 671 students in five comparison schools.

The five *SFA*® schools served between 97–100% of African-American students, and 83–98% of students qualified for free or reduced-price lunch. In comparison schools, at least 75% of students qualified for free or reduced-price lunch. The comparison schools received funding under federal programs for low-achieving disadvantaged students.

Intervention group

The study included two variants of the *SFA*® program, which the study authors referred to as full implementation (two schools) and dropout prevention (three schools).[18] Intervention students in the full implementation version received the typical *SFA*® program, including the *SFA*® reading curriculum, tutoring for students in grades 1–3, quarterly assessments, family support teams for students' parents, a full-time facilitator who worked with school personnel, and training for all intervention teachers. Intervention schools in the dropout prevention version had a half-time

facilitator and a reduced number of tutors and family support staff. Chapter I funds supported a dropout prevention program. Although the two program variants provided different schoolwide components, the components of the *SFA*® reading curricula were similar, with each school receiving the same training, coaching support, and materials.

Comparison group

The comparison condition included schools that implemented a traditional reading program built around the *Macmillan Connections* basal series. Comparison schools largely used their Chapter I funds to reduce first- through third-grade class sizes and to provide low-achieving students with traditional group-based pullout services.

Outcomes and measurement

Outcomes were measured at five points in time: spring 1989, spring 1990, spring 1991, spring 1992, and spring 1993. Primary findings in the alphabetics, comprehension, and general reading achievement domains, collected in spring 1991, reflect 3 years of exposure to the *SFA*® intervention for students across different cohorts/grades. Primary findings in the reading fluency and comprehension domains, collected in spring 1993, reflect 5 years of exposure to the *SFA*® intervention for Cohort 2 students in grade 4.

The following assessments were administered in the study over years: the California Achievement Test (CAT), the CTBS, the DARD, the GORT, the Woodcock Language Proficiency Battery (WLPB), and the WRMT. The WWC reviewed the WLPB Letter-Word Identification and Word Attack subtests under the alphabetics domain. The GORT Passage subtest falls under the reading fluency domain. The WRMT Passage Comprehension, DARD Silent Reading, CAT Reading, and CTBS scores on Total Reading, Reading Comprehension, and Reading Vocabulary, all fall under the comprehension domain. DARD Oral Reading and CTBS Total Language were reviewed in the general reading achievement domain.

The schools were matched on CAT scores from spring 1987 or fall 1988.[19] The CAT pretest was administered in 1988 and 1989, and the CTBS was administered in 1990. Pretests were administered in the spring of students' kindergarten year by district. For a more detailed description of these outcome measures, see Appendix B.

Supplemental findings are presented for (1) the full student samples after 1 and 2 years of exposure to the *SFA*® intervention, (2) after 5 years of *SFA*® exposure for Cohort 2 students on two subtests of the CTBS (reading comprehension and reading vocabulary), and (3) for subgroups of low-achieving students (that is, students scoring in the lowest 25% on a standardized test of reading achievement) with different levels of intervention implementation (from 1 to 4 years). These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

The study also examined student performance on the following outcomes: the Test of Language Development (picture vocabulary and sentence imitation scales), the Merrill Language Screening Test, and the Maryland School Performance Assessment Program. However, the corresponding analysis samples were not shown to be equivalent at baseline across the intervention and comparison groups and, therefore, are not included in this review. Grade retention (the number of students retained each year) and school attendance (yearly attendance rates) were also collected from school records (presented in Madden et al., 1993) but are not eligible under the Beginning Reading review protocol, version 3.0.

Support for implementation

The teachers and tutors were regular certified teachers. They received detailed teacher's manuals supplemented by 2 to 3 days of in-service training at the beginning of the school year. For teachers of grades 1–3 and for reading tutors, these training sessions focused on the implementation of the reading program. Preschool and kindergarten teacher's and teachers aides were trained in the use of the thematic units and other aspects of the preschool and kindergarten models. School facilitators also organized information sessions to allow teachers to share problems and solutions, suggest changes, and discuss the progress of individual children.

Appendix A.4: Research details for Ross et al. (1998)

Ross, S. M., Alberg, M., McNelis, M., & Rakow, J. (1998). *Evaluation of elementary school school-wide programs: Clover Park School District year 2: 1997–98*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Additional source:

Ross, S. M., Alberg, M., & McNelis, M. (1997). *Evaluation of elementary school school-wide programs: Clover Park School District, year 1: 1996–97*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Table A4. Summary of findings

Meets WWC group design standards with reservations

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Alphabetics	5 schools/128 students	–2	No
Comprehension	5 schools/128 students	–11	No
General reading achievement	5 schools/128 students	–4	No

Setting

The study was conducted in 19 schools in Clover Park, Washington.

Study sample

The study compared whole-school improvement programs, including *SFA*®, *Accelerated Schools*, and locally developed programs, in 19 schools for students in grades 1–2. Schools were divided into four groups based on their similarity on several school characteristics, including enrollment, percentage of minority students, percentage of students eligible for free or reduced-price lunch, and initial academic performance. Only one group (referred to as "cluster 2A" by the study authors), which was the third highest with respect to socioeconomic status, meets WWC group design standards. This group included three *SFA*® schools and two *Accelerated Schools*.[21] The percentage of minority students in the three intervention schools was between 47% and 63%. In the comparison schools, the percentage of minority students ranged from 42% to 54%. The percentage of students eligible for free or reduced-price lunch varied from 63% to 66% in intervention schools, and from 66% to 71% in comparison schools. For the effectiveness ratings, the WWC focused on findings from the sample of 128 second graders, who completed 2 years of the program. After 2 years, three *SFA*® schools with 86 students and two *Accelerated Schools* with 42 students remained in the analytic sample.

Intervention group

Intervention students received the typical *SFA*® program, including the *SFA*® reading curriculum, tutoring for students in grades 1–2, quarterly assessments, family support teams for students' parents, a facilitator who worked with school personnel, and training for all intervention teachers.

Comparison group

Accelerated Schools is a comprehensive school reform program that is designed to close the achievement gap between at-risk and not-at-risk children. The program redesigns and integrates curricular, instructional, and organizational practices to improve the achievement of at-risk students.

Outcomes and measurement

Outcomes were measured at two points in time: spring 1997 and spring 1998, and the pretest was administered in fall 1996 when study students were in grade 1. Primary findings, collected in May 1998, reflect 2 years of exposure to the *SFA*® intervention for students at the end of second grade. The DARD Oral Reading subtest and three subtests of the WRMT—Word Identification, Word Attack, and Passage Comprehension—were administered at the end of each school year. The WWC reviewed WRMT Word Identification and Word Attack under the alphabetics domain. WRMT Passage Comprehension was reviewed in the comprehension domain, and DARD Oral Reading was reviewed in the general reading achievement domain. For a more detailed description of these outcome measures, see Appendix B.

The Peacock Picture Vocabulary Test (PPVT) was administered as the pretest to participating first graders. The authors also used a writing measure in the study; however, this writing test was outside of the scope of the Beginning Reading review protocol.

Supplemental findings are presented for first graders after 1 year of exposure to the *SFA*® intervention. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Support for implementation

No information on training for the specific teachers in this study was provided.

Appendix A.5: Research details for Ross and Casey (1998a)

Ross, S. M., & Casey, J. (1998a). *Longitudinal study of student literacy achievement in different Title I school-wide programs in Fort Wayne Community Schools Year 2: First grade results.* Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Table A5. Summary of findings

Meets WWC group design standards with reservations

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Alphabetics	7 schools/288 students	+6	No
Comprehension	7 schools/288 students	+3	No
General reading achievement	7 schools/288 students	+5	No

Setting

The study took place in seven Title I elementary schools in Fort Wayne, Indiana.

Study sample

This study examined the effects of *SFA*® in two Title I schools. Five Title I schools that were implementing locally developed schoolwide programs were used as a comparison group. The study was conducted in fall 1996 through spring 1998 and reports on first-grade outcomes of students who were in kindergarten at the start of the study. The analysis sample included 288 students: 83 students in the *SFA*® schools and 205 students in comparison schools. The student-level analysis sample demonstrated equivalence on the PPVT. School populations ranged between 31% and 50% minority students; between 62% and 81% of students received free or reduced-price lunch. The study also reported on an additional intervention school that supplemented *SFA*® with another branded intervention (*Reading Recovery*), but results from this portion of the study are ineligible for review.

Intervention group

Intervention students received the typical *SFA*® curriculum, including the *Reading Roots* reading curriculum in grade 1 and the *Reading Wings* reading curriculum in grade 2, one-to-one tutoring for the lowest-achieving students by certified teacher tutors, quarterly assessments, family support teams for students' parents, a facilitator who worked with school personnel, and training for all intervention teachers.

Comparison group

The five comparison schools implemented locally developed schoolwide programs. The schools were comparable with *SFA*® schools on pretest PPVT measures, free or reduced-price lunch status, and ethnicity. Four out of the five local school programs incorporated components of other branded programs, including *Reading Recovery*, *Accelerated Reader*, *Four-Block*, and *STAR*. These curricula place considerable emphasis on reading, use of basal readers, and multifaceted reading activities.

Outcomes and measurement

Primary findings reflect 2 years of exposure to the intervention for students in first grade. Three subtests of the WRMT were administered: Word Identification, Word Attack, and Passage Comprehension. Word Identification and Word Attack were reviewed in the alphabetics domain, while Passage Comprehension was reviewed in the comprehension domain. Outcomes in the general reading achievement domain were measured using the DARD Oral Reading subtest. The study also administered the PPVT to students in the fall of kindergarten as the pretest measure. For a more detailed description of these measures, see Appendix B.

Supplemental findings are presented for low-achieving students in grade 1 (that is, lowest 25% on a standardized test of reading achievement). These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Support for implementation

A full-time facilitator worked with staff to ensure fidelity of implementation in the intervention schools. No information on training for the specific teachers was provided in this study.

Appendix A.6: Research details for Ross and Casey (1998b)

Ross, S. M., & Casey, J. (1998b). *Success For All evaluation: 1997–1998 Tigard-Tualatin School District.* Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Table A6. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Alphabetics	4 schools/265 students	+9	No
Comprehension	4 schools/265 students	0	No
General reading achievement	4 schools/265 students	+5	No

Setting

The study was conducted in four elementary schools located in the Tigard-Tualatin School District in Oregon.

Study sample

This study examined the effects of *SFA*® in two elementary schools in the Tigard-Tualatin School District and used two elementary schools in the same school district as a comparison group. The study took place over 1 school year (1997–98) and included kindergarten and first-grade students. Students at the same grade levels in four schools were described as demographically similar.

The WWC based its effectiveness rating on the kindergarten sample because comparisons of first graders did not satisfy the baseline equivalence requirement and therefore did not meet WWC group design standards. The analytic sample included 156 kindergarten students in the *SFA*® group and 109 kindergarten students in the comparison group.

The schools in the intervention and comparison groups had low proportions of minority students, as well as low proportions of students receiving free or reduced-price lunch. All study schools had between 12% and 17% minority enrollment, contained less than 1,000 students, and between 11% and 21% of students received free or reduced-price lunches.

Intervention group

No description of *SFA*® as implemented in the study is provided in the text.

Comparison group

The comparison group received the district's standard reading program for kindergarten. No other information was provided on the comparison curriculum.

Outcomes and measurement

Primary findings reflect 1 year of exposure to the intervention for students in kindergarten. Three subtests of the WRMT were administered: Word Identification, Word Attack, and Passage Comprehension. Word Identification and Word Attack were reviewed in the alphabetics domain, while Passage Comprehension was reviewed in the comprehension domain. Outcomes in the general reading achievement domain were measured using the DARD Oral Reading subtest. The study used the PPVT as the pretest measure. For a more detailed description of these outcome measures, see Appendix B.

