Appendix

Appendix A1.1 Study characteristics: Mooney, 2003

Characteristic	Description
Study citation	Mooney, P. J. (2003). An investigation of the effects of a comprehensive reading intervention on the beginning reading skills of first graders at risk for emotional and behavioral disorders (Doctoral dissertation, University of Nebraska–Lincoln). *Dissertation Abstracts International, 64*(05A), 85–1599.
Participants	The study included first-grade students who were screened prior to treatment and determined to be at risk for emotional and behavioral disorders. All of the students were systematically screened using a modified version of the first two steps of the Systematic Screening for Behavioral Disorders and met criteria for either internalizing or externalizing behavioral disorders.
Setting	Study participants were enrolled in seven elementary schools in Lincoln, Nebraska.
Intervention	Children in the experimental group received the standard beginning reading instruction provided in the classroom in addition to *Sound Partners*. The general first-grade literacy curriculum included the phonics component of the *Open Court* reading program and various teacher-designed reading, listening, and writing activities. Students in the experimental group received approximately 30 minutes of tutoring in reading 5 times weekly throughout the majority of the 2002–03 school year (i.e., mid-September through mid-April). The mean number of *Sound Partners* lessons completed by participants in the experimental condition was 68.2 (range 2 to 100). Of the 28 first-graders who began the intervention, seven (25%) completed all 100 lessons, while four (14%) completed less than half of the lessons.
Comparison	Children in the comparison group received the standard beginning reading instruction provided in the classroom and a home-school intervention designed to improve social skills known as *First Step to Success*. All 19 participants in the comparison group completed the *First Step to Success* program.
Primary outcomes and measurement	The study reports the total reading scores on the Woodcock Reading Mastery Test–Revised/Normative Update (WMRT-R/NU). The total reading score combines the scores from the Word Attack, Word Identification, Word Comprehension, and Passage Comprehension subtests. The study also includes the combined Word Attack and Word Identification scores and the combined Word Comprehension and Passage Comprehension scores, which are presented in Appendix A2.4. In addition, the study presents the scores from three subtests of the Dynamic Indicator of Basic Early Literacy Skills (DIBELS): Phoneme Segmentation, Nonsense Word Fluency, and Oral Reading Fluency. For a more detailed description of these outcome measures, see Appendices A2.1 and A2.2.
Staff/teacher training	A total of 14 tutors (two at each of the seven schools) implemented the *Sound Partners* program. Tutors were identified and selected by the research team at the University of Nebraska–Lincoln's Center for At-Risk Children's Services. A five-step training strategy was used to train tutors to implement the *Sound Partners* program: (1) a presentation to tutors on the theory and rationale for *Sound Partners*; (2) a demonstration involving live modeling of skills; (3) simulated testing conditions to provide practice for the tutors until a high level of skill performance was obtained; (4) structured feedback to tutors on how proficiently they performed during simulated practice conditions (tutors were observed on at least three occasions before beginning tutoring with children); and (5) following training, observation of tutors on a regular basis until a satisfactory maintenance level was achieved.

Appendix A1.2 Study characteristics: Vadasy et al., 1997a

Characteristic	Description
Study citation	Vadasy, P. F., Jenkins, J. R., Antil, L. R., Wayne, S. K., & O'Connor, R. E. (1997a). The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers. *Learning Disability Quarterly, 20*(1), 126–139.
Participants	After prescreening and pretesting 229 first-graders, the 46 students scoring lowest on the pretests were stratified and randomly assigned to intervention and control groups, with 23 students in each group. At study completion, 20 students remained in each group, for a total of 40 students.[1] Ninety-five percent of the study students were of minority background.
Setting	The study includes first-grade children from four schools in a large urban school district in Washington state. Forty-five percent of students in the four schools were eligible for free or reduced-price lunch. Students from 13 classrooms were in the final analytic sample of 40 students.
Intervention	A set of 100, thirty-minute *Sound Partners* lessons, each including six to eight activities, was administered to students in the intervention group. Some activities were phased out once students mastered the target skills. Other activities were initiated only after most letter sounds had been introduced, and they continued throughout the lessons. Students received reading tutoring after school for 30 minutes per day, four days per week, for 23 weeks. Tutors were provided with lessons to guide the sessions, which focused for specific amounts of time on instruction in letter names and sounds, sound categorization, rhyming exercises, onset-rime segmentation, auditory blending, spelling, writing, and reading from Bob Books®.
Comparison	The control group students received only the regular reading instruction in their classrooms.
Primary outcomes and measurement	For both pre- and posttests, the authors administered a test of alphabetics, the Wide Range Achievement Test–Revised Reading subtest. Alphabetics achievement was further assessed using the Dolch Word Recognition test, the Woodcock-Johnson Psycho-Educational Battery–Revised Word Attack subtest, the Bryant Pseudoword Test, an additional pseudoword list, and the Yopp-Singer Segmentation Task. The authors assessed reading fluency using the primary and first-grade passages of the Analytical Reading Inventory. The authors also used spelling and writing assessments, but they were not included in this review because they are outside the scope of the Beginning Reading review protocol. For a more detailed description of the included outcome measures, see Appendices A2.1 and A2.2.
Staff/teacher training	Tutors (nonprofessional educators who were community members) were trained as a group two weeks before they began tutoring. Six hours of training were provided at that time and included an introduction to the goals and methods of the tutoring lessons, a presentation and practice role-playing on each lesson component, general information on tutoring, suggestions for behavior management and safety, and record-keeping tasks. Three hours of follow-up training were provided after the tutoring began.

