Appendix

Appendix A1.1 Study characteristics: Borman, Slavin, Cheung, Chamberlain, Madden, & Chambers, 2006 (randomized controlled trial)

Characteristic	Description
Study citation	Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2006). *Final reading outcomes of the national randomized field trial of Success for All.* Retrieved from Success for All Web site: http://www.successforall.net/_images/pdfs/Third_Year_Results_06.doc
Participants	The study piloted the *SFA*® program in fall 2001, when three schools were randomly assigned to the *SFA*® and three schools 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 serve as comparisons.[1] The study presented findings after the intervention students completed one, two, and three years of the program. For the effectiveness ratings, the WWC focused on findings from the longitudinal sample, that is, schools and students who completed three years of the program.[2] After three years, 18 *SFA*® schools with 707 students and 17 comparison schools with 718 students remained in the longitudinal sample.
Setting	The analysis sample included 35 elementary schools across 14 states located in rural and small towns in the South and urban areas of the Midwest.
Intervention	Intervention students received the *SFA*® 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. Intervention schools implemented *SFA*® in grades K–2 and used their previously planned curriculum in grades 3–5. Some schools took a year to fully implement the program.
Comparison	Comparison schools continued using their regular, previously planned curriculum for grades K–2 (though *SFA*® was implemented in grades 3–5). Authors conducted observations at all schools and indicated that there was no evidence that when *SFA*® was implemented in grades 3–5, students in grades K–2 were also exposed to *SFA*®. All sample students were pretested with the Peabody Picture Vocabulary Test (PPVT) prior to *SFA*® implementation, and school-wide PPVT scores show equivalence between the program and comparison schools. Researchers also use information from the Common Core of Data (a database maintained by the National Center for Education Statistics) at several points over the course of the study to demonstrate the equivalence between the program and comparison schools on race/ethnicity, gender, English as a second language, special education, and free and reduced-price lunch. All equivalency tests were assessed at the school level and no statistically significant differences were found.
Primary outcomes and measurement	Three subtests of the Woodcock Reading Mastery Test were administered during the period reflected in the intervention rating: Word Identification, Word Attack, and Passage Comprehension.[3] (See Appendices A2.1–2.3 for more detailed descriptions of outcome measures.)
Teacher training	SFA® teachers received three days of training during the summer and approximately eight 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.

1. The 17 additional comparison schools implemented *SFA*® in grades 3–5 but students in grades K–2—the focus of this study and the WWC review—did not receive the intervention.
2. The study provided analysis for two samples, the "longitudinal sample" which included students who participated in the program for all three years, and the "in-mover sample" which included the longitudinal sample plus students who transferred into the school. The WWC analysis focuses on the longitudinal sample. The WWC prioritized outcomes that reflected students' exposure to the intervention for the longest period of time available. Findings reflecting students' outcomes after shorter periods of implementation can be found in Appendices A4.1–A4.9.
3. One additional subtest of the Woodcock Reading Mastery Test (Letter Identification) was administered during an earlier time period and is presented as an additional finding in Appendix A4.1

Appendix A1.2 Study characteristics: Dianda & Flaherty, 1995 (quasi-experimental design)

Characteristic	Description
Study citation	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.
Participants	This study involved seven elementary schools in California where the majority of students were English language learners. Six schools remained by the third year of program implementation. Students were grouped into four language categories and received instruction in English, Spanish, or "Sheltered English."[1] Only the English-speaking subsample was reviewed.[2] The report includes three cohorts of students who began participating in the study as kindergarteners in 1992 (99 intervention and 120 comparison students), 1993 (105 intervention and 62 comparison students), or 1994 (94 intervention and 59 comparison students), for a total of 539 participants. For the effectiveness rating, the WWC used data that reflected students' exposure to the intervention for the longest period of time, which varied for the different cohorts and domains.[3] Exact attrition rates are not known for this study, however the post-attrition intervention and comparison samples were equivalent for the English speaking subgroup. In the overall sample, the percent of students eligible for free lunch varied from 70 to 98 in intervention schools, and from 47 to 80 in comparison schools. The percentages of minority students were between 50 and 70 for each study condition.
Setting	The analysis sample included seven elementary schools in California.
Intervention	Intervention students received the typical *SFA*® curriculum 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.
Comparison	Comparison schools continued using their regular, previously planned curriculum. Each comparison school was matched with a *SFA*® school in the same district with students that had similar demographics and pretest scores on the Peabody Picture Vocabulary Test measure.
Primary outcomes and measurement	Three subtests of the Woodcock Language Proficiency Battery were administered: Letter-Word Identification, Word Attack, and Passage Comprehension. The authors presented findings from each Woodcock subtest separately and also pooled findings from the Woodcock Letter-Word Identification subtests (see Appendices A2.1–2.3 for more detailed descriptions of outcome measures).
Teacher training	*SFA*® teachers received three days of training during the summer and approximately eight 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. Specially trained certified teachers or qualified aides work one-to-one with the students.

1. English language learners participate in *SFA*® in English alongside their English-dominant classmates during a common period in the morning. During the rest of the day, they receive sheltered-content instruction or ESL instruction, depending on their level of English proficiency.
2. The WWC Beginning Reading topic focuses only on students learning to read in English (see Beginning Reading Protocol).
3. Findings include outcomes after two years of exposure for the alphabetics and comprehension domains; and after two (1994 cohort), three (1993 cohort), and four (1992 cohort) years of exposure for the general reading domain. Findings reflecting students' outcomes after shorter periods of implementation can be found in Appendix A4.3.

Appendix A1.3 Study characteristics: Madden, N. A., Slavin, R. E., Karweit, N., Dolan, L., & Wasik, B. A., 1993 (quasi-experimental design)

Characteristic	Description
Study citation	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.
Participants	The study investigated the effects of three versions of the *SFA*® program: full implementation, curriculum only,[1] and dropout prevention.[2] The WWC focused on the full implementation portion of the study, which included two intervention schools and two matched comparison schools. Within each comparison school, one third of the students were randomly selected for testing purposes. The study focused on cohorts of students who started *SFA*® in pre-kindergarten, kindergarten, and first grade and received the intervention for multiple years. To determine the effectiveness ratings, the WWC focused on the latest term results available. The third-year analytic sample included 268 students within two *SFA*® schools and 268 students within two comparison schools spread across three grade levels.[3] African-American students constituted 97–99% of students in two intervention schools, with 83–97% of students qualified for free lunch. In comparison (Chapter 1) schools, at least 75% of students qualified for free lunch.
Setting	The analysis sample included four elementary schools in Baltimore, Maryland.
Intervention	Intervention students 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 facilitator who worked with school personnel, and training for all intervention teachers.
Comparison	The comparison condition included schools that implemented a traditional reading program built around Macmillan Connections basal series. Each comparison school was matched with an intervention school based on the percentage of students getting free or reduced-price lunch and historical achievement level. Students were then individually matched on a standardized test given by the school district. Pretest scores on WRMT Letter-Word Identification, Word Attack, and Durrell Oral Reading subtests served as covariates in analyses.
Primary outcomes and measurement	Two subtests of the Woodcock Language Proficiency Battery were administered: Letter-Word Identification and Word Attack. Additional measures included Durrell Analysis of Reading Difficulty Silent Reading and Oral Reading subtests and the California Achievement Test (CAT) Total Reading (see Appendices A2.1–2.3 for more detailed descriptions of outcome measures).
Teacher training	The teachers and tutors were regular certified teachers. They received detailed teacher's manuals supplemented by two to three days of in-service 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 teachers and aids were trained in the use of the thematic units, and other aspects of the preschool and kindergarten models. School facilitators also organized many information sessions to allow teachers to share problems and solutions, suggest changes, and discuss individual children.