For the above outcomes, the findings were also presented as dichotomous test scores (with "0" indicating no correct responses and "1" indicating at least one correct response). These outcome measures are not featured in this WWC report because they do not contribute unique information about the intervention's effectiveness that is not also captured in reported findings based on the established scales from the standardized tests.

Supplemental findings are presented for low-achieving students in kindergarten (that is, students with the lowest 25% of scores on a standardized test of reading achievement). These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Support for implementation

No information on training for the specific teachers in this study was provided

Appendix A.7: Research details for Ross et al. (1995)

Ross, S. M., Smith, L. J., & Casey, J. (1995). *Final report: 1994–1995 Success for All program in Fort Wayne, Indiana*. Memphis: TN: University of Memphis, Center for Research in Educational Policy.

Additional sources:

Casey, J., Smith, L. J., & Ross, S. M. (1994). *Final report: 1993–1994 Success for All program in Fort Wayne, Indiana*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., & Casey, J. (1997). Preventing early school failure: Impacts of Success for All on standardized test outcomes, minority group performance, and school effectiveness. *Journal of Education for Students Placed at Risk, 2*(1), 29–53.

Ross, S. M., Smith, L. J., & Casey, J. (1999). "Bridging the gap": The effects of the Success for All program on elementary school reading achievement as a function of student ethnicity and ability level. *School Effectiveness and School Improvement, 10*(2), 129–150

Ross, S. M., Smith, L. J., Casey, J., & Johnson, B. (1993). *Final report: 1992–93 Success for All program in Ft. Wayne, Indiana.* Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., Casey, J., Johnson, B., & Bond, C. (1994, April). Using *"Success For All" to restructure elementary schools: A tale of four cities.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Smith, L. J., Ross, S. M., & Casey, J. (1996) Multi-site comparison of the effects of Success for All on reading achievement. *Journal of Literacy Research, 28*(3), 329–353.

Smith, L. J., Ross, S. M., Faulks, A., Casey, J., Shapiro, M., & Johnson, B. (1993). *1991–1992 Ft. Wayne, Indiana Success for All results*. Memphis, TN: University of Memphis, Center for Research in Educational Policy.

Table A7. Summary of findings **Meets WWC group design standards with reservations**

Outcome domain	Sample size	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Alphabetics	4 schools/205 students	+8	No
Comprehension	4 schools/205 students	+1	No
General reading achievement	4 schools/205 students	–5	No

Setting

The study took place in four elementary schools located in the same district in Fort Wayne, Indiana.

Study sample

This study included students who were enrolled at two *SFA*® schools and two comparison schools. Comparison schools were matched to the intervention schools based on poverty level, prior achievement level, and ethnicity; pairs of students were then matched on PPVT pretest scores.

The study included three cohorts of students, and intervention students in each cohort received *SFA*® for up to 4 years. The WWC based its effectiveness rating on spring 1995 findings (after 3 or 4 years of exposure) from 205 students in the three analytic samples that were found to be equivalent at baseline:

Cohort 1: 54 students in the *SFA*® group and 20 students in the comparison group—these students began using the reading program in the 1991–92 school year and were followed from kindergarten to third grade;

Cohort 2: 45 students in the *SFA*® group and 32 students in the comparison group—these students began using the reading program in the 1991–92 school year and were followed from first to fourth grade; and

Cohort 3: 29 students in the *SFA*® group and 25 students in the comparison group—these students began using the reading program in the 1992–93 school year and were followed from kindergarten to second grade. The analytic sample for Cohort 3 that the WWC used for the intervention's effectiveness rating included only ethnic minority students (comprised largely of African-American students). Results for the full sample of Cohort 3 students are not included in this report because the intervention and comparison group students in that sample were not equivalent on key characteristics at baseline.

The percentage of Caucasian students in the four study schools was between 40% and 68%. The percentage of African-American students ranged from 27% to 45%. The percentage of Hispanic students ranged from 8% to 9%.

Intervention group

Intervention students received the typical *SFA*® program, including the *SFA*® reading curriculum, tutoring for students, quarterly assessments, family support teams for students' parents, a facilitator who worked with school personnel, and training for all intervention teachers. Students were grouped into cross-grade reading groups based on reading level. These groups met for 90 minutes a day and used the *Reading Roots* and *Reading Wings* curricula. Students who were struggling to keep up with their reading group were provided with one-on-one tutoring, and students were regrouped on a regular basis. *SFA*® was coordinated at the school level by a full- or part-time program coordinator.

Comparison group

Comparison schools continued using their regular curriculum. One school used a reading program based on basal readers, with a strong focus on phonics. The other placed some emphasis on phonics and whole-language instruction, and introduced individual tutoring and regrouping in the later years of the study.

Outcomes and measurement

Outcomes were measured in spring 1992, spring 1993, spring 1994, and spring 1995; the pretest was administered in fall 1991. Primary findings, collected in spring 1995, reflect 4 years of exposure to the *SFA*® intervention for students in grades 3 and 4 and 3 years of *SFA*® exposure for minority students in grade 2. Three subtests of the WRMT were administered: Word Identification, Word Attack, and Passage Comprehension. Word Identification and Word Attack were reviewed by the WWC in the alphabetics domain, while Passage Comprehension was reviewed in the comprehension domain. Outcomes in the general reading achievement domain were measured using the DARD Oral Reading subtest (in grades 2–3) and the GORT (in grades 4). The study used the PPVT as the pretest measure.

Supplemental findings are presented for (1) fourth-grade students in Cohort 2 that scored below 25% on the pretest, (2) minority students in grades 2–4 from Cohorts 1 through 3, and (3) nonminority students in grades 3–4 from Cohorts 1 and 2. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Support for implementation

Teachers in their first year of teaching *SFA*® classes received 3 days of summer training and 2–4 additional in-service days during the school year. A school facilitator monitored and provided feedback throughout the year. Twice a year, trainers provided by the developer visited and observed teachers. After the first year, training was reinforced by regular in-service training, an annual *SFA*® conference, and implementation checks for the facilitators and trainers.

Appendix A.8: Research details for Skindrud and Gersten (2006)

Skindrud, K., & Gersten, R. (2006). An evaluation of two contrasting approaches for improving reading achievement in a large urban district. *Elementary School Journal, 106*(5), 389–407.

Table A8. Summary of findings **Meets WWC group design standards with reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
General reading achievement	12 schools/531 students	–7	No

Setting

The study was conducted in 12 schools in the Sacramento City Unified School District (SCUSD), a large urban district in northern California.

Study sample

Under California's interpretation of *Reading First*, all 59 elementary schools in SCUSD were required to implement one of two models of reading instruction, *SFA*® or *Open Court Reading*©. In the fall of 1997, four schools implemented *SFA*®. A matched sample of *Open Court Reading*© schools were created by rank-ordering SCUSD schools by poverty level (measured by the percentage of students eligible for free or reduced-price meals and percentage of students on Aid to Families with Dependent Children), and selecting two comparison schools for each *SFA*® school—those ranked just above and just below each *SFA*® school. The study included two cohorts of students: students in Cohort 1 began using the reading programs in grade 2, while students in Cohort 2 started in grade 3. A total of 936 students in Cohort 1 and Cohort 2 participated in the study. The WWC based its effectiveness rating on findings from 531 students from the two analytic samples that were found to be equivalent at baseline:

Cohort 1: 142 students in the *SFA*® group and 292 students in the comparison group—these students were followed from second to third grade; and

Cohort 2: 36 students in the *SFA*® group and 61 students in the comparison group—these students were followed through third grade. The analytical sample for Cohort 2 includes only low-achieving students (that is, lowest 25% on a standardized test of reading achievement). Results for the full sample of Cohort 2 students are not included in this report because, based on information obtained from the authors, that sample of students was not equivalent on key characteristics at baseline.

Intervention group

Students in the intervention group received reading instruction through *SFA*®. Students were put into homogeneous groups, across classrooms and grades, based on reading skills. They received 90 minutes of reading instruction daily, outside of their homerooms. *SFA*® also prescribes additional writing instruction outside of these groups. The *SFA*® training consultants monitored implementation fidelity and observed additional writing instruction in all study schools during both study years. The authors noted that teachers in *SFA*® schools frequently included additional spelling and grammar, along with writing instruction, outside of the 90-minute reading block. *SFA*® prescribes a core reading curriculum only in grades K–1; in grades 2–6, the schools can choose their own reading curricula. The authors state that the materials and guidelines for instruction (*Reading Root*s for grade 1 and *Reading Wings* for grades 2–4), as well as the professional development, tutoring, and the *SFA*® school facilitator and regional consultant oversight procedures, all followed those outlined by the developers of the curriculum.

Comparison group

Students in the comparison group received reading instruction using *Open Court Reading*©, a systematic approach to teaching alphabetics, print knowledge, and phonemic awareness. For this study, the district used the 1996 version of the curriculum, *Open Court Collections for Young Scholars*. Two hours of daily whole-class reading instruction was followed by 30 minutes of small-group instruction and/or independent work. All study students received a condensed selection of instructional content to "catch-up" students to *Open Court Reading*© content that they had not received in prior years (since they began using the curriculum in either second or third grade).

Outcomes and measurement

Outcomes were measured in spring 1998 and spring 1999; the pretest was administered in fall 1997. Primary findings reflect 2 years of exposure to the *SFA*® intervention for students in Cohort 1 (collected in spring 1999), and 1 year of *SFA*® exposure for low-achieving students in Cohort 2 (spring 1998). Two subtests of the SAT-9 were administered: Reading and Language. The WWC reviewed these outcomes under the general reading achievement domain. The study also used two subtests of the Iowa Test of Basic Skills, Reading and Language, as the pretest measures. The authors converted all measures to Normal Curve Equivalent scores. For a more detailed description of these outcome measures, see Appendix B.

Supplemental findings are presented for second graders from Cohort 1 after 1 year of exposure to the *SFA*® intervention, and subsamples of low-achieving students (that is, lowest 25% on a standardized test of reading achievement) after 1 and 2 years of exposure to *SFA*® for Cohort 1.[22] These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Support for implementation

At *SFA*® schools, training and technical assistance were provided by *SFA*® consultants from a regional *SFA*® office. The *SFA*® consultants assessed implementation fidelity and rated it as a typical level of implementation when compared with national implementation averages.

At *Open Court Reading*© schools, teachers received 4 days of basic grade-level training in Year 1, followed by 4 days of advanced grade-level training in Year 2. Each *Open Court Reading*© school received a reading coach (either full-time or part-time, depending on school size). Curriculum experts met monthly with reading coaches and administrators to refine instruction and supervision and to solve problems. Reading coaches collected implementation information but were prohibited from sharing the information with the study authors; the district-level reading coordinator indicated that although some schools had implementation problems at the beginning of the study, these were resolved by the second study year.

Appendix A.9: Research details for Tracey et al. (2014)

Tracey, L., Chambers, B., Slavin, R. E., Madden, N. A., Cheung, A., & Hanley, P. (2014). *Success for All in England: Results from the third year of a national evaluation. SAGE Open, 4*(3), 1–10.

Table A9. Summary of findings **Meets WWC group design standards with reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Alphabetics	35 schools/886 students	+8	Yes
Reading fluency	35 schools/880 students	+5	No
Comprehension	35 schools/868 students	+2	No

Setting

The study was conducted in 35 schools in England.

Study sample

Schools were recruited in spring 2008 to participate in the study, which began in fall 2008 at the start of the 2008–09 school year. All 20 intervention schools were already implementing *SFA*®. Once 20 *SFA*® schools were recruited, recruitment began for comparison schools with similar demographic and achievement characteristics; matching criteria included school-level achievement, percentage of students eligible for free school meals, and the percentage of students with English as an Additional Language (EAL). The percentage of students with EAL in 20 intervention schools was 45%, and in 20 comparison schools it was 22%. The percentage of students eligible for free school meals was 44% in intervention schools and 33% in comparison schools.