1. Information about the sample size of 46 students at baseline was received by the WWC through communication with the author.

Appendix A1.3 Study characteristics: Vadasy & Sanders, 2008

Characteristic	Description
Study citation	Vadasy, P. F., & Sanders, E. A. (2008). Code-oriented instruction for kindergarten students at risk for reading difficulties: A replication and comparison of instructional grouping. *Reading and Writing: An Interdisciplinary Journal, 21*(9), 929–963.
Participants	Full-day kindergarten teachers in 13 urban public elementary schools were asked to identify students who would benefit from intensive additional reading instruction. Of the referred students with parental consent, 99 met eligibility criteria based on scoring below cutoff scores on DIBELS tests. After dropping one student (who was the only student in one classroom to be eligible), the other 98 students who met eligibility standards were randomly assigned to one of two intervention groups (one in which tutoring occurred one-on-one, and one in which tutoring occurred in pairs) or to the comparison group using an algorithm that compensated for the fact that students in the pair tutoring group needed to be assigned in pairs within the same classroom. Pretests were given in December and posttests at the end of the school year.
Setting	The study took place in 13 urban public elementary schools.
Intervention	Paraeducators, equipped with 70 scripted lessons with seven to eight activities per lesson, worked with students individually for 30 minutes a day, four days a week, for 18 weeks. Tutoring was conducted during the school day in a quiet nearby school space. Typically, 20 minutes were devoted to phonics and 10 minutes to oral reading practice using Bob Books®, although the tutors were free to adjust this to meet individual student needs. For tutoring in pairs, the same general approach was followed, but two students were tutored at once. If one student was ahead of the other, then the tutor focused on the student who was behind while the other student read silently for part of the time. This review focuses on the combined effect of the two tutoring groups compared to the group that did not receive tutoring. The study does not identify the intervention as *Sound Partners*, although the developer verified that this study included *Sound Partners* instruction.
Comparison	Control group students received a variety of Title I, ESL, and special education services available to all students in the study schools. The control students did not receive supplemental tutoring.
Primary outcomes and measurement	The study addresses the alphabetics domain using a set of standardized tests (DIBELS, Comprehensive Test of Phonological Processing [CTOPP], Woodcock Reading Mastery Test, Test of Word Reading Efficiency [TOWRE]), the reading comprehension domain using a standardized test (Woodcock Reading Mastery Test–Revised/Normative Update [WRMT-R/NU] Passage Comprehension subtest), and the reading fluency domain using an author-developed measure that is similar to standardized tests of reading fluency. The study also includes a spelling assessment, but it is not included in this review because it is outside the scope of the Beginning Reading review protocol. For a more detailed description of the included outcome measures, see Appendices A2.1–A2.3.
Staff/teacher training	Twenty-one paraeducators were hired by schools on the basis of their interest in working with children, prior tutoring experience, and scheduling flexibility. The paraeducators averaged 3.3 years of prior tutoring experience. They were trained in an initial two-hour session. Follow-up training was provided throughout the intervention, along with coaching for paraeducators with less experience and/or low initial intervention fidelity ratings.