1. The curriculum-only portion (a version of the *SFA*® program that only uses the beginning reading curriculum rather than the whole school reform) of the study included only one school in the comparison condition making it impossible to separate the effect of the school from the effect of the regular reading curriculum.
2. The dropout prevention version was designed to operate within schools that do not have the funding to implement the full *SFA*® program. The dropout prevention program had a reduced number of tutors and family support staff. Chapter 1 monies supported the program. The dropout prevention portion is not included in the intervention rating because it differs from the standard implementation of the program. However, findings for the dropout prevention portion of *SFA*® can be found in Appendices A4.7–4.9
3. Additional findings reflecting students' outcomes after shorter periods of implementation can be found in Appendices A4.1–A4.9, along with findings for a subsample of low-achieving students.

Appendix A1.4 Study characteristics: Ross, Alberg, & McNelis, 1997 (quasi-experimental design)

Characteristic	Description
Study citation	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: The University of Memphis, Center for Research in Education Policy.
Participants	The study compared whole-school improvement programs, *Success for All*®, Accelerated Schools, and locally-developed programs, in 19 schools. Schools were divided into four groups based on the similarity of several school characteristics, including enrollment, percentage of minority students, percentage of students eligible for free/reduced lunch, and initial academic performance. WWC focused on only one group, "cluster 2A", the third highest with respect to socio-economic status, which included three *SFA*® schools and three Accelerated Schools, with a total number of 252 first-grade students (148 students that attended *SFA*® schools; 104 students that attended Accelerated Schools).[1] The study included data that reflected students' outcomes after one year of program implementation. In the overall sample, the percent of minority students in three intervention schools was between 47 and 63. In three the comparison schools, the range was between 42 and 54%. The percent of students eligible for free/reduced lunch varied from 63 to 66 in intervention schools, and from 66 to 71 in comparison schools.
Setting	The analysis sample included six elementary schools in Clover Park, Washington.
Intervention	Intervention students 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 facilitator who worked with school personnel, and training for all intervention teachers.
Comparison	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 so that they provide enrichment for at-risk students.
Primary outcomes and measurement	Three subtests of the Woodcock Reading Mastery Test were administered: Word Identification, Word Attack, and Passage Comprehension. The Durrell Analysis of Reading Difficulty Oral Reading subtest was also used (see Appendices A2.1–2.3 for more detailed descriptions of outcome measures).
Teacher training	No information on training for the specific teachers in this study was provided.

1. An additional group included one *SFA*® school and three comparison schools (one school used Accelerated Schools design, and the other two locally developed programs), but this comparison did not meet WWC evidence screens because the effect of *SFA*® cannot be separated from the effect of that school.

Appendix A1.5 Study characteristics: Ross & Casey, 1998 (quasi-experimental design)

Characteristic	Description
Study citation	Ross, S. M., & Casey, J. (1998). *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: The University of Memphis, Center for Research in Education Policy.
Participants	This study examines the effects of *SFA*® in two Title I schools by comparing them with five other Title I schools that were implementing locally developed school-wide programs.[1] The study did not report on the initial sample size, but 288 students in kindergarten (83 students in the *SFA*® schools; 205 students at comparison schools) were included in the final analysis sample and the post-attrition intervention and comparison samples were equivalent on the achievement pretest measure (PPVT). The study included data that reflected students' outcomes after two years of program implementation.[2] School populations ranged between 31 and 50% minority enrollment; between 62 and 81% of students received free or reduced-price lunch.
Setting	The analysis sample included seven Title I elementary schools in Fort Wayne, Indiana.
Intervention	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	The five comparison schools implemented locally developed school-wide programs. The schools were comparable with *SFA*® schools on pretest PPVT measures, socio-economic status, and ethnicity. Four out of the five local school programs incorporate 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 multi-faceted reading activities.
Primary outcomes and measurement	Three subtests of the Woodcock Reading Mastery Test were administered: Word Identification, Word Attack, and Passage Comprehension. The study presented a combined measure of Word Identification and Word Attack. The Durrell Analysis of Reading Difficulty Oral Reading subtest was also used (see Appendices A2.1–2.3 for more detailed descriptions of outcome measures).
Teacher training	No information on training for the specific teachers was provided in this study.

1. The article reported on an additional intervention school that supplemented *SFA*® with another branded intervention (*Reading Recovery*), but results from this portion of the study do not meet WWC evidence standards because the effect of *SFA*® cannot be separated from the effect of *Reading Recovery.*
2. Additional findings for a subsample of low-achieving students (i.e., lowest 25% with respect to reading achievement) are reported in Appendices A4.1–A4.9.

Appendix A1.6 Study characteristics: Ross, McNelis, Lewis, & Loomis, 1998 (quasi-experimental design)

Characteristic	Description
Study citation	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: The University of Memphis, Center for Research in Education Policy.
Participants	This study involved 97 first-grade students with both pretest and posttest data in four schools. Two schools implemented the *Success for All*® program (40 students) and two schools were selected as their matched comparison schools (47 students). The *SFA*® schools and the comparison schools were similar in poverty level, achievement level, and enrollment. The study reported data on students' outcomes after one year of program implementation.
Setting	The study took place in four elementary schools in Little Rock, Arkansas.
Intervention	Intervention students 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 facilitator who worked with school personnel, and training for all intervention teachers.
Comparison	No information was provided on the nature of the comparison curriculum. The two comparison schools were matched to the *SFA*® schools based on poverty level, achievement level, and enrollment. Pretest PPVT scores were used as a covariate to adjust for differences in students' abilities.
Primary outcomes and measurement	Three subtests of the Woodcock Reading Mastery Test were administered: Word Identification, Word Attack, and Passage Comprehension. The Durrell Analysis of Reading Difficulty Oral Reading subtest was also used (see Appendices A2.1–2.3 for more detailed descriptions of outcome measures).
Teacher training	No information on training for the teachers in this study was provided.