Students in the sample began the study in the Reception year (pre-K) and were followed for 3 years, through Year 2—the equivalent of first grade in the United States. The WWC based effectiveness ratings on findings after 3 years of exposure from the analytic sample of 886 students in 17 intervention and 18 comparison schools: 415 students in the *SFA*® group and 471 in the comparison group.

Intervention group

Students in the intervention group received reading instruction through *SFA-UK*®. The instruction was aligned with normal *SFA*® practices that include the *SFA*® reading curriculum, tutoring for students, quarterly assessments, a facilitator who worked with school personnel, and training for all intervention teachers. The family services component of *SFA*® was underutilized, with the emphasis being on within-school practices. Intervention schools were already implementing *SFA*®, and the study was conducted over the entire school year for 3 successive school years.

Comparison group

Students in the comparison group continued using their regular, previously planned curricula (i.e., *Letters and Sounds*; *Jolly Phonics*; *Read, Write Inc*.). No other information was provided on the comparison curricula.

Outcomes and measurement

Outcomes were measured during June–July 2011; the pretest was administered in September 2008. Primary findings reflect 3 years of exposure to the *SFA*® intervention for students at the end of Year 2 —the equivalent of first grade in the United States. Two subtests of the WRMT were administered: Word Identification and Word Attack. The WWC reviewed these outcomes under the alphabetics domain. The study also used three subtests from the York Assessment of Reading Comprehension (YARC): Rate Ability, Comprehension, and Accuracy. The WWC reviewed the Rate Ability and Accuracy subtests in the reading fluency domain and the Comprehension subtest in the comprehension domain. The study also used the British Picture Vocabulary Scale-Second Edition (BPVS-II) as the pretest measure. Post-testing occurred in the spring of 2009, spring 2010, and spring 2011, at the conclusion of the Reception, Year 1, and Year 2 grades, respectively. Only results from the conclusion of Year 2 are reported in the study. For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

At *SFA*® schools, classroom observations were conducted to produce general assessment of implementation fidelity, and trainers from *SFA-UK*® made their normal implementation visits throughout each year of the study.

Appendix B: Outcome measures for each domain

Alphabetics	
Letter knowledge	
Woodcock Reading Mastery Test (WRMT) Letter Identification subtest	This standardized test measures the number of letters that students are able to identify correctly (as cited in Borman et al., 2007). This outcome is only reported as a supplemental finding.
Phonics	
Test of Word Reading Efficiency (TOWRE)	This standardized test is a nationally-normed, age-based measure of word reading accuracy and fluency. The assessment consists of two subtests: Sight Word Efficiency (SWE) and Phonetic Decoding Efficiency (PDE). The SWE subtest assesses the number of real printed words that can be accurately identified within 45 seconds. The PDE subtest measures the number of pronounceable printed nonwords that can be accurately decoded within 45 seconds. Reliability estimate is above 0.90 (as cited in Quint et al., 2015).
Woodcock-Johnson III (WJ-III) Tests of Achievement and Woodcock Language Proficiency Battery (WLPB) Letter-Word Identification subtest	This standardized test requires the child to identify letters that appear in large type, and then to pronounce words correctly. Items become increasingly difficult as the selected words appear less and less frequently in written English. Reliability estimates range from 0.97 to 0.99 for ages 5–7 (as cited in Madden et al., 1993; Quint et al., 2015).
WJ-III Tests of Achievement Word Attack subtest	This standardized test requires the child to produce the sounds for individual letters, then read aloud letter combinations that are regular patterns in English but are nonwords or low-frequency words. Reliability estimates are 0.92 to 0.99 for ages 5–7 (as cited in Quint et al., 2015).
WRMT and WLPB Word Attack subtest	This standardized test measures phonemic decoding skills by asking students to read pseudowords. Students are aware that the words are not real. They cannot read the pseudowords by sight and must rely on phonological processes to decode them (as cited in Borman et al., 2007; Madden et al., 1993; Ross et al, 1997; Ross & Casey, 1998a; Ross & Casey, 1998b; Ross et al., 1998; Ross, Smith & Casey, 1995; Tracey et al., 2014).
WRMT Word Identification subtest	This standardized test measures basic word reading skills and requires the child to read aloud isolated words that range in frequency and difficulty (as cited in Borman et al., 2007, Ross & Casey, 1998a; Ross & Casey, 1998b; Ross et al., 1995; Tracey et al., 2014).
Reading fluency	
Gray Oral Reading Test (GORT) Passage subtest	This standardized test provides a measure of students' oral reading performance. The Passage subtest is a combination of scores on the Rate and Accuracy scales. The Rate subtest measures speed of oral reading, and the Accuracy subtest measures the accuracy of a student's oral reading (as cited in Madden et al., 1993).
York Assessment of Reading Comprehension (YARC) Accuracy subtest	This standardized test is an individually administered assessment of a student's decoding and sight reading ability, their reading fluency, and how well they understand what they have read. The test comprises both fiction and nonfiction texts, and measures the accuracy, rate, and comprehension of oral reading skills in children between ages 5 years to 11 years 11 months. Accuracy is measured by total number and percentage of errors and across different types of errors (i.e., mispronunciation, substitutions, omissions, etc.). The reliability coefficient for this assessment is .95 (as cited in Tracey et al., 2014).
YARC Reading Rate subtest	This standardized test is an individually administered assessment of a student's decoding and sight reading ability, their reading fluency and how well they understand what they have read. The test comprises both fiction and non-fiction texts, and measures the accuracy, rate, and comprehension of oral reading skills in children between 5 to 11 years 11 months. Reading rates are measured as the number of words read correctly per minute. The reliability coefficient for this assessment is .87 (as cited in Tracey et al., 2014).
Comprehension	
California Achievement Test (CAT) Total Reading	This standardized test is a norm and criterion referenced annual assessment. The Reading Composite includes the Vocabulary and Comprehension subtests. The Vocabulary subtest measures student word knowledge, given limited context, as well as the ability to identify missing words within a longer passage or sentence. The Comprehension subtest measures information recall, meaning construction, form analysis, and meaning evaluation of seven different selections. Passages reflect a wide range of narrative, expository, contemporary, and traditional texts (as cited in Madden et al., 1993). This outcome is only reported as a supplemental finding.

Comprehensive Tests of Basic Skills (CTBS) Total Reading	This standardized test is a group-administered assessment that provides three reading scores: Reading Comprehension, Vocabulary, and Total Reading (as cited in Madden et al., 1993). Also known as the Terra Nova, this assessment combines selected-response items with constructed-response items that allow students to produce short and extended responses. The Reading composite score is the average of Reading Comprehension and Vocabulary subtest scores. The Reading Comprehension subtest items focus on five objectives: oral comprehension of passages read aloud; basic understanding of literal meanings of passages; analyzing text; evaluating and extending meaning; and identifying reading strategies. The Vocabulary subtest focuses on three objectives: understanding word meaning; identifying multi-meaning words; and inferring words in context.
Gates-MacGinitie Reading Test (4th Edition, Level 3, Form S)	This standardized test has two components which independently assess reading vocabulary and comprehension skills. The Vocabulary subtest measures each student's reading vocabulary by asking the student to choose one word or phrase that means most nearly the same as a presented word. The test contains 45 questions. The Comprehension subtest measures each student's ability to read and understand different types of prose. The test contains 11 passages of various lengths and subjects and 48 questions. The scores from the two tests can be combined to give an overall reading score that can be reported in terms of a grade-equivalent score. Internal consistency reliabilities in levels 3–5 range from .95 to .96, and test-retest reliabilities range from .89 to .93 (as cited in Borman et al., 2007).
Reading comprehension	
CTBS Reading Comprehension subtest	This standardized test is a group-administered assessment of reading comprehension (as cited in Madden et al., 1993). Also known as the Terra Nova, this assessment combines selected-response items with constructed-response items that allow students to produce short and extended responses. The Reading Comprehension subtest items focus on five objectives: oral comprehension of passages read aloud; basic understanding of literal meanings of passages; analyzing text; evaluating and extending meaning; and identifying reading strategies. This outcome is only reported as a supplemental finding.
Durrell Analysis of Reading Difficulty (DARD) Silent Reading subtest	This standardized test is an individually administered diagnostic assessment of reading accuracy, reading rate, and oral reading comprehension. Silent comprehension of paragraphs is assessed by having the student read graded selections and then answer a series of questions (as cited in Madden et al., 1993).
GORT Comprehension subtest	This standardized test provides a measure of students' oral reading performance. The comprehension subtest requires a student to respond to five multiple choice questions following each story; a variety of literal, inferential, and critical questions are included (as cited in Madden et al., 1993).
WJ-III Tests of Achievement Passage Comprehension subtest	This standardized test measures comprehension by asking students to match pictographic representations of words with actual pictures of the object, choose pictures represented by a phrase, and read several short passages and identify missing key words. Reliability estimate is 0.96 for ages 5–7 (as cited in Quint et al., 2015).
WRMT and WLPB Passage Comprehension subtest	This standardized test measures comprehension by having students read silently and fill in missing words in a short paragraph (as cited in Madden et al., 1993; Ross & Casey, 1998a; Ross & Casey, 1998b; Ross et al., 1995).
YARC Comprehension subtest	This standardized test is an individually administered assessment of students' decoding and sight reading ability, their reading fluency, and how well they understand what they have read. The test comprises both fiction and nonfiction texts, and measures the accuracy, rate, and comprehension of oral reading skills in children between ages 5 years to 11 years 11 months. The questions that are linked to each passage demand the use of deduction and inference (cohesive device, knowledge-based, and elaborative) to arrive at the answers. Reading comprehension is measured by asking the subject to read a passage and then answer questions about it. The reliability coefficient for this assessment is .62 (as cited in Tracey et al., 2014).
Vocabulary development	
CTBS Reading Vocabulary subtest	The standardized test is a group-administered assessment of vocabulary (as cited in Madden et al., 1993). Also known as the Terra Nova, this assessment combines selected-response items with constructed-response items that allow students to produce short and extended responses. The Vocabulary subtest focuses on three objectives: understanding word meaning; identifying multi-meaning words; and inferring words in context. This outcome is only reported as a supplemental finding.

General reading achievement	
CTBS Total Language	This standardized test is a group-administered assessment of language (as cited in Madden et al., 1993). Also known as the Terra Nova, this assessment combines selected-response items with constructed-response items that allow students to produce short and extended responses. The Language composite score is the average of scores on the Language and Language Mechanics subtests. The Language subtest covers four objectives: introduction to print; understanding sentence structure; writing strategies; and editing skills. The Language Mechanics subtest focuses on three objectives: appropriate construction of sentences, phrases, and clauses; appropriate writing conventions; and editing skills.
DARD Oral Reading subtest	This standardized diagnostic test is an individually administered assessment of reading accuracy, reading rate, and oral reading comprehension. Oral Reading is assessed by having the student read aloud graded passages and then answer a series of comprehension questions (as cited in Madden et al., 1993; Ross et al., 1998; Ross & Casey, 1998a; Ross & Casey, 1998b; Ross et al., 1995).
GORT	This standardized test provides a measure of students' oral reading performance and is calculated by combining scores on the Passage and Comprehension subtests. The Passage subtest is a combination of scores on the Rate and Accuracy subtests (the Rate subtest measures speed of oral reading, and the Accuracy subtest measures the accuracy of a student's oral reading). The Comprehension subtest requires a student to respond to five multiple choice questions following each story (a variety of literal, inferential, and critical questions are included) (as cited in Ross et al., 1995).
Stanford Achievement Test, 9th Edition (SAT-9) Reading	This standardized norm referenced test assesses comprehension of three types of reading material: textual (nonfiction, general information); recreational (fiction); and functional (material encountered in everyday life, such as advertisements). Test questions tap various comprehension skills from the basic literal level up to the inferential and critical levels of reading comprehension. The authors converted all assessment scores to normal curve equivalent scores (as cited in Skindrud & Gersten, 2006).
SAT-9 Language	This standardized norm referenced test assesses punctuation and capitalization skills and the ability to apply grammatical concepts correctly. Test questions also assess language expression, or the ability to manipulate words, phrases, and clauses, and the ability to recognize correct, effective sentence structure and writing style. The authors converted all assessment scores to normal curve equivalent scores (as cited in Skindrud & Gersten, 2006).