Appendix A1.4 Study characteristics: Vadasy, Sanders, & Peyton, 2006

Characteristic	Description
Study citation	Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology, 98*(3), 508–528.
Participants	Seventy-five kindergarten students were recruited to participate in the study after having been identified by their teachers as needing additional reading instruction. Students also had to meet eligibility screens for the study by receiving low scores on a range of reading pretests. Thirty-nine students were randomly assigned to the intervention group, and 36 were assigned to the comparison group. Three students from the intervention group and five students from the comparison group dropped out of the study, yielding a final analysis sample of 36 students in the intervention group and 31 students in the comparison group. Outcomes were assessed immediately after the 18-week intervention period and again one year later, during the spring of the students' first-grade year. However, the first-year follow-up results do not meet WWC evidence standards because the intervention is confounded with another mentoring program.
Setting	The study was conducted in 19 full-day kindergarten classrooms in 9 elementary schools.
Intervention	Students in the intervention group received individualized reading instruction from a trained paraeducator for 30 minutes a day, four days per week, for 18 weeks. Paraeducators taught students using a series of 62 scripted lessons, with three to four activities per lesson. The first 20 minutes of tutoring focused on phonics activities from the scripted lessons. During the last 10 minutes of tutoring, the students read aloud from Bob Books®. Most children read independently, but some read the story with the tutors (either echo reading or partner reading). Students completed an average of 47 lessons during the 18 weeks.
Comparison	Students in the comparison group received their regular reading instruction and services.
Primary outcomes and measurement	Outcomes were assessed on eight measures: (1) DIBELS Letter Name Fluency subtest, (2) CTOPP phonological awareness composite, (3) Word Reading Accuracy subtest of the WRMT-R/NU, (4) TOWRE, (5) DIBELS Phoneme Segmentation Fluency subtest, (6) DIBELS Nonsense Word Fluency subtest, (7) an oral reading fluency test, and (8) the Passage Comprehension subtest of the WRMT-R/NU. For a more detailed description of these outcome measures, see Appendices A2.1– A2.3. The study also assessed outcomes on the Revised Spelling subtest of the Wide Range Achievement Test–Revised (WRAT-R), but that outcome is excluded from this review because it falls outside the scope of the Beginning Reading review protocol.
Staff/teacher training	The 11 paraeducators in this study were hired as employees of the school district. All but four had prior tutoring experience, and five had prior experience working with kindergarten students. Their average education level was 14 years, and six tutors had more than a high school education.

Appendix A1.5 Study characteristics: Jenkins et al., 2004

Characteristic	Description
Study citation	Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading, 8*(1), 53–86.
Participants	Teachers identified first-graders from 26 classrooms in 11 schools whom they considered at risk for reading failure. The researchers then identified 121 students who scored at or below the 25th percentile on the Reading subtest of the WRAT-R as eligible for inclusion in the study. The treatment and comparison groups were formed partly by convenience and partly through random assignment, with some schools agreeing to allow students to serve only as the comparison group.[1] After attrition, the analysis sample included 79 students (in 21 classes) in the treatment condition and 20 students (in 10 classrooms) in the comparison condition. The study was conducted in a single school year.
Setting	The study was conducted in 11 public schools in an urban area.
Intervention	The tutoring lessons in phonics were drawn from *Sound Partners*. They targeted letter-sound correspondences, blending letters into sounds, reading and spelling phonetically regular words, and reading nondecodable and high-frequency words scheduled to appear in the text portion of the lesson. Tutors also worked with students who read from storybooks that had varying degrees of decodability, with one of the treatment groups reading from books with highly decodable words and the other treatment group reading from books with high-frequency but less decodable words. The WWC considers the two treatment groups to be variants of the *Sound Partners* intervention and so presents them as a single treatment group. Lessons were scripted, and all tutoring was one-on-one. Lessons were provided 30 minutes a day, four days a week, for 25 weeks.
Comparison	Children in the control group received typical classroom instruction only, without tutoring in phonics or story reading.
Primary outcomes and measurement	At the conclusion of the intervention, the students were given the Phonemic Decoding and Sight Word reading subtests of the TOWRE; the Word Attack, Word Identification, and Passage Comprehension subtests of the Woodcock Reading Mastery Test–Revised; the Bryant Pseudoword Test; the Reading subtest of the Wide Range Achievement Test–Revised; and fluency and accuracy reading tests from passages with highly decodable words, as well as passages with less decodable words. The study includes a text reading list that contained words that the students read as part of the *Sound Partners* curriculum. The WWC determined that this outcome was overaligned with the intervention and is therefore not included in this review. Students also took two spelling tests that are not included in this review because they are outside the scope of the Beginning Reading review protocol. For a more detailed description of the included outcome measures, see Appendices A2.1–A2.3.
Staff/teacher training	Tutors received scripted phonics lessons, directions for book reading, attendance forms and recording sheets for each student's lesson coverage, and a set of books for reading practice. Research staff provided tutors with three hours of formal training in lesson procedures, conducted weekly observations, provided ongoing coaching in lesson delivery, and held monthly follow-up meetings.

1. Information about how students were assigned to treatment and control conditions was received by the WWC