Appendix A1.7 Study characteristics: Smith, Ross, Faulks, Casey, Shapiro, & Johnson, 1993 (quasi-experimental design)

Characteristic	Description
Study citation	Smith, L. J., Ross, S. M., Faulks, A., Casey, J., Shapiro, M., & Johnson, B. (1993). 1991-1992 Ft. Wayne, Indiana *SFA* results. Memphis, TN: The University of Memphis, Center for Research in Education Policy.
Participants	This study involved approximately 286 students in kindergarten and first grade in four elementary schools in Fort Wayne, Indiana. Two schools implemented the *SFA*® program. Two comparison schools were matched to the intervention schools based on poverty level, historical achievement level, and ethnicity; then pairs of students were matched on PPVT pretest scores. There were 74 kindergarteners and 69 first-grade students in the intervention group and 74 kindergarteners and 69 first-grade students in the comparison group. Exact student attrition rates are not known for this study; however, the post-attrition intervention and comparison samples were equivalent on achievement pretest. School level data—poverty level, achievement, and enrollment—were similar across all schools. The study included data on students' outcomes after one year of program implementation.[1]
Setting	The study took place in four elementary schools in Fort Wayne, Indiana.
Intervention	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.
Comparison	Comparison schools continued using their regular, previously planned curriculum. No other information was provided on the comparison curriculum.
Primary outcomes and measurement	Four subtests of the Woodcock Reading Mastery Test were used: Letter Identification, Word Identification, Word Attack, and Passage Comprehension. Additional measures included the Peabody Picture Vocabulary Test and Durrell Analysis of Reading Difficulty Oral Reading subtest. The Merrill Language Screening Test and the Test of Language Development were also administered, but have not been included in this review because they were outside the scope of the Beginning Reading review (see Appendices A2.1–2.3 for more detailed descriptions of outcome measures).
Teacher training	Teachers in their first year of teaching *SFA*® classes received three days of summer training and two to four 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-services, an annual *SFA*® conference, and implementation checks for the facilitators and trainers.

1. Additional findings for a low-achieving subset of students (lowest 25% with respect to reading achievement) are presented in Appendices A41–A4.9.

Appendix A2.1 Outcome measures in the alphabetics domain by construct

Outcome measure	Description
Letter knowledge	
Woodcock Reading Mastery Test (WRMT): Letter Identification subtest	The standardized test measures the number of letters that students are able to identify correctly (Smith et al., 1993).
Phonics	
WRMT: Word Identification subtest	The Word Identification subtest is a test of decoding skills. The standardized test requires the child to read aloud isolated real words that range in frequency and difficulty (as cited in Borman et al., 2006; Ross & Casey, 1998; Ross, Alberg, & McNelis, 1997; Ross et al., 1998; Smith et al., 1993).
Woodcock Language Proficiency Battery (WLPB): Letter-Word Identification subtest	The Letter/Word Identification subtest is a standardized test that requires the child to read aloud isolated letters and real words that range in frequency and difficulty (as cited in Dianda & Flaherty, 1995, and Madden et al., 1993).
WRMT and WLPB: Word Attack subtest	The standardized test measures phonemic decoding skills by asking students to read pseudowords. Students are aware that the words are not real (as cited in Borman et al., 2006; Dianda & Flaherty, 1995; Ross & Casey, 1998; Ross, Alberg, & McNelis, 1997; Ross et al., 1998; Madden et al., 1993; Smith et al., 1993).

Appendix A2.2 Outcome measures in the comprehension domain by construct

Outcome measure	Description
Reading comprehension	
WRMT and WLPB: Passage Comprehension subtest	In this standardized test, comprehension is measured by having students fill in missing words in a short paragraph (as cited in Borman et al., 2006; Dianda & Flaherty, 1995; Ross & Casey, 1998; Ross, Alberg, & McNelis, 1997; Ross et al., 1998; Smith et al., 1993).
Durrell Analysis of Reading Difficulty (DARD): Silent Reading Test	An individually-administered, standardized diagnostic test that measures reading rate while students read passages silently and answer comprehension questions (as cited in Madden et al., 1993).
Vocabulary development	
Peabody Picture Vocabulary Test (PPVT)	A standardized, receptive vocabulary test that asks students to choose which one of four pictures corresponds to a test word spoken aloud (as cited in Smith et al., 1993).

Appendix A2.3 Outcome measures in the general reading domain by construct

Outcome measure	Description
California Achievement Test (CAT) Total Reading	A group-administered, standardized assessment battery comprised of numerous reading and language-oriented subtests (as cited in Madden et al., 1993).
DARD Oral Reading Test	An individually administered, standardized diagnostic test that measures reading accuracy, reading rate, and oral reading comprehension (as cited in Ross, Albert, & McNelis, 1997; Ross & Casey, 1998; Ross et al., 1998; Madden et al., 1993; Smith et al., 1993).

Appendix A3.1 Summary of findings for all domains[1]

Outcome measure	Domain: Alphabetics – Letter knowledge	Alphabetics – Phonics	Comprehension – Reading comprehension	Comprehension – Vocabulary development	General reading achievement
Met evidence standards					
Borman et al., 2006	nr	+	+	nr	nr
Met evidence standards with reservations					
Dianda & Flaherty, 1995	nr	(+)	(+)	nr	(+)
Madden et al., 1993	nr	(+)	nr	nr	(+)
Ross, Alberg, & McNelis, 1997	nr	ind	ind	nr	ind
Ross & Casey, 1998	nr	ind	ind	nr	ind
Ross et al., 1998	nr	(+)	ind	nr	ind
Smith et al., 1993	(+)	(+)	(+)	ind	(+)
Rating of effectiveness	Potentially positive		Mixed effects		Potentially positive

nr = no reported outcomes under this construct

+ = study average finding was positive and statistically significant

(+) = study average finding was positive and substantively important, but not statistically significant

ind = study average finding was indeterminate, that is, neither substantively important nor statistically significant

1. This appendix reports summary findings of study averages that were considered for the effectiveness rating and the improvement index in each domain. More detailed information on findings for all measures within the domains and the constructs that factor into the domains can be found in Appendices A3.2–A3.4.