Appendix C.1: Findings included in the rating for the alphabetics domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Borman et al. (2007)[a]			**3 years of intervention**					
WRMT Word Attack	Grade 2	35 schools/ 1,936 students	493.48 (17.39)	486.26 (19.20)	7.22	0.39	+15	.01
WRMT Word Identification	Grade 2	35 schools/ 1,928 students	462.34 (25.68)	455.12 (28.72)	7.22	0.27	+10	.09
Domain average for alphabetics (Borman et al., 2007)						**0.33**	**+13**	**Statistically significant**
Quint et al. (2015)[b]			**3 years of intervention**					
Test of Word Reading Efficiency	Grade 2	37 schools/ 2,873 students	46.96 (nr)	46.15 (15.82)	0.81	0.05	+2	.39
Woodcock-Johnson III (WJ-III) Tests of Achievement Letter-Word Identification	Grade 2	37 schools/ 2,902 students	39.99 (nr)	39.18 (8.81)	0.82	0.09	+4	.15
WJ-III Word Attack	Grade 2	37 schools/ 2,907 students	15.53 (nr)	14.37 (6.81)	1.15	0.17	+7	< .01
Domain average for alphabetics (Quint et al., 2015)						**0.10**	**+4**	**Statistically significant**
Madden et al. (1993)[c]			**3 years of intervention**					
Woodcock Language Proficiency Battery (WLPB) Letter-Word Identification	Grade 1/ Cohort 1	10 schools/ 492 students	18.53 (5.34)	15.91 (6.59)	2.62	0.44	+17	> .05
WLPB Word Attack	Grade 1/ Cohort 1	10 schools/ 492 students	5.46 (4.11)	2.25 (3.55)	3.21	0.83	+30	< .01
WLPB Letter-Word Identification	Grade 2/ Cohort 2	10 schools/ 440 students	25.09 (6.65)	21.54 (6.72)	3.55	0.53	+20	> .05
WLPB Word Attack	Grade 2/ Cohort 2	10 schools/ 440 students	8.63 (6.27)	5.21 (4.76)	3.42	0.61	+23	< .05
WLPB Letter-Word Identification	Grade 3/ Cohort 3	10 schools/ 410 students	28.69 (6.72)	25.56 (6.19)	3.12	0.48	+19	> .05
WLPB Word Attack	Grade 3/ Cohort 3	10 schools/ 410 students	10.77 (6.94)	7.02 (5.49)	3.74	0.60	+23	< .05
Domain average for alphabetics (Madden et al., 1993)						**0.58**	**+22**	**Statistically significant**
Ross et al. (1998)[d]			**2 years of intervention**					
WRMT Word Attack	Grade 2	5 schools/ 128 students	23.62 (9.65)	23.69 (10.16)	–0.07	–0.01	0	> .05
WRMT Word Identification	Grade 2	5 schools/ 128 students	51.94 (13.38)	52.95 (14.79)	–1.01	–0.07	–3	> .05
Domain average for alphabetics (Ross et al., 1998)						**–0.04**	**–2**	**Not statistically significant**

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation) Intervention group	Mean (standard deviation) Comparison group	WWC calculations Mean difference	WWC calculations Effect size	WWC calculations Improvement index	*p*-value
Ross & Casey (1998a)[e]			**2 years of intervention**					
WRMT Word Attack	Grade 1	7 schools/ 288 students	12.25 (7.36)	10.39 (8.59)	1.86	0.22	+9	< .01
WRMT Word Identification	Grade 1	7 schools/ 288 students	32.14 (14.63)	31.32 (14.72)	0.82	0.06	+2	> .05
Domain average for alphabetics (Ross & Casey, 1998a)						**0.14**	**+6**	**Not statistically significant**
Ross & Casey (1998b)[f]			**1 year of intervention**					
WRMT Word Attack	Kindergarten	4 schools/ 265 students	3.64 (5.32)	2.39 (4.67)	1.25	0.25	+10	> .05
WRMT Word Identification	Kindergarten	4 schools/ 265 students	8.38 (12.28)	5.70 (10.22)	2.68	0.23	+9	> .05
Domain average for alphabetics (Ross & Casey, 1998b)						**0.24**	**+9**	**Not statistically significant**
Ross et al. (1995)[g]			**4 years of intervention**					
WRMT Word Attack	Grade 3/ Cohort 1	4 schools/ 74 students	27.16 (11.80)	26.78 (11.28)	0.38	0.03	+1	> .05
WRMT Word Identification	Grade 3/ Cohort 1	4 schools/ 74 students	60.73 (11.42)	60.28 (15.54)	0.45	0.04	+1	> .05
WRMT Word Attack	Grade 4/ Cohort 2	4 schools/ 77 students	27.11 (10.72)	24.83 (12.48)	2.28	0.20	+8	> .05
WRMT Word Identification	Grade 4/ Cohort 2	4 schools/ 77 students	63.56 (9.67)	62.03 (18.27)	1.53	0.11	+4	> .05
			3 years of intervention					
WRMT Word Attack	Grade 2/ minority Cohort 3	4 schools/ 54 students	23.58 (8.80)	19.66 (11.44)	3.92	0.38	+15	> .05
WRMT Word Identification	Grade 2/ minority Cohort 3	4 schools/ 54 students	53.67 (10.03)	47.82 (12.55)	5.85	0.51	+20	> .05
Domain average for alphabetics (Ross et al., 1995)						**0.21**	**+8**	**Not statistically significant**
Tracey et al. (2014)[h]			**3 years of intervention**					
WRMT Word Attack	Year 2	35 schools/ 886 students	29.22 (8.72)	27.65 (8.66)	1.57	0.18	+7	< .01
WRMT Word Identification	Year 2	35 schools/ 886 students	64.66 (15.40)	61.47 (15.00)	3.19	0.21	+8	< .05
Domain average for alphabetics (Tracey et al., 2014)						**0.20**	**+8**	**Statistically significant**
Domain average for alphabetics across all studies						**0.22**	**+9**	**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 the study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable. nr = not reported.

[a] For Borman et al. (2007), corrections for clustering and multiple comparisons were needed for the two measures of alphabetics and resulted in a WWC-computed critical *p*-value of .025 for the *WRMT Word Attack* measure; therefore, the WWC finds the result for the *WRMT Word Attack* outcome to be statistically significant. The *p*-values presented here were calculated by the WWC, unadjusted group means and standard deviations were obtained for the combined longitudinal and in-mover sample (of students who did not have any data imputed) from the study authors. This study is characterized as having a statistically significant positive effect, because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[b] For Quint et al. (2015), a correction for multiple comparisons was needed and resulted in a WWC-computed critical *p*-value of .017 for the *WJ-III Word Attack* measure; therefore, the WWC finds the result for the *WJ-III Word Attack* outcome to be statistically significant. The *p*-values and effect sizes presented here were reported in the original study. The authors used standard deviations for the comparison group means to calculate an effect sizes. This study is characterized as having a statistically significant effect because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[c] For Madden et al. (1993), the *p*-values presented here were calculated by the WWC. Corrections for clustering and multiple comparisons were needed for the six measures of alphabetics and resulted in a WWC-computed critical *p*-value of .0083 for the *WRMT Word Attack* measure for the Cohort 1 group (*p*=006); therefore, the WWC finds the result for the *WRMT Word Attack* outcome to be statistically significant. None of the other findings were statistically significant after the adjustment for multiple comparisons. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools. The intervention and comparison group means reported in this table are analysis of covariance (ANCOVA)-adjusted. This study is characterized as having a statistically significant positive effect because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons (and correcting for clustering when not properly aligned). For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[d] For Ross et al. (1998), corrections for clustering and multiple comparisons were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated across participating schools. The reported group means are based on an ANCOVA, which adjusted for pretest. This study is characterized as having an indeterminate effect, because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[e] For Ross and Casey (1998a), a correction for clustering was needed and resulted in a WWC-computed *p*-value larger than .05 for the WRMT Word Attack outcome; therefore, the WWC does not find the result to be statistically significant. The *p*-values presented here were reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated by the WWC across participating schools. The intervention and comparison group means reported in this table are multivariate analysis of covariance (MANCOVA)-adjusted. This study is characterized as having an indeterminate effect because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[f] For Ross and Casey (1998b), corrections for clustering and multiple comparisons were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are MANCOVA-adjusted. This study is characterized as having an indeterminate effect because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[g] For Ross et al. (1995), corrections for clustering and multiple comparisons were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values for Cohort 2 outcomes were reported in the original study, and the WWC calculated *p*-values for Cohort 1 and Cohort 3 outcomes. The intervention and comparison group means reported in this table are MANCOVA-adjusted for Cohorts 1 and 2 and unadjusted for Cohort 3. This study is characterized as having an indeterminate effect because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[h] For Tracey et al. (2014), a correction for multiple comparisons was needed for the two measures of alphabetics and resulted in WWC-computed critical *p*-values of .025 for the WRMT Word Attack measure and .05 for the WRMT Word Identification measure; therefore, the WWC finds the results for two outcomes to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted, as reported by the authors in response to a query from the WWC. This study is characterized as having a statistically significant positive effect because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

Appendix C.2: Findings included in the rating for the reading fluency domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Madden et al. (1993)[a]			**5 years of intervention**					
Gray Oral Reading Test (GORT) Passage subtest	Grade 4/ Cohort 2	10 schools/ 306 students	30.33 (18.23)	22.27 (15.73)	8.06	0.47	+18	< .01
Domain average for reading fluency (Madden et al., 1993)						**0.47**	**+18**	**Not statistically significant**
Tracey et al. (2014)[b]			**3 years of intervention**					
The York Assessment of Reading Comprehension (YARC) Accuracy	Year 2	35 schools/ 880 students	47.50 (9.78)	46.64 (9.89)	0.86	0.09	+3	> .05
YARC Reading Rate	Year 2	35 schools/ 737 students	60.97 (14.26)	58.37 (15.05)	2.60	0.18	+7	> .05
Domain average for reading fluency (Tracey et al., 2014)						**0.13**	**+5**	**Not statistically significant**
Domain average for reading fluency across all studies						**0.30**	**+12**	**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 the study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable.