Appendix A3.2 Summary of findings for alphabetics domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Borman et al., 2006 (randomized controlled trial)[8]—Three years of intervention									
WRMT: Word ID subtest[9]	Phonics	Kindergarten	35/1,425	462.96 (23.56)	457.41 (25.72)	5.55	0.22	Statistically significant	+9
WRMT: Word Attack subtest[9]	Phonics	Kindergarten	35/1,425	493.43 (16.45)	487.73 (17.64)	5.70	0.33	Statistically significant	+13
Madden et al., 1993 (quasi experimental design)[8, 10]—Three years of intervention									
WLPB): Letter-Word ID subtest	Phonics	Pre-kindergarten (Cohort 1)	4/210	18.25 (5.20)	16.10 (6.69)	2.14	0.36	ns	+ 14
WLPB: Word Attack subtest	Phonics	Pre-kindergarten (Cohort 1)	4/210	5.41 (4.25)	2.29 (3.55)	3.12	0.79	ns	+ 29
WLPB: Letter-Word ID subtest	Phonics	Kindergarten (Cohort 2)	4/148	24.50 (5.93)	21.08 (6.61)	3.42	0.54	ns	+21
WLPB: Word Attack subtest	Phonics	Kindergarten (Cohort 2)	4/148	7.74 (6.00)	5.67 (4.69)	2.08	0.38	ns	+15
WLPB: Letter-Word ID subtest	Phonics	Grade 1 (Cohort 3)	4/178	28.09 (7.30)	25.28 (5.97)	2.81	0.42	ns	+16
WLPB: Word Attack subtest	Phonics	Grade 1 (Cohort 3)	4/178	11.47 (7.40)	6.52 (4.87)	4.95	0.79	ns	+18
Dianda & Flaherty, 1995 (quasi experimental design)[8]—Two years of intervention									
WLBP: Letter-Word ID subtest	Phonics	English-speaking kindergarten (1992 cohort)	7/219	nr	nr	na	0.34[11]	ns	+13
WLBP: Word Attack subtest	Phonics	English-speaking kindergarten (1992 cohort)	7/219	nr	nr	na	0.26[11]	ns	+10

(continued)

Appendix A3.2 Summary of findings for alphabetics domain[1] *(continued)*

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Ross & Casey, 1998 (quasi experimental design)[8]—Two years of intervention									
WRMT: Word ID subtest	Phonics	Kindergarten	7/288	32.14 (14.63)	31.30 (14.20)	0.84	0.06	ns	+2
WRMT: Word Attack subtest	Phonics	Kindergarten	7/288	12.25 (7.36)	10.40 (8.20)	1.85	0.23	ns	+9
Ross, Alberg, & McNelis, 1997 (quasi experimental design)[8]—One year of intervention									
WRMT: Word ID subtest	Phonics	Grade 1	6/252	nr	nr	na	–0.01[12]	ns	0
WRMT: Word Attack subtest	Phonics	Grade 1	6/252	18.35	15.86	2.49 (8.89)[13]	0.28[12]	ns	+11
Ross et al., 1998 (quasi experimental design)[8]—One year of intervention									
WRMT) Word ID subtest	Phonics	Grade 1	4/97	38.27	36.21	2.06 (12.31)[14]	0.17	ns	+7
WRMT: Word Attack subtest	Phonics	Grade 1	4/97	15.17	11.19	3.98 (8.89)[14]	0.44	ns	+17
Smith et al., 1993 (quasi experimental design)[8]—One year of intervention									
WRMT: Word ID subtest	Phonics	Kindergarten (Cohort 1)	4/148	10.26 (9.82)	3.15 (4.95)	7.11	0.91	ns	+32
WRMT: Letter ID subtest	Letter Knowledge	Kindergarten (Cohort 1)	4/148	32.43 (4.28)	29.36 (7.81)	3.07	0.48	ns	+19
WRMT: Letter ID subtest[9]	Letter Knowledge	Grade 1 (Cohort 2)	4/138	nr	nr	na	0.08[11]	ns	+3
WRMT: Word ID subtest	Phonics	Grade 1 (Cohort 2)	4/138	35.04 (10.63)	28.00 (14.70)	7.04	0.55	ns	+21
WRMT: Word Attack subtest	Phonics	Grade 1 (Cohort 2)	4/138	12.60 (7.43)	7.90 (7.91)	4.70	0.61	ns	+23

(continued)

Appendix A3.2 Summary of findings for alphabetics domain[1] *(continued)*

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Averages for alphabetics[15]									
Borman et al., 2006—Three years of intervention							0.28	Statistically significant	+11
Madden et al., 1993—Three years of intervention							0.55	ns	+21
Dianda & Flaherty, 1995—Two years of intervention							0.30	ns	+12
Ross & Casey, 1998—Two years of intervention							0.14	ns	+6
Ross, Alberg, & McNelis, 1997—One year of intervention							0.13	ns	+5
Ross et al., 1998—One year of intervention							0.31	ns	+12
Smith et al., 1993—One year of intervention							0.56	ns	+21
Domain average for alphabetics across all studies							0.32	na	+13
Averages by years of *SFA®* implementation									
Average of results from studies with three years of intervention (two studies)							0.38	na	+15
Average of results from studies with two years of intervention (two studies)							0.22	na	+9
Average of results from studies with one year of intervention (three studies)							0.33	na	+13

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

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. Earlier findings from longitudinal studies are not included in these ratings, but are reported in Appendix A4.1. Subgroup findings from the studies are not included in these ratings, but are reported in Appendix A4.4.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Borman et al. (2006), a correction for multiple comparisons was needed so the significance levels may differ from those reported in the original study. There was no need to adjust for clustering because the findings were based on HLM analyses. In the case of the six other studies, corrections for both clustering and multiple comparisons were needed so the significance levels may differ from those reported in the original studies.

(continued)

Appendix A3.2 Summary of findings for alphabetics domain[1] *(continued)*

9. Standard deviations and adjusted means have been received through communication with the author (G. Borman, personal communication, 2006).
10. WWC combined means and standard deviations for two *SFA*® schools (Abbottston and City Springs) and their counterparts. Adjusted posttest means (with pretests standard scores as covariates) were used for effect size calculations. Kindergarten and grade 1 cohorts from Abbottston elementary school received four years of intervention.
11. Authors reported effect sizes that used comparison group standard deviation in the denominator (Glass's delta). Effect size was computed by subtracting the comparison group mean from the intervention group mean and dividing the result by the comparison group standard deviation.
12. Authors reported effect sizes adjusted for PPVT pretest scores.
13. The WWC derived pooled standard deviation from the reported means and effect size.
14. Authors reported pooled standard deviation.
15. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.

Appendix A3.3 Summary of findings for comprehension domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Borman et al., 2006 (randomized controlled trial)[8]—Three years of intervention									
WRMT: Passage Comprehension subtest[9]	Reading comprehension	Kindergarten	35/1,425	481.41 (14.20)	478.33 (15.33)	3.08	0.21	Statistically significant	+8
Dianda & Flaherty, 1995 (quasi-experimental design)[8]—Two years of intervention									
WLPB: Passage Comprehension subtest	Reading comprehension	English-speaking kindergarten (1992 cohort)	7/219	nr	nr	na	0.44	ns	+17
Ross & Casey, 1998 (quasi-experimental design)[8]—Two years of intervention									
WRMT: Passage Comprehension subtest	Reading comprehension	Kindergarten	7/288	16.09 (8.46)	15.40 (8.70)	0.69	0.08	ns	+3
Ross, Alberg, & McNelis, 1997 (quasi-experimental design)[8]—One year of intervention									
WRMT: Passage Comprehension subtest	Reading comprehension	Grade 1	6/252	nr	nr	na	0.01[11]	ns	0
Ross et al., 1998 (quasi-experimental design)[8]—One year of intervention									
WRMT: Passage Comprehension subtest	Reading comprehension	Grade 1	4/97	19.19	17.73	1.46 (8.19)[12]	0.18	ns	+7
Smith et al., 1993 (quasi-experimental design)[8]—One year of intervention									
Peabody Picture Vocabulary Test	Vocabulary development	Kindergarten (Cohort 1)	4/148	nr	nr	na	0.17[10]	ns	+7
WRMT: Passage Comprehension subtest	Reading comprehension	Grade 1 (Cohort 2)	4/136	16.37 (8.07)	13.91 (9.31)	2.46	0.28	ns	+11
Averages for comprehension[13]									
Borman et al., 2006—Three years of intervention							0.21	Statistically significant	+8
Dianda & Flaherty, 1995—Two years of intervention							0.44	ns	+17
Ross & Casey, 1998—Two years of intervention							0.08	ns	+3

(continued)

Appendix A3.3 Summary of findings for comprehension domain[1] *(continued)*

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Ross, Alberg, & McNelis, 1997—One year of intervention							0.01	ns	0
Ross et al., 1998—One year of intervention							0.18	ns	+7
Smith et al., 1993—One year of intervention							0.23	ns	+9
Domain average for comprehension across all studies							0.19	na	+8
Averages by years of *SFA®* implementation:									
Results from study with three years of intervention (one study)							0.21	Statistically significant	+8
Average of results from studies with two years of intervention (two studies)							0.26	na	+10
Average of results from studies with one year of intervention (three studies)							0.14	na	+6

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

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. Earlier findings from longitudinal studies are not included in these ratings, but are reported in Appendix A4.2. Subgroup findings from the studies are not included in these ratings, but are reported in Appendix A4.5
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Borman et al. (2006), there was no need to adjust for clustering because the findings were based on HLM analyses. In the case of Dianda and Flaherty (1995), Ross & Casey (1998), Ross, Alberg, & McNelis (1997), and Ross et al. (1998), a correction for clustering was needed so the significance levels may differ from those reported in the original study. In the case of Smith et al. (1993), correction for both clustering and multiple comparisons were needed so the significance levels may differ from those reported in the original studies.
9. Standard deviations and adjusted means have been received through communication with the author.
10. Authors reported effect sizes that used comparison group standard deviation in the denominator (Glass's delta). Effect size was computed by subtracting the comparison group mean from the intervention group mean and dividing the result by the comparison group standard deviation.
11. Authors reported effect sizes adjusted for PPVT pretest scores.
12. Authors reported pooled standard deviation.
13. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.

Appendix A3.4 Summary of findings for general reading achievement domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Dianda & Flaherty, 1995 (quasi-experimental design)[8, 9]									
Four years of intervention									
3 WLPB subtests and Durrell Reading subtest combined	General reading	English-speaking kindergarten (1992 cohort)	6/136	nr	nr	na	0.23[10]	ns	+9
Three years of intervention									
3 WLPB subtests and Durrell Reading subtest combined	General reading	English-speaking kindergarten (1993 cohort)	6/167	nr	nr	na	0.34[10]	ns	+13
Two years of intervention									
3 WLPB subtests and Durrell Reading subtest combined	General reading	English-speaking kindergarten (1994 cohort)	6/153	nr	nr	na	0.27[10]	ns	+11
Madden et al., 1993 (quasi-experimental design)[9, 11]—Three years of intervention									
Durrell Oral Reading subtest	General reading	Pre-kindergarten (Cohort 1)	4/210	5.45 (4.73)	4.46 (5.58)	0.99	0.19	ns	+8
Durrell Oral Reading subtest	General reading	Kindergarten (Cohort 2)	4/148	12.35 (7.77)	8.51 (5.06)	3.84	0.58	ns	+22
Durrell Oral Reading subtest	General reading	Grade 1 (Cohort 3)	4/178	16.74 (7.07)	12.92 (6.99)	3.82	0.54	ns	+21
Ross & Casey, 1998 (quasi-experimental design)[9]—Two years of intervention									
Durrell Oral Reading subtest	General reading	Kindergarten	7/288	5.35 (4.63)	4.7 0 (4.30)	0.65	0.15	ns	+6
Ross, Alberg, & McNelis, 1997 (quasi-experimental design)[9]—One year of intervention									
Durrell Oral Reading subtest	General reading	Grade 1	6/252	nr	nr	na	0.04[12]	ns	+2

(continued)

Appendix A3.4 Summary of findings for general reading achievement domain[1] *(continued)*

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Ross et al., 1998 (quasi-experimental design)[9]—One year of intervention									
Durrell Oral Reading subtest	General reading	Grade 1	4/97	7.01	6.46	0.55 (3.52)[13]	0.16	ns	+6
Smith et al., 1993 (quasi-experimental design)[9]—One year of intervention									
Durrell Oral Reading subtest	General reading	Grade 1	4/138	6.74 (4.25)	4.68 (3.83)	2.06	0.51	ns	+19
Averages for general reading achievement[14]									
Dianda & Flaherty, 1995[10]—Two to four years of intervention							0.28	ns	+11
Madden et al., 1993—Three years of intervention							0.44	ns	+17
Ross & Casey, 1998—Two years of intervention							0.15	ns	+6
Ross, Alberg, & McNelis, 1997—One year of intervention							0.04	ns	+2
Ross et al., 1998—One year of intervention							0.16	ns	+6
Smith et al., 1993—One year of intervention							0.51	ns	+19
Domain average for general reading achievement across all studies							0.26	na	+10
Averages by years of *SFA®* implementation									
Results from study with four year of intervention (one study)							0.23	ns	+9
Average of results from studies with three years of intervention (two studies)							0.39	na	+15
Average of results from studies with two years of intervention (two studies)							0.21	ns	+8
Average of results from studies with one year of intervention (three studies)							0.24	na	+9

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

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. Earlier findings from longitudinal studies are not included in these ratings, but are reported in Appendix A4.3. Subgroup findings from the studies are not included in these ratings, but are reported in Appendix A4.6
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.