[a] For Madden et al. (1993), a correction for clustering was needed and resulted in a WWC-computed *p*-value of .13 for the GORT Passage outcome; therefore, the WWC does not find the result to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted. This study is characterized as having a substantively important positive effect because the estimated effect size for the outcome in this domain is positive and not statistically significant but is substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[b] For Tracey et al. (2014), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted, as reported by the authors in response to a query from the WWC. This study is characterized as having an indeterminate effect because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

Appendix C.3: Findings included in the rating for the comprehension domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Borman et al. (2007)[a]			**3 years of intervention**					
WRMT Passage Comprehension	Grade 2	35 schools/ 1,935 students	480.54 (16.08)	476.66 (16.96)	3.88	0.23	+9	< .05
			1 year of intervention					
Gates-MacGinitie Reading Test (level 3)	Grade 3	35 schools/ 2,420 students	451.60 (34.90)	451.60 (37.70)	0.00	0.00	0	> .05
Domain average for comprehension (Borman et al., 2007)						**0.12**	**+5**	**Not statistically significant**

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Quint et al. (2015)[b]			**3 years of intervention**					
Woodcock-Johnson III (WJ-III) Passage Comprehension	Grade 2	37 schools/ 2,894 students	21.03 (nr)	20.88 (4.83)	0.15	0.03	+1	.56
Domain average for comprehension (Quint et al., 2015)						**0.03**	**+1**	**Not statistically significant**
Madden et al. (1993)[c]			**5 years of intervention**					
Comprehensive Tests of Basic Skills (CTBS) Total Reading	Grade 4/ Cohort 2	9 schools/ 254 students	661.30 (52.63)	649.00 (56.97)	12.30	0.23	+9	< .10
Gray Oral Reading Test (GORT) Comprehension	Grade 4/ Cohort 2	10 schools/ 306 students	20.97 (9.55)	17.48 (10.44)	3.49	0.35	+14	< .01
			4 years of intervention					
Woodcock Reading Mastery Test (WRMT) Passage Comprehension	Grade 2/ lowest 25%/ Cohort 1	10 schools/ 104 students	16.44 (8.50)	10.48 (6.43)	5.96	0.79	+29	< .01
			2 years of intervention					
Durrell Analysis of Reading Difficulty (DARD) Silent Reading	Grade 2/ Cohort 3	10 schools/ 320 students	8.16 (6.63)	5.89 (5.35)	2.27	0.38	+15	> .05
Domain average for comprehension (Madden et al., 1993)						**0.49**	**+19**	**Statistically significant**
Ross et al. (1998)[d]			**2 years of intervention**					
WRMT Passage Comprehension	Grade 2	5 schools/ 128 students	27.43 (8.13)	29.65 (8.49)	–2.22	–0.27	–11	> .05
Domain average for comprehension (Ross et al., 1998)						**–0.27**	**–11**	**Not statistically significant**
Ross & Casey (1998a)[e]			**2 years of intervention**					
WRMT Passage Comprehension	Grade 1	7 schools/ 288 students	16.09 (8.46)	15.44 (8.96)	0.65	0.07	+3	> .05
Domain average for comprehension (Ross & Casey, 1998a)						**0.07**	**+3**	**Not statistically significant**
Ross & Casey (1998b)[f]			**1 year of intervention**					
WRMT Passage Comprehension	Kindergarten	4 schools/ 265 students	3.71 (5.16)	3.66 (5.78)	0.05	0.01	0	> .05
Domain average for comprehension (Ross & Casey, 1998b)						**0.01**	**0**	**Not statistically significant**

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Ross et al. (1995)[g]			**4 years of intervention**					
WRMT Passage Comprehension	Grade 3/ Cohort 1	4 schools/ 74 students	33.41 (6.90)	35.02 (8.68)	–1.61	–0.21	–9	> .05
WRMT Passage Comprehension	Grade 4/ Cohort 2	4 schools/ 77 students	33.28 (6.06)	33.00 (11.49)	0.28	0.03	+1	> .05
			3 years of intervention					
WRMT Passage Comprehension	Grade 2/ minority Cohort 3	4 schools/ 54 students	27.58 (4.47)	26.20 (7.94)	1.38	0.22	+9	> .05
Domain average for comprehension (Ross et al., 1995)						**0.01**	**+1**	**Not statistically significant**
Tracey et al. (2014)[h]			**3 years of intervention**					
York Assessment of Reading Comprehension (YARC) Comprehension	Year 2	35 schools/ 868 students	53.04 (10.51)	52.61 (9.29)	0.43	0.04	+2	> .05
Domain average for comprehension (Tracey et al., 2014)						**0.04**	**+2**	**Not statistically significant**
Domain average for comprehension across all studies						**0.06**	**+3**	**na**

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on 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 the study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable. nr = not reported.

[a] For Borman et al. (2007) *third-year* outcomes, a correction for clustering was needed and resulted in a WWC-computed *p*-value of .14 for the WRMT Passage Comprehension outcome; therefore, the WWC does not find the result to be statistically significant. The *p*-value presented here was reported in the original study. Unadjusted group means and standard deviations were obtained for the combined longitudinal and in-mover sample (of students who did not have any data imputed) from the study authors.

For the *first-year* outcome, the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value and effect size presented here were reported in the original study. Group means and standard deviations were obtained through the author query. The reported intervention group means are calculated as the comparison group means plus the HLM level-2 coefficient.

This study is characterized as having an indeterminate effect because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[b] For Quint et al. (2015), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value and effect size presented here were reported in the original study. The authors used standard deviations for the full comparison group to calculate an effect size. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[c] For Madden et al. (1993) *fifth-year* outcomes, a correction for clustering was needed and resulted in a WWC-computed *p*-value of .22 for the GORT Comprehension outcome; therefore, the WWC does not find the result to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

For the *fourth-year* outcome, a correction for clustering was needed and resulted in a WWC-computed *p*-value of .02 for the Passage Comprehension outcome; therefore, the WWC finds the result to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

For the *second-year* outcome, the *p*-value presented here was calculated by the WWC. A correction for clustering was needed, and the WWC did not find the result to be statistically significant. The intervention and comparison group means reported in this table are ANCOVA-adjusted. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

This study is characterized as having a statistically significant positive effect because at least one measure is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons and correcting for clustering. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[d] For Ross et al. (1998), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated across participating schools. The reported group means are based on an ANCOVA and adjusted for pretest. This study is characterized as having a substantively important negative effect because the estimated effect for the outcome in this domain is negative, not statistically significant after any necessary adjustments, and is substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[e] For Ross and Casey (1998a), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated by the WWC across participating schools. The intervention and comparison group means reported in this table are MANCOVA-adjusted. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[f] For Ross and Casey (1998b), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are MANCOVA-adjusted. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[g] For Ross et al. (1995), corrections for clustering were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values for Cohort 2 outcomes were reported in the original study, and the WWC calculated *p*-values for Cohort 1 and Cohort 3 outcomes. The intervention and comparison group means reported in this table are MANCOVA-adjusted for Cohorts 1 and 2, and unadjusted for Cohort 3. This study is characterized as having an indeterminate effect because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[h] For Tracey et al. (2014), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted, as reported by the authors in response to a query from the WWC. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in the comprehension domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

Appendix C.4: Findings included in the rating for the general reading achievement domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Madden et al. (1993)[a]			**5 years of intervention**					
Comprehensive Tests of Basic Skills (CTBS) Total Language	Grade 4/ Cohort 2	9 schools/ 255 students	677.49 (47.38)	660.86 (42.98)	16.63	0.36	+14	< .01
			3 years of intervention					
Durrell Analysis of Reading Difficulty (DARD) Oral Reading	Grade 1/ Cohort 1	10 schools/ 492 students	5.59 (4.78)	4.26 (5.16)	1.33	0.27	+11	> .05
DARD Oral Reading	Grade 3/ Cohort 3	10 schools/ 410 students	16.66 (7.00)	13.25 (7.13)	3.41	0.48	+19	< .05
Domain average for general reading achievement (Madden et al., 1993)						**0.37**	**+14**	**Not statistically significant**
Ross et al. (1998)[b]			**2 years of intervention**					
DARD Oral Reading	Grade 2	5 schools/ 128 students	11.93 (6.47)	12.63 (6.42)	–0.70	–0.11	–4	> 0.05
Domain average for general reading achievement (Ross et al., 1998)						**–0.11**	**–4**	**Not statistically significant**
Ross & Casey (1998a)[c]			**2 years of intervention**					
DARD Oral Reading	Grade 1	7 schools/ 288 students	5.35 (4.63)	4.74 (4.52)	0.61	0.13	+5	> .05

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation) Intervention group	Mean (standard deviation) Comparison group	WWC calculations Mean difference	WWC calculations Effect size	WWC calculations Improvement index	*p*-value
Domain average for general reading achievement (Ross & Casey, 1998a)						**0.13**	**+5**	**Not statistically significant**
Ross & Casey (1998b)[d]			**1 year of intervention**					
DARD Oral Reading	Kindergarten	4 schools/ 265 students	0.78 (2.60)	0.49 (1.89)	0.29	0.12	+5	> .05
Domain average for general reading achievement (Ross & Casey, 1998b)						**0.12**	**+5**	**Not statistically significant**
Ross et al. (1995)[e]			**4 years of intervention**					
DARD Oral Reading	Grade 3/ Cohort 1	4 schools/ 74 students	19.80 (6.49)	22.44 (9.64)	–2.64	–0.35	–14	> .05
Gray Oral Reading Test (GORT-3)	Grade 4/ Cohort 2	4 schools/ 77 students	83.80 (13.39)	90.54 (23.43)	–6.74	–0.37	–14	> .05
3 years of intervention								
DARD Oral Reading	Grade 2/ minority Cohort 3	4 schools/ 54 students	14.64 (6.14)	12.60 (5.66)	2.04	0.34	+13	> .05
Domain average for general reading achievement (Ross et al., 1995)						**–0.13**	**–5**	**Not statistically significant**
Skindrud & Gersten (2006)[f]			**2 years of intervention**					
Stanford Achievement Test, 9th Edition (SAT-9) Reading	Grade 3, Cohort 1	12 schools/ 434 students	38.60 (18.50)	43.90 (16.50)	–5.30	–0.31	–12	< .01
			1 year of intervention					
SAT-9 Language	Grade 3/ lowest 25% Cohort 2	12 schools/ 97 students	28.80 (12.30)	29.60 (12.80)	–0.80	–0.06	–3	> .05
Domain average for general reading achievement (Skindrud & Gersten, 2006)						**–0.19**	**–7**	**Not statistically significant**
Domain average for general reading achievement across all studies						**0.03**	**+1**	**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 the study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable.

[a] For Madden et al. (1993), a correction for clustering was needed and resulted in WWC-computed *p*-values larger than .05 for the *CTBS Total Language* outcome and for the *DARD Oral Reading* outcome in grade 3; therefore, the WWC does not find the results to be statistically significant. The *p*-values presented here for *outcomes in grades 3 and 4* were reported in the original study, while the *p*-value for the outcome in grade 1 was calculated by the WWC. Because study authors analyzed each matched pair of schools separately, the WWC aggregated means and standard deviations for the *SFA®* schools and for the comparison schools for the DARD Oral Reading outcomes. The intervention and comparison group means reported in this table are ANCOVA-adjusted. This study is characterized as having a substantively important positive effect because the mean effect size for the outcomes in this domain is positive and not statistically significant but is substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[b] For Ross et al. (1998), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated across participating schools. The intervention and comparison group means reported in this table are ANCOVA-adjusted. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[c] For Ross and Casey (1998a), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated by the WWC across participating schools. The group means are MANCOVA-adjusted. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[d] For Ross and Casey (1998b), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are MANCOVA-adjusted. This study is characterized as having an indeterminate effect because the estimated effect for the outcome in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[e] For Ross et al. (1995), corrections for clustering were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values for Cohort 2 outcomes were reported in the original study, and the WWC calculated *p*-values for Cohort 1 and Cohort 3 outcomes. The intervention and comparison group means reported in this table are MANCOVA-adjusted for Cohorts 1 and 2 and unadjusted for Cohort 3. This study is characterized as having an indeterminate effect because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook, version 3.0, page 26.

[f] For Skindrud and Gersten (2006), a correction for clustering was needed and resulted in a WWC-computed *p*-value of .30 for the Cohort 1 SAT-9 outcome; therefore, the WWC does not find the result to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted. This study is characterized as having an indeterminate effect, because the mean effect reported is neither statistically significant nor substantively important. For more information, please refer to the WWC Standards and Procedures Handbook, version 3.0, p. 26.