(continued)

Appendix A3.4 Summary of findings for general reading achievement domain[1] *(continued)*

6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. Data are taken from Livingston & Flaherty (1997).
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Dianda & Flaherty (1995), Madden et al. (1993), and Smith et al. (1993), a correction for clustering and multiple comparisons was needed so the significance levels may differ from those reported in the original study. In the case of Ross & Casey (1998), Ross, Alberg, & McNelis (1997), and Ross et al. (1998), a correction for clustering was needed so the significance levels may differ from those reported in the original study.
10. Authors reported effect sizes that used comparison group standard deviation in the denominator (Glass's delta). Effect size was computed by subtracting the comparison group mean from the intervention group mean and dividing the result by the comparison group standard deviation.
11. WWC combined means and standard deviations for two *SFA*® schools (Abbottston and City Springs) and their counterparts. Adjusted posttest means (with pretests standard scores as covariates) were used for effect size calculations. Kindergarten and grade 1 cohorts from Abbottston elementary school received four years of intervention.
12. Authors reported effect sizes adjusted for PPVT pretest scores.
13. Authors reported pooled standard deviation.
14. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.

Appendix A4.1 Summary of earlier findings from longitudinal studies for alphabetics domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Borman et al., 2006 (randomized controlled trial)[8]—Two years of intervention									
WRMT: Letter ID subtest	Letter knowledge	Kindergarten and Grade 1	38/3,353	451.42 (14.08)	449.46 (11.19)	1.96	0.15	ns	+6
WRMT: Word ID subtest	Phonics	Kindergarten and Grade 1	38/3,353	449.52 (28.31)	444.82 (29.18)	4.70	0.16	ns	+6
WRMT: Word Attack subtest	Phonics	Kindergarten and Grade 1	38/3,353	487.92 (18.20)	483.29 (19.82)	4.63	0.24	Statistically significant	+10

ns = not statistically significant

1. This appendix presents earlier longitudinal findings for measures that fall in the alphabetics domain. Data that reflected students' exposure to the intervention for the longest period of time were used for intervention rating purposes and are presented in Appendix A3.2.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not applied to findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Borman et al. (2006), there was no need to adjust for clustering because the data were based on HLM analyses.

Appendix A4.2 Summary of earlier findings from longitudinal studies for comprehension domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Borman et al., 2006 (randomized controlled trial)[8]—Two years of intervention									
WRMT: Passage Comprehension subtest	Reading comprehension	Kindergarten and Grade 1	38/3,353	472.00 (18.29)	469.87 (19.53)	2.13	0.11	ns	+4

ns = not statistically significant

1. This appendix presents earlier longitudinal findings for measures that fall in comprehension domain. Data that reflected students' exposure to the intervention for the longest period of time were used for intervention rating purposes and are presented in Appendix A3.3.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Borman et al. (2006), there was no need to adjust for clustering because the findings were based on HLM analyses.

Appendix A4.3 Summary of earlier findings from longitudinal studies for general reading achievement domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Dianda & Flaherty, 1995 (quasi-experimental design)[8, 9]									
Three years of intervention									
3 WLPB subtests and Durrell Reading subtest combined	General reading	English-speaking kindergarten (1992 cohort)	6/136	nr	nr	na	0.44[10]	ns	+17
Two years of intervention									
3 WLPB subtests and Durrell Reading subtest combined	General reading	English-speaking kindergarten (1993 cohort)	6/167	nr	nr	na	0.87[10]	Statistically significant	+31

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

1. This appendix presents earlier longitudinal findings for measures that fall in general reading domain. Data that reflected students' exposure to the intervention for the longest period of time were used for intervention rating purposes and are presented in Appendix A3.4.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. Data are taken from Livingston & Flaherty (1997).
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Dianda & Flaherty (1995), a correction for clustering was needed so the significance levels may differ from those reported in the original study.
10. Authors reported effect sizes that used comparison group standard deviation in the denominator (Glass's delta). Effect size was computed by subtracting the comparison group mean from the intervention group mean and dividing the result by the comparison group standard deviation.

Appendix A4.4 Summary of subgroup findings for alphabetics domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Madden et al., 1993 (quasi-experimental design)[8, 9]—Three years of intervention									
WLPB: Letter-Word ID subtest	Phonics	Pre-kindergarten/ lowest 25% (Cohort 1)	4/54	16.37 (4.88)	10.86 (5.72)	5.51	1.02	ns	+35
WLPB: Word Attack subtest	Phonics	Pre-kindergarten/ lowest 25% (Cohort 1)	4/54	4.55 (4.44)	0.78 (2.41)	3.78	1.04	ns	+35
WLPB: Letter-Word ID subtest	Phonics	Kindergarten/lowest 25% (Cohort 2)	4/38	21.05 (4.54)	14.47 (6.34)	6.58	1.17	Statistically significant	+38
WLPB: Word Attack subtest	Phonics	Kindergarten/lowest 25% (Cohort 2)	4/38	5.21 (3.26)	1.84 (2.48)	3.37	1.14	ns	+37
WLPB: Letter-Word ID subtest	Phonics	Grade 1/lowest 25% (Cohort 3)	4/44	24.14 (7.06)	20.73 (4.87)	3.41	0.55	ns	+21
WLPB: Word Attack subtest	Phonics	Grade 1/lowest 25% (Cohort 3)	4/44	8.27 (7.18)	2.86 (3.93)	5.41	0.92	ns	+32
Ross & Casey, 1998 (quasi-experimental design)[9]—Two years of intervention									
WRMT: Word ID subtest	Phonics	Kindergarten/ lowest 25%	7/79	27.10 (14.25)	25.10 (13.40)	2.00	0.15	ns	+6
WRMT: Word Attack subtest	Phonics	Kindergarten/ lowest 25%	7/79	10.11 (6.13)	7.80 (8.10)	2.31	0.30	ns	+12
Smith et al., 1993 (quasi-experimental design)[9]—One year of intervention									
WRMT: Letter ID subtest	Letter Knowledge	Kindergarten/lowest 25% (Cohort 1)	4/38	nr	nr	na	0.38[10]	ns	+15
WRMT: Word ID subtest	Phonics	Kindergarten/lowest 25% (Cohort 1)	4/38	nr	nr	na	2.56[10]	Statistically significant	+49
WRMT: Letter ID subtest	Letter Knowledge	Grade 1/lowest 25% (Cohort 2)	4/38	nr	nr	na	–0.07[10]	ns	–3
WRMT: Word ID subtest	Phonics	Grade 1/lowest 25% (Cohort 2)	4/38	28.16 (10.02)	18.53 (12.78)	9.63	0.82	ns	+29

(continued)

Appendix A4.4 Summary of subgroup findings for alphabetics domain[1] *(continued)*

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
WRMT: Word Attack subtest	Phonics	Grade 1/lowest 25% (Cohort 2)	4/38	9.05 (5.37)	4.68 (5.76)	4.37	0.77	ns	+28

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

1. This appendix presents subgroup findings (students in the lowest 25% of their grades) for measures that fall in the alphabetics domain. Total group scores were used for rating purposes and are presented in Appendix A3.2.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. WWC combined means and standard deviations for two *SFA®* schools (Abbottston and City Springs) and their counterparts. Adjusted posttest means (with pretests standard scores as covariates) were used for effect size calculations. Kindergarten and grade 1 cohorts from Abbottston elementary school received four years of intervention.
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Ross & Casey (1998), Madden et al. (1993), and Smith et al. (1993), a correction for clustering was needed, so the significance levels may differ from those reported in the original study.
10. Authors reported effect sizes that used comparison group standard deviation in the denominator (Glass's delta).