Appendix D.1: Description of supplemental findings for the alphabetics domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Borman et al. (2007)[a]			**2 years of intervention**					
Woodcock Reading Mastery Test (WRMT) Letter Identification	Grade 1	38 schools	nr (nr)	nr (nr)	nr	0.18	+7	< .05
WRMT Word Attack	Grade 1	38 schools	nr (nr)	nr (nr)	nr	0.19	+8	> .05
WRMT Word Identification	Grade 1	38 schools	nr (nr)	nr (nr)	nr	0.15	+6	> .05
Quint et al. (2015)[b]			**2 years of intervention**					
Test of Word Reading Efficiency	Grade 1	37 schools/ 2,802 students	29.50 (nr)	28.73 (16.00)	0.77	0.05	+2	.41
Woodcock-Johnson III (WJ-III) Letter-Word Identification	Grade 1	37 schools/ 2,952 students	30.34 (nr)	29.80 (8.84)	0.54	0.06	+2	.26
WJ-III Word Attack	Grade 1	37 schools/ 2,962 students	12.36 (nr)	10.51 (6.05)	1.85	0.31	+12	< .01
			1 year of intervention					
WJ-III Letter-Word Identification	Kindergarten	37 schools/ 2,893 students	19.67 (nr)	19.74 (6.98)	–0.07	–0.01	0	.90
WJ-III Word Attack	Kindergarten	37 schools/ 2,893 students	5.74 (nr)	5.21 (3.07)	0.53	0.18	+7	.03
Madden et al. (1993)[c]			**4 years of intervention**					
WRMT Word Attack	Grade 2/ lowest 25%/ Cohort 1	10 schools/ 104 students	11.36 (8.47)	1.80 (3.14)	9.56	1.53	+44	< .01
WRMT Word Identification	Grade 2/ lowest 25%/ Cohort 1	10 schools/ 104 students	36.12 (13.35)	21.08 (10.40)	15.04	1.26	+40	< .01
			3 years of intervention					
Woodcock Language Proficiency Battery (WLPB) Letter-Word Identification	Grade 1/ lowest 25%/ Cohort 1	10 schools/ 126 students	16.65 (5.34)	12.56 (6.66)	4.09	0.67	+25	< .05
WLPB Word Attack	Grade 1/ lowest 25%/ Cohort 1	10 schools/ 126 students	4.92 (4.38)	1.52 (3.39)	3.40	0.86	+31	.01
WLPB Letter-Word Identification	Grade 2/ lowest 25%/ Cohort 2	10 schools/ 112 students	19.19 (4.80)	15.50 (5.54)	3.69	0.71	+26	.04
WLPB Word Attack	Grade 2/ lowest 25%/ Cohort 2	10 schools/ 112 students	4.73 (3.68)	1.48 (2.17)	3.25	1.07	+36	< .01

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
WLPB Letter-Word Identification	Grade 3/ lowest 25%/ Cohort 3	10 schools/ 104 students	25.06 (6.85)	21.31 (4.75)	3.75	0.63	+24	.07
WLPB Word Attack	Grade 3/ lowest 25%/ Cohort 3	10 schools/ 104 students	7.85 (6.52)	4.02 (4.02)	3.83	0.70	+26	.04
2 years of intervention								
WRMT Letter-Word Identification	Grade 1/ Cohort 2	10 schools/ 338 students	18.23 (5.82)	16.51 (5.30)	1.72	0.31	+12	.31
WRMT Word Attack	Grade 1/ Cohort 2	10 schools/ 338 students	6.15 (5.01)	2.62 (3.79)	3.53	0.79	+29	.01
WRMT Letter-Word Identification	Grade 1/ lowest 25%/ Cohort 2	10 schools/ 86 students	13.94 (4.64)	12.18 (3.16)	1.76	0.44	+17	.21
WRMT Word Attack	Grade 1/ lowest 25%/ Cohort 2	10 schools/ 86 students	3.06 (3.11)	1.15 (2.30)	1.91	0.69	+26	> .05
WRMT Letter-Word Identification	Grade 2/ Cohort 3	10 schools/ 320 students	24.08 (7.14)	21.03 (6.40)	3.05	0.45	+17	.15
WRMT Word Attack	Grade 2/ Cohort 3	10 schools/ 320 students	8.11 (5.82)	4.49 (4.87)	3.62	0.67	+25	.03
WRMT Letter-Word Identification	Grade 2/ lowest 25%/ Cohort 3	10 schools/ 78 students	19.84 (5.88)	17.02 (4.80)	2.82	0.52	+20	.15
WRMT Word Attack	Grade 2/ lowest 25%/ Cohort 3	10 schools/ 78 students	5.00 (3.46)	1.44 (2.07)	3.56	1.24	+39	< .01
1 year of intervention								
WRMT Letter-Word Identification	Grade 1/ Cohort 3	14 schools/ 584 students	19.27 (6.22)	17.43 (6.29)	1.84	0.29	+12	> .05
WRMT Word Attack	Grade 1/ Cohort 3	14 schools/ 584 students	5.92 (4.83)	3.49 (4.63)	2.43	0.51	+20	.04
WRMT Letter-Word Identification	Grade 1/ lowest 25%/ Cohort 3	10 schools/ 118 students	14.32 (5.74)	11.78 (5.05)	2.54	0.47	+18	> .05
WRMT Word Attack	Grade 1/ lowest 25%/ Cohort 3	10 schools/ 118 students	3.46 (3.32)	0.97 (2.10)	2.49	0.89	+31	< .01
Ross et al. (1998)[d]			**1 year of intervention**					
WRMT Word Attack	Grade 1	6 schools/ 252 students	18.35 (nr)	15.86 (nr)	2.49	0.28	+11	.03
WRMT Word Identification	Grade 1	6 schools/ 252 students	nr (nr)	nr (nr)	nr	–0.01	0	> .05

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Ross & Casey (1998a)[e]			**2 years of intervention**					
WRMT Word Attack	Grade 1/ lowest 25%	7 schools/ 79 students	10.11 (6.13)	7.53 (7.85)	2.58	0.35	+14	> .05
WRMT Word Identification	Grade 1/ lowest 25%	7 schools/ 79 students	27.10 (14.25)	25.73 (13.57)	1.37	0.10	+4	> .05
Ross & Casey (1998b)[f]			**1 year of intervention**					
WRMT Word Attack	Kindergarten/ lowest 25%	4 schools/ 69 students	1.03 (1.96)	0.31 (0.95)	0.72	0.44	+17	> .05
WRMT Word Identification	Kindergarten/ lowest 25%	4 schools/ 69 students	3.18 (6.33)	0.62 (1.35)	2.56	0.51	+19	> .05
Ross et al. (1995)[g]			**4 years of intervention**					
WRMT Word Attack	Grades 3–4/ nonminority/ Cohorts 1–2	4 schools/ 81 students	–0.07 (0.90)	0.04 (1.00)	–0.11	–0.12	–5	> .05
WRMT Word Identification	Grades 3–4/ nonminority/ Cohorts 1–2	4 schools/ 81 students	0.01 (0.88)	0.09 (1.29)	–0.08	–0.07	–3	> .05
WRMT Word Attack	Grade 4/ lowest 25% Cohort 2	4 schools/ 19 students	25.09 (8.25)	20.25 (13.66)	4.84	0.43	+17	> .05
WRMT Word Identification	Grade 4/ lowest 25% Cohort 2	4 schools/ 19 students	61.45 (5.97)	48.55 (20.15)	12.90	0.90	+32	> .05
			4 years and 3 years of intervention					
WRMT Word Attack	Grades 2–4/ minority/ Cohorts 1–3	4 schools/ 123 students	0.08 (0.89)	–0.29 (1.06)	0.37	0.39	+15	< .05
WRMT Word Identification	Grades 2–4/ minority/ Cohorts 1–3	4 schools/ 123 students	0.12 (0.78	–0.29 (1.04)	0.41	0.47	+18	< .05

Table Notes: The supplemental findings presented in this table are additional findings 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.

[a] For Borman et al. (2007), a correction for multiple comparisons was needed for the three measures of alphabetics (2 years of intervention) and resulted in a WWC-computed critical *p*-value of .017 for the *WRMT Letter Identification* measure, which is within the authors reported *p*-value range of < .05; therefore, the WWC does not make a determination about statistical significance of the effect. The *p*-values and effect sizes presented here were reported in the original study.

[b] For Quint et al. (2015) second-year outcomes, a correction for multiple comparisons was needed for the three measures of alphabetics and resulted in a WWC-computed critical *p*-value of .017 for the *WJ-III Word Attack* measure; therefore, the WWC finds the result to be statistically significant for the *first-grade WJ-III Word Attack* outcome. The *p*-values and effect sizes presented here were reported in the original study.

For *first-year* outcomes, the authors applied a correction for multiple comparisons for the two measures of alphabetics and did not find the result for the *WJ-III Word Attack* measure for kindergarten to be statistically significant. The WWC confirmed this result. The *p*-values and effect sizes presented here were reported in the original study.

[c] For Madden et al. (1993) *fourth-year* outcomes, corrections for clustering and multiple comparisons were needed but did not affect whether any of the contrasts ceased to be statistically significant. The *p*-values presented here were reported in the original study, and the WWC confirms statistical significance of both findings. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

For *third-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering and six multiple comparisons were needed and resulted in WWC-computed critical *p*-values of .025 for the WLPB *Letter-Identification* measure for the Cohort 2 subgroup (*p*=.038), .033 for the *WLPB Word Attack* measure for the *Cohort 3 subgroup p*=.042, and .042 for the *WLPB Letter-Identification measure* for the Cohort 1 subgroup (*p*=.045); therefore, the WWC did not find these results to be statistically significant. The intervention and comparison group means reported in this table are ANCOVA-adjusted. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

For *second-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering and eight multiple comparisons were needed and resulted in a WWC-computed critical *p*-value of .019 for the *WRMT Word Attack* measure for the *Cohort 3 second-grade subgroup (p=.03)*; therefore, the WWC did not find the result for the *Cohort 3 second-grade subgroup* to be statistically significant. The intervention and comparison group means reported in this table are ANCOVA-adjusted. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

For *first-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering and four multiple comparisons were needed and resulted in a WWC-computed critical *p*-value of .025 for the *WRMT Word Attack* measure (*Cohort 3; p=.044*); therefore, the WWC did not find the result for *the Cohort 3 subgroup* to be statistically significant. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

[d] For Ross et al. (1998), a correction for clustering was needed and resulted in a WWC-computed *p*-value of .25 for the WRMT Word Attack outcome; therefore, the WWC does not find the result to be statistically significant. The *p*-values and effect sizes presented here were reported in the original study.

[e] For Ross and Casey (1998a), corrections for clustering were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated by the WWC across participating schools. The intervention and comparison group means reported in this table are MANCOVA-adjusted.

[f] For Ross and Casey (1998b), corrections for clustering and multiple comparisons were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are MANCOVA-adjusted.