Appendix A4.5 Summary of subgroup findings for comprehension domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Ross & Casey, 1998 (quasi-experimental design)[8]—Two years of intervention									
WRMT: Passage Comprehension subtest	Reading comprehension	Kindergarten/ lowest 25%	7/79	12.29 (7.79)	11.20 (8.20)	1.09	0.13	ns	+5
Smith et al., 1993 (quasi-experimental design)[8]—One year of intervention									
Peabody Picture Vocabulary Test	Vocabulary development	Kindergarten/lowest 25% (Cohort 1)	4/38	nr	nr	na	0.26[9]	ns	+10
WRMT: Passage Comprehension subtest	Reading comprehension	Grade 1/lowest 25% (Cohort 2)	4/38	9.84 (6.18)	8.11 (7.13)	1.73	0.25	ns	+10

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

1. This appendix presents subgroup findings (students in the lowest 25% of their grades) for measures that fall in the comprehension domain. Total group scores were used for rating purposes and are presented in Appendix A3.3.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Ross & Casey (1998) and Smith et al. (1993), a correction for clustering was needed so the significance levels may differ from those reported in the original study.
9. Authors reported effect sizes that used comparison group standard deviation in the denominator (Glass's delta). Effect size was computed by subtracting the comparison group mean from the intervention group mean and dividing the result by the comparison group standard deviation.

Appendix A4.6 Summary of subgroup findings for general reading achievement domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Madden et al., 1993 (quasi-experimental design)[8, 9]—Three years of intervention									
Durrell Oral Reading subtest	General reading	Pre-kindergarten/ lowest 25% (Cohort 1)	4/54	3.78 (4.05)	0.97 (2.62)	2.82	0.81	ns	+29
Durrell Oral Reading subtest	General reading	Kindergarten/lowest 25% (Cohort 2)	4/38	7.79 (5.25)	4.21 (3.83)	3.58	0.76	ns	+28
Durrell Oral Reading subtest	General reading	Grade 1/lowest 25% (Cohort 3)	4/44	14.00 (6.42)	7.63 (4.89)	6.36	1.10	ns	+36
Ross & Casey, 1998 (quasi-experimental design)[9]—Two years of intervention									
Durrell Oral Reading subtest	General reading	Kindergarten/ lowest 25%	7/79	4.14 (3.84)	3.00 (3.60)	1.14	0.31	ns	+12

ns = not statistically significant

1. This appendix presents subgroup findings (students in the lowest 25% of their grades) for measures that fall in the general reading domain. Total group scores were used for rating purposes and are presented in Appendix A3.4.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. WWC combined means and standard deviations for two *SFA®* schools (Abbottston and City Springs) and their counterparts. Adjusted posttest means (with pretests standard scores as covariates) were used for effect size calculations. Kindergarten and Grade 1 cohorts from Abbottston elementary school received four years of intervention.
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Madden et al. (1993) and Ross & Casey (1998), a correction for clustering was needed so the significance levels may differ from those reported in the original study.

Appendix A4.7 Summary of alternative groups findings for alphabetics domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Madden et al, 1993 (quasi-experimental design)[8, 9]—Dropout version, three years of intervention									
WLPB: Letter-Word ID subtest	Phonics	Pre-kindergarten (Cohort 1)	6/282	18.74 (5.44)	15.77 (6.53)	2.97	0.49	ns	+ 19
WLPB: Word Attack subtest	Phonics	Pre-kindergarten (Cohort 1)	6/282	5.50 (4.01)	2.23 (3.56)	3.27	0.86	Statistically significant	+31
WLPB: Letter-Word ID subtest	Phonics	Kindergarten (Cohort 2)	6/292	25.39 (6.89)	21.77 (6.78)	3.62	0.53	ns	+20
WLPB: Word Attack subtest	Phonics	Kindergarten (Cohort 2)	6/292	9.08 (6.37)	4.98 (4.79)	4.10	0.72	ns	+27
WLPB: Letter-Word ID subtest	Phonics	Grade 1 (Cohort 3)	6/232	29.14 (6.24)	25.78 (6.37)	3.36	0.53	ns	+20
WLPB: Word Attack subtest	Phonics	Grade 1 (Cohort 3)	6/232	10.22 (6.54)	7.42 (5.92)	2.81	0.45	ns	+17
Madden et al., 1993 (quasi-experimental design)[10]—Dropout version, one year of intervention									
WRMT: Combined Letter ID and Word ID subtests	Phonics	Kindergarten (Cohort 1)	8/256	18.75 (5.86)	17.46 (6.58)	1.29	0.21	ns	+8
WRMT: Word Attack subtest	Phonics	Kindergarten (Cohort 1)	8/256	5.05 (4.54)	3.77 (4.94)	1.28	0.27	ns	+11
WRMT: Word Attack subtest	Phonics	Grade 1 (Cohort 2)	8/216	7.77 (5.70)	8.41 (6.14)	–0.64	–0.11	ns	–4
WRMT: Word ID subtest	Phonics	Grade 1 (Cohort 2)	8/216	24.95 (6.25)	25.41 (6.41)	–0.46	–0.07	ns	–3
WRMT: Word Attack subtest	Phonics	Grade 2 (Cohort 3)	8/106	11.52 (7.32)	10.11 (6.07)	1.41	0.21	ns	+8
WRMT: Word ID subtest	Phonics	Grade 2 (Cohort 3)	8/106	30.42 (4.82)	28.49 (5.80)	1.93	0.36	ns	+14

ns = not statistically significant

1. This appendix presents findings for dropout version of *SFA®* for measures that fall in alphabetics domain. Data for the full implementation model of *SFA®* that reflected students' exposure to the intervention for the longest period of time were used for intervention rating purposes and are presented in Appendix A3.2.

(continued)

Appendix A4.7 Summary of alternative groups findings for alphabetics domain[1] *(continued)*

2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. WWC combined means and standard deviations for three *SFA*® schools (Dallas Nicholas, Harriet Tubman, and Dr. Bernard Harris) and their counterparts. Adjusted posttest means (with pretests standard scores as covariates) were used for effect size calculations.
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not applied to findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Madden et al. (1993), a correction for clustering was needed so the significance levels may differ from those reported in the original study.
10. Data are taken from Slavin et al. (1990).