[g] For Ross et al. (1995), corrections for clustering were needed and resulted in WWC-computed *p*-values of .39 and .47, respectively, for the *WRMT Word Identification* and *Word Attack* outcomes *(Grades 2–4/ minority/Cohorts 1–3)*; therefore, the WWC does not find the results to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

D.2: Description of supplemental findings for the comprehension domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Borman et al. (2007)[a]			**2 years of intervention**					
WRMT Passage Comprehension	Grade 1	38 schools	nr (nr)	nr (nr)	nr	0.09	+4	> .05
			1 year of intervention					
Gates-MacGinitie Reading Test (level 3)	Grade 3/at grade level	35 schools/ 662 students	487.65 (34.10)	487.70 (35.80)	–0.05	0.00	0	> .05
Gates-MacGinitie Reading Test (level 3)	Grade 3/ below grade level	35 schools/ 1,474 students	435.79 (26.40)	435.80 (25.60)	–0.01	0.00	0	> .05
Quint et al. (2015)[b]			**2 years of intervention**					
Woodcock-Johnson III (WJ-III) Passage Comprehension	Grade 1	37 schools/ 2,957 students	14.69 (nr)	14.57 (5.36)	0.12	0.02	+1	.69
Madden et al. (1993)[c]			**5 years of intervention**					
Comprehensive Tests of Basic Skills (CTBS) Comprehension Subtest	Grade 4/ Cohort 2	9 schools/ 255 students	676.63 (57.62)	653.95 (66.12)	22.68	0.37	+14	< .01
CTBS Reading Vocabulary subtest	Grade 4/ Cohort 2	9 schools/ 255 students	645.64 (58.65)	643.61 (55.47)	2.03	0.04	+1	> .05

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
2 years of intervention								
Durrell Analysis of Reading Difficulty (DARD) Silent Reading	Grade 1/ Cohort 2	10 schools/ 338 students	4.90 (5.85)	2.67 (4.03)	2.23	0.44	+17	> .05
DARD Silent Reading	Grade 1/ lowest 25%/ Cohort 2	10 schools/ 86 students	1.57 (2.69)	0.61 (1.39)	0.96	0.44	+17	> .05
DARD Silent Reading	Grade 2/ lowest 25%/ Cohort 3	10 schools/ 78 students	5.08 (5.27)	3.18 (3.33)	1.90	0.43	+17	> .05
1 year of intervention								
California Achievement Test (CAT) Total Reading	Grade 1/ Cohort 3	14 schools/ 584 students	479.51 (107.53)	481.76 (101.87)	–2.25	–0.02	0	> .05
DARD Silent Reading	Grade 1/ Cohort 3	14 schools/ 584 students	4.01 (4.18)	3.28 (4.49)	0.73	0.17	+7	> .05
CAT Total Reading	Grade 1/ Cohort 3/ lowest 25%	10 schools/ 118 students	380.27 (100.78)	406.34 (92.48)	–26.07	–0.27	–11	> .05
DARD Silent Reading	Grade 1/ Cohort 3/ lowest 25%	10 schools/ 118 students	1.57 (2.88)	0.55 (1.83)	1.02	0.42	+16	> .05
Ross et al. (1998)[d]			**1 year of intervention**					
WRMT Passage Comprehension	Grade 1	6 schools/ 252 students	nr (nr)	nr (nr)	nr	0.01	0	> .05
Ross & Casey (1998a)[e]			**2 years of intervention**					
WRMT Passage Comprehension	Grade 1/ lowest 25%	7 schools/ 79 students	12.29 (7.79)	11.17 (8.03)	1.12	0.14	+6	> .05
Ross & Casey (1998b)[f]			**1 year of intervention**					
WRMT Passage Comprehension	Kindergarten/ lowest 25%	4 schools/ 69 students	1.50 (2.26)	1.00 (1.20)	0.50	0.26	+10	> .05
Ross et al. (1995)[g]			**4 years of intervention**					
Woodcock Reading Mastery Test (WRMT) Passage Comprehension	Grade 4/ lowest 25% Cohort 2	4 schools/ 19 students	30.28 (2.59)	25.51 (14.03)	4.77	0.49	+19	> .05
WRMT Passage Comprehension	Grades 3–4/ nonminority/ Cohorts 1–2	4 schools/ 81 students	–0.08 (0.92)	0.18 (1.30)	–0.26	–0.23	–9	> .05
4 years and 3 years of intervention								
WRMT Passage Comprehension	Grades 2–4/ minority/ Cohorts 1–3	4 schools/ 123 students	0.04 (0.70)	–0.22 (1.04)	0.28	0.31	+12	> .05

Table Notes: The supplemental findings presented in this table are additional findings 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.

[a] For Borman et al. (2007) second-year outcome, the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value and effect size presented here were reported in the original study.

For *first-year* outcomes, the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here and effect sizes (Cohen's *d*) were reported in the original study. Group means and standard deviations were obtained through the author query. The reported intervention group means are calculated as comparison groups means plus the HLM level-2 coefficient.

[b] For Quint et al. (2015), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value and effect size presented here were reported in the original study.

[c] For Madden et al. (1993) fifth-year outcomes, a correction for clustering was needed and resulted in a WWC-computed *p*-value of .27 for the *CTBS Reading Comprehension* outcome; therefore, the WWC does not find the result to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

For *second-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering were needed, and the WWC did not find the results to be statistically significant. The intervention and comparison group means reported in this table are ANCOVA-adjusted. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

For *first-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering were needed and the WWC did not find the results to be statistically significant. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

[d] For Ross et al. (1998), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value and the effect size presented here were reported in the original study.

[e] For Ross and Casey (1998a), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated by the WWC across participating schools The intervention and comparison group means reported in this table are MANCOVA-adjusted.

[f] For Ross and Casey (1998b), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are MANCOVA-adjusted.

[g] For Ross et al. (1995), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

Appendix D.3: Description of supplemental findings for the general reading achievement domain

Domain and utcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Madden et al. (1993)[a]			**4 years of intervention**					
Durrell Analysis of Reading Difficulty (DARD) Oral Reading	Grade 2/ lowest 25%/ Cohort 1	10 schools/ 104 students	7.20 (4.75)	2.44 (3.18)	4.76	1.19	+38	< .01
			3 years of intervention					
DARD Oral Reading	Grade 1/ lowest 25%/ Cohort 1	10 schools/ 126 students	4.35 (4.30)	1.81 (3.66)	2.54	0.63	+24	> .05
DARD Oral Reading	Grade 2/ Cohort 2	10 schools/ 440 students	11.99 (7.28)	8.84 (6.05)	3.15	0.47	+18	> .05
DARD Oral Reading	Grade 2/ lowest 25%/ Cohort 2	10 schools/ 112 students	6.04 (4.62)	3.32 (3.37)	2.72	0.67	+25	> .05
DARD Oral Reading	Grade 3/ lowest 25%/ Cohort 3	10 schools/ 104 students	12.92 (6.39)	8.08 (4.87)	4.85	0.85	+30	.02

Domain and utcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
2 years of intervention								
DARD Oral Reading	Grade 1/ Cohort 2	10 schools/ 338 students	6.01 (6.58)	4.84 (4.87)	1.17	0.20	+8	> .05
DARD Oral Reading	Grade 1/ lowest 25%/ Cohort 2	10 schools/ 86 students	1.42 (2.85)	1.48 (2.57)	–0.06	–0.02	–1	> .05
DARD Oral Reading	Grade 2/ Cohort 3	10 schools/ 320 students	11.85 (8.23)	8.60 (6.47)	3.25	0.44	+17	> .05
DARD Oral Reading	Grade 2/ lowest 25%/ Cohort 3	10 schools/ 78 students	6.82 (5.06)	4.72 (3.71)	2.10	0.47	+18	> .05
1 year of intervention								
DARD Oral Reading	Grade 1/ Cohort 3	14 schools/ 584 students	5.32 (4.07)	4.78 (3.91)	0.54	0.14	+5	> .05
DARD Oral Reading	Grade 1/ lowest 25%/ Cohort 3	10 schools/ 118 students	3.02 (3.08)	1.90 (2.45)	1.12	0.40	+16	> .05
Ross et al. (1998)[b]	**1 year of intervention**							
DARD Oral Reading	Grade 1	6 schools/ 252 students	nr (nr)	nr (nr)	nr	0.04	+2	> .05
Ross & Casey (1998a)[c]	**2 years of intervention**							
DARD Oral Reading	Grade 1/ lowest 25%	7 schools/ 79 students	4.14 (3.84)	3.18 (3.55)	0.96	0.26	+10	> .05
Ross & Casey (1998b)[d]	**1 year of intervention**							
DARD Oral Reading	Kindergarten/ lowest 25%	4 schools/ 69 students	0.20 (0.97)	0.00 (0.00)	0.20	0.26	+10	> .05
Ross et al. (1995)[e]	**4 years of intervention**							
DARD Oral Reading/Gray Oral Reading Test (GORT)	Grades 3–4/ nonminority/ Cohorts 1–2	4 schools/ 81 students	–0.11 (0.87)	0.36 (1.30)	–0.47	–0.43	–17	> .05
GORT	Grade 4/ lowest 25% Cohort 2	4 schools/ 19 students	75.86 (11.25)	78.90 (25.30)	–3.04	–0.16	–3	> .05
4 years and 3 years of intervention								
DARD Oral Reading/ GORT-3	Grades 2–4/ minority/ Cohorts 1	4 schools/ 123 students	–0.08 (0.80)	–0.23 (0.92)	0.15	0.18	+7	> .05
Skindrud & Gersten (2006)[f]	**2 years of intervention**							
Stanford Achievement Test (SAT-9) Language	Grade 3/ lowest 25% Cohort 1	12 schools/ 114 students	29.50 (10.20)	38.30 (14.50)	–8.80	–0.66	–25	< .01

Domain and utcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
SAT-9 Reading	Grade 3/ lowest 25% Cohort 1	12 schools/ 108 students	25.40 (14.20)	34.60 (13.10)	–9.20	–0.68	–25	< .01
1 year of intervention								
SAT-9 Language	Grade 2/ lowest 25% Cohort 1	12 schools/ 114 students	22.15 (11.30)	29.80 (16.30)	–7.30	–0.49	–19	< .01
SAT-9 Language	Grade 2 Cohort 1	12 schools/ 434 students	37.20 (16.80)	44.30 (17.10)	–7.10	–0.42	–16	< .01
SAT-9 Language	Grade 2/ lowest 25% Cohort 1	12 schools/ 108 students	25.80 (5.90)	33.60 (13.70)	–7.80	–0.66	–24	< .01

Table Notes: The supplemental findings presented in this table are additional findings that meet WWC design standards with 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.

[a] For Madden et al. (1993) *fourth-year* outcome, a correction for clustering was needed and resulted in a WWC-computed *p*-value of .001 for the *DARD Oral Reading* outcome; therefore, the WWC finds the result to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

For *third-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering and four multiple comparisons were needed and resulted in a WWC-computed critical *p*-value of .013 for the *DARD Oral Reading* measure for the Cohort 3 subgroup (p=.015); therefore, the WWC did not find the result for the *Cohort 3 subgroup* to be statistically significant. The intervention and comparison group means reported in this table are ANCOVA-adjusted. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

For *second-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering were needed, and the WWC did not find the results to be statistically significant. The intervention and comparison group means reported in this table are ANCOVA-adjusted. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

For *first-year* outcomes, the *p*-values presented here were calculated by the WWC. Corrections for clustering were needed and the WWC did not find the results to be statistically significant. Because study authors analyzed each matched pair of schools separately, the WWC combined means and standard deviations for the *SFA®* schools and for the comparison schools.

[b] For Ross et al. (1998), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value and the effect size presented here were reported in the original study.

[c] For Ross and Casey (1998a), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. Authors presented outcome statistics for each school separately. The intervention and comparison group means and standard deviations reported in this table are aggregated by the WWC across participating schools. The intervention and comparison group means reported in this table are MANCOVA-adjusted.

[d] For Ross and Casey (1998b), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. The intervention and comparison group means reported in this table are MANCOVA-adjusted.

[e] For Ross et al. (1995), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means reported in this table are ANCOVA-adjusted.

[f] For Skindrud and Gersten (2006), the *p*-values presented here were reported in the original study. Note that the authors did not conduct univariate statistical tests for all reported outcomes. For example, the two bottom quartile reading outcomes (in grade 2 and grade 3; Cohort 1) were jointly significant at $p < .001$. Similarly, the two bottom quartile language outcomes (in grade 2 and grade 3; Cohort 1) were jointly significant at $p < .001$. For *second-year* outcomes, the WWC found the results for the two analyses , SAT-9 Reading and SAT-9 Language, to be negative and statistically significant after the corrections for clustering and multiple comparisons adjustment were performed. For *first-year outcomes*, corrections for clustering and multiple comparisons were needed, and the WWC did not find the results to be statistically significant. The intervention and comparison group means reported in this table are ANCOVA-adjusted. The effect sizes reported here differ from those reported in the original study due to differences in the effect-size formulas used; WWC uses Hedges' *g* statistic, while the study appears to use the Cohen's *d* statistic to calculate effect sizes.

Endnotes

[1] The descriptive information for this intervention comes from publicly available sources: the program's website (http://www.successforall.org/) and from the previous WWC Beginning Reading intervention report of *SFA*® (https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_sfa_081109.pdf). The What Works Clearinghouse (WWC) requests developers review the intervention description sections for accuracy from their perspective. The WWC provided the developer with the intervention description in August 2014, and the WWC incorporated feedback from the developer. Further verification of the accuracy of the descriptive information for this intervention is beyond the scope of this review.

[2] The literature search reflects documents publicly available through March 1, 2016. Reviews of the studies in this report used the standards from the WWC Procedures and Standards Handbook (version 3.0) and the Beginning Reading review protocol (version 3.0). The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

This updated report includes reviews of 111 studies that the previous WWC intervention report for this intervention, released in August 2009, did not include. Of the additional studies, 102 were not within the scope of the review protocol for the Beginning Reading topic area, and seven were within the scope of the review protocol for the Beginning Reading but did not meet WWC group design standards. Two studies, Quint et al. (2015) and Tracey et al. (2014), meet WWC group design standards. A complete list and disposition of all studies reviewed are available in the references. The current report, which includes reviews of all previous studies, resulted in a revised disposition of five studies: Dianda and Flaherty (1995), Ross et al. (1998), Ross and Casey (1998b), Ross et al. (1999), and Skindrud and Gersten (2006).

Two studies currently do not meet WWC group design standards, whereas previously they met WWC evidence standards with reservations:

(1) The previous review of Dianda and Flaherty (1995) used version 1.0 WWC evidence standards; the current review used version 3.0 standards, which clarify that the study must demonstrate equivalence on the analytic sample. There was a discrepancy between the sample size reported for all students at baseline (n=319) and the sum of the baseline subsamples reported in the study (n=366). The authors did not respond to the WWC's request for clarification of the analytic samples or data that could demonstrate equivalence, so now the study does not meet WWC group design standards.

(2) The previous review of Ross et al. (1998) used version 1.0 WWC evidence standards; the current review used version 3.0 standards, which clarify that analysis of covariance (ANCOVA) that adjusts for pretest difference (of unknown magnitude) cannot demonstrate equivalence of the analytic sample. The authors did not provide data that could demonstrate equivalence in response to the WWC's request, so now the study does not meet WWC group design standards.

Three studies now meet WWC group design standards with reservations, whereas previously they did not meet WWC evidence standards:

(1) The previous rating for Ross and Casey (1998b) is based on the sample of first-grade students, whereas the current rating is based on the sample of kindergarteners. The current review confirmed the previous rating for first-grade students but revised the disposition for the kindergarten analysis. The previous rating for the kindergarten sample is based on the transformed Woodcock Reading Mastery Test (WRMT) posttest scores that did not meet WWC outcome reliability requirements. (The study authors assigned dichotomous scores to all outcome measures, with "0" indicating no correct responses and "1" indicating at least one correct response.) The current disposition is based on findings using original continuous WRMT scores (reported in Tables 2 and 4 of the study). The continuous WRMT scores meet WWC outcome reliability requirements because they are scaled scores based on established scoring procedures from a standardized test.

(2) The previous rating for Ross et al. (1999) is based on the full sample of students, whereas the current effectiveness rating is based on the following subgroups: minority students in grades 2 through 4 and nonminority students in grades 3 and 4. The current review confirmed the previous rating for the full sample but added dispositions for subgroup analyses that the previous report did not rate. (Note that Ross et al. [1999] is an additional source for the Ross et al. [1995] study featured in Appendix A.7 of this intervention report.)

(3) The previous review of Skindrud and Gersten (2006) used version 1.0 WWC evidence standards; the current review used version 3.0 standards, which clarify that the study must demonstrate equivalence on the analytic sample. The previous review found that the analytic intervention and comparison groups for the analysis of the Stanford Achievement Test, 9th Edition (SAT-9) Language Subtest did not demonstrate baseline equivalence and thus rated all analytic samples in the study on that basis. In this report, the WWC establishes baseline equivalence separately for the SAT-9 Reading Subtest full sample (n=434; Cohort 1; d=.21), low-achieving subgroups for both (Language and Reading) subtests, but not for the SAT-9 Language Subtest full sample (n=428; Cohort 1; d=.36). The current review confirmed the previous rating for the SAT-9 Language Subtest full sample analysis but added dispositions for the SAT-9 Reading Subtest full sample and subgroup analyses.

[3] Please see the Beginning Reading review protocol (version 3.0) for a list of all the outcome domains.

[4] For criteria used to determine the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 70. These improvement index numbers show the average and range of individual-level improvement indices for all findings across the studies.

[5] In addition to the analysis of the schoolwide outcomes, Borman et al. (2007) examined the intervention's effects on students in the longitudinal (that is, long-term) sample, which included students who were present in schools at the time of baseline and outcome assessments. In this analysis, which made inferences at the student level, the integrity of the study's random assignment was jeopardized because the study defined the student sample after school random assignment (that is, some students joined study schools after random assignment but before the baseline assessment). The impact analysis on these outcomes for the longitudinal sample does not meet WWC group design standards because the study did not establish baseline equivalence for the intervention and comparison groups.

[6] Within each study, the findings the WWC considered for the domain effectiveness rating are those measured at the period closest to the end of the intervention and that reflect the maximum exposure of students to the program. The Beginning Reading review protocol (version 2.1) documents this decision.

[7] For analyses of students in grades K–2, the study lost six schools to attrition (that is, the outcome variable is not available for all participants initially assigned to the intervention and comparison groups) and reduced the third-year analytic sample to 35 schools (Borman et al., 2007).

[8] In addition to analyzing the schoolwide outcomes, Quint et al. (2015) examined the effects of the intervention on students in the main sample, which included students who were present in schools at the time of baseline and outcome assessments. For this analysis, which made inferences at the student level, the integrity of the study's random assignment was jeopardized because the student sample was defined after school random assignment. The impact analysis on these outcomes does not meet WWC group design standards because the study did not establish baseline equivalence for the intervention and comparison groups.

[9] The dropout prevention *SFA*® model was named after a U.S. Department of Education dropout prevention grant, from which the three study schools received funds.

[10] For Madden et al. (1993), findings reported in Appendix C (primary findings) in the reading fluency domain reflect 5 years of program involvement for students in Cohort 2. Primary findings in the comprehension domain reflect 5, 4, and 2 years of *SFA*® involvement for students in Cohorts 2, 1, and 3, respectively. For the general reading achievement domain, primary findings reflect 5 years of program involvement for students in Cohort 2 and 3 years of program involvement for students in Cohorts 1 and 3.

[11] For Ross et al. (1995), the most recent study period reflects 4 years of program involvement for students in Cohorts 1 and 2. However, the most recent results that meet WWC group design standards reflect 3 years of *SFA*® involvement for minority students in Cohort 3, because analyses based on the full Cohort 3 sample did not meet WWC group design standards.

[12] For Skindrud and Gersten (2006), the most recent study period reflects 2 years of program involvement for students in Cohort 1. However, the latest, most recent results that meet WWC group design standards reflect 1 year of *SFA*® involvement for students in the lowest quartile from Cohort 2, because analyses based on the full Cohort 2 sample did not meet WWC group design standards.

[13] Madden et al. (1993) reported *p*-values (indicators of statistical significance) for five pairwise (matched) school comparisons at each grade level: 1, 2, and 3. For the WLPB Word Attack subtest, 13 (of 15) pairwise comparisons were positive and statistically significant, including all five comparisons in grade 1. For the WLPB Letter-Word subtest, 11 (of 15) pairwise comparisons were positive and statistically significant, including all five comparisons in grade 2 (Madden et al., 1993: pp. 134–139). Note that the WWC combined findings across schools because reported pairwise analyses did not meet WWC group design standards; because each condition has only a single school, it was impossible to separate the effect of the intervention from the effect of the schools on the findings.

[14] Note that the WWC adjusted for multiple comparisons within exposure level for each study domain.

[15] Ross and Casey (1998a: p. 19) reported *p*-values for univariate analyses (that is, analyses based on one variable) based on the three intervention schools, while the WWC excluded from its review results from one study school that supplemented *SFA*® with another branded intervention (*Reading Recovery*).

[16] Madden et al. (1993) reported *p*-values for five pairwise school comparisons at each grade level: 1, 2, and 3. For the DARD Oral Reading subtest, 11 (of 15) pairwise comparisons were positive and statistically significant, including all five comparisons in grade 3 (Madden et al., 1993: pp. 134–139). Note that the WWC combined findings across schools because the reported pairwise analyses did not meet WWC group design standards; because each condition had only a single school, it was impossible to separate the effect of the intervention from the effect of the schools on the findings.

[17] The WWC guidance (version 3.0) indicates that if the authors of a cluster randomized controlled trial study characterize the intervention as having effects on student scores (rather than only on cluster-level scores), and some students enter clusters after random assignment, then the study must demonstrate equivalence on the analytic sample.

[18] For Madden et al. (1993), combined results for the three *SFA*® versions (encompassing all eight *SFA*® schools) are reported in supplemental Appendices D for Cohort 3 students in grade 1 (after 1 year of intervention implementation). The *SFA*® sample included two full implementation schools, three dropout prevention schools, and three curriculum-only schools. The three *SFA*® versions varied in the number of personnel used to implement *SFA*®, particularly tutors and family support staff. Also, the curriculum-only schools had no facilitator.

[19] For Madden et al. (1993), one *SFA*® school (Abbottston Elementary) was matched with a comparison school on spring 1987 scores, and otherwise the matching was performed on fall 1988 scores.

[20] For Madden et al. (1993), although intervention students in Cohorts 1 and 2 were exposed to the intervention in pre-K and K, the baseline assessment was measured in the spring of kindergarten.

[21] For Ross, Alberg, McNelis, and Rakow (1998), an additional group ("cluster 2B") included one *SFA*® school and three comparison schools (one school used *Accelerated Schools* design, and the other two used locally developed programs). However, this comparison did not meet WWC group design standards because the effect of *SFA*® could not be separated from the effect of the single intervention school.

[22] Analyses of low-achieving students in grade 4 were ineligible for review under the Beginning Reading review protocol, version 3 (p. 4).

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2017, March). *Beginning Reading intervention report: Success for All®.* 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 analytic sample groups are similar on observed characteristics defined in the review area protocol.

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.

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. 70.

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.

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

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

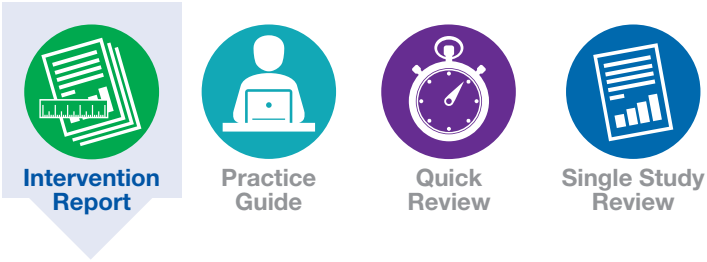
Statistical significance Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < .05$).

Study rating The result of the WWC assessment of a study. The rating is based on the strength of the evidence of the effectiveness of the educational intervention. Studies are given a rating of meets WWC design standards without reservations, meets WWC design standards with reservations, or does not meet WWC design standards, based on the assessment of the study against the appropriate design standards. The WWC has design standards for group design, single-case design, and regression discontinuity design studies.

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

Systematic review A review of existing literature on a topic that is identified and reviewed using explicit methods. A WWC systematic review has five steps: 1) developing a review protocol; 2) searching the literature; 3) reviewing studies, including screening studies for eligibility, reviewing the methodological quality of each study, and reporting on high quality studies and their findings; 4) combining findings within and across studies; and, 5) summarizing the review.

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.