Appendix A4.8 Summary of alternative groups findings for comprehension domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Madden et al., 1993 (quasi-experimental design)[8, 9]—Dropout version, one year of intervention									
Durrell Silent Reading subtest	Reading comprehension	Kindergarten (Cohort 1)	8/256	3.77 (3.95)	3.50 (4.64)	0.27	0.06	ns	+2
Durrell Silent Reading subtest	Reading comprehension	Grade 1 (Cohort 2)	8/216	8.42 (6.14)	7.75 (5.20)	0.67	0.12	ns	+5
Durrell Silent Reading subtest	Reading comprehension	Grade 2 (Cohort 3)	8/106	15.07 (5.25)	11.84 (5.49)	3.23	0.60	ns	+22

ns = not statistically significant

1. This appendix presents findings for dropout version of *SFA®* for measures that fall in comprehension domain. Data for the full implementation model of *SFA®* that reflected students' exposure to the intervention for the longest period of time were used for intervention rating purposes and are presented in Appendix A3.3.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. Data are taken from Slavin et al. (1990).
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not applied to findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Madden et al. (1993), a correction for clustering was needed so the significance levels may differ from those reported in the original study.

Appendix A4.9 Summary of alternative groups findings for general reading achievement domain[1]

Outcome measure	Construct	Study sample[3]	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Success for All®* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[4] (*SFA®* – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Madden et al., 1993 (quasi-experimental design)[8, 9]—Dropout version, three years of intervention									
Durrell Oral Reading subtest	General reading	Pre-kindergarten (Cohort 1)	6/282	5.70 (4.83)	4.11 (4.83)	1.59	0.33	ns	+13
Durrell Oral Reading subtest	General reading	Kindergarten (Cohort 2)	6/292	11.81 (7.04)	9.00 (6.50)	2.81	0.41	ns	+16
Durrell Oral Reading subtest	General reading	Grade 1 (Cohort 3)	6/232	16.60 (6.97)	13.50 (7.25)	3.10	0.44	ns	+17
Madden et al., 1993 (quasi-experimental design)[10]—Dropout version, one year of intervention									
CAT Total Reading	General reading	Kindergarten (Cohort 1)	8/256	470.28 (105.92)	485.13 (107.52)	–14.85	–0.14	ns	-6
Durrell Oral Reading Subtest	General reading	Kindergarten (Cohort 1)	8/256	4.69 (3.94)	4.89 (4.03)	–0.20	–0.05	ns	–2
CAT Total Reading	General reading	Grade 1 (Cohort 2)	8/216	348.67 (47.31)	360.67 (49.99)	–12	–0.25	ns	–10
Durrell Oral Reading Subtest	General reading	Grade 1 (Cohort 2)	8/216	10.09 (5.74)	9.34 (4.33)	0.75	0.15	ns	+6
CAT Total Reading	General reading	Grade 2 (Cohort 3)	8/106	387.44 (36.27)	388.15 (33.75)	–0.71	–0.02	ns	–1
Durrell Oral Reading Subtest	General reading	Grade 2 (Cohort 3)	8/106	16.02 (6.52)	12.13 (4.22)	3.89	0.70	ns	+26

ns = not statistically significant

1. This appendix presents findings for the dropout version of *SFA®* for measures that fall in general reading achievement domain. Data for the full implementation model of *SFA®* that reflected students' exposure to the intervention for the longest period of time were used for intervention rating purposes and are presented in Appendix A3.4.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The cohort is defined by the time pretest is administered. For example, kindergarten cohort describes students who completed pretest measures in kindergarten.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.

(continued)

Appendix A4.9 Summary of alternative groups findings for general reading achievement domain[1] *(continued)*

8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not applied to findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Madden et al. (1993), a correction for clustering was needed so the significance levels may differ from those reported in the original study.
9. WWC combined means and standard deviations for three *SFA*® schools (Dallas Nicholas, Harriet Tubman, and Dr. Bernard Harris) and their counterparts. Adjusted posttest means (with pretests standard scores as covariates) were used for effect size calculations.
10. Data are taken from Slavin et al. (1990).

Appendix A5.1 *Success for All*® rating for the alphabetics domain

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

For the outcome domain of alphabetics, the WWC rated *Success for All*® as having potentially positive effects. It did not meet the criteria for positive effects because only one study showed a statistically significant positive effect. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *Success for All*® was assigned the highest applicable rating.

Rating received

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

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

  Met. One study that met standards for a strong design showed a statistically significant positive effect. Four studies that met standards with reservations showed substantively important positive effects.

AND

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

  Met. No studies showed statistically significant or substantively important negative effects. Two out of the seven studies showed indeterminate effects.

Other ratings considered

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

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

  Not met. Only one study showed a statistically significant positive effect.

AND

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

  Met. No studies showed statistically significant or substantively important negative effects.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A5.2 *Success for All*® rating for the comprehension domain

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

For the outcome domain of comprehension, the WWC rated *Success for All*® as having mixed effects. It did not meet the criteria for positive effects because only one study showed statistically significant positive effects. In addition, it did not meet the criteria for potentially positive effects because more studies showed indeterminate effects than substantively important or statistically significant positive effects. The remaining ratings (no discernible effects, potentially negative effects, and negative effects) were not considered because *Success for All*® was assigned the highest applicable rating.

Rating received

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

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

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

OR

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

  Met. One study showed a statistically significant positive effect, one study showed a substantively important positive effect, and four studies showed indeterminate effects.

Other ratings considered

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

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

  Not met. Only one study had a statistically significant positive effect in this domain.

AND

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

  Met. No studies showed statistically significant or substantively important negative effects in this domain.

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

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

  Met. One study had a statistically significant positive effect, and one study had a substantively important positive effect in this domain.

AND

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

  Not met. No studies showed statistically significant or substantively important negative effects in this domain, and more studies showed indeterminate effects (four) than statistically significant (one) or substantively important positive effects (one) in this domain.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A5.3 *Success for All*® rating for the general reading achievement domain

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

For the outcome domain of general reading achievement, the WWC rated *Success for All*® as having potentially positive effects. It did not meet the criteria for positive effects because only one study showed a statistically significant positive effect. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *Success for All*® was assigned the highest applicable rating.

Rating received

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

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

  Met. Three studies showed substantively important positive effects.

AND

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

  Met. No studies showed statistically significant or substantively important negative effects. Three studies showed indeterminate effects and three studies showed substantively important positive effects.

Other ratings considered

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

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

  Not met. No studies showed a statistically significant positive effect.

AND

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

  Met. No studies showed statistically significant or substantively important negative effects.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain-level effect. The WWC also considers the size of the domain-level effect for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A6 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Alphabetics	7	67	3,103	Moderate to large
Fluency	0	0	0	na
Comprehension	6	65	2,565	Moderate to large
General reading achievement	6	31	1,767	Moderate to large

na = not applicable/not studied

1. A rating of "moderate to large" requires at least two studies and two schools across studies in one domain and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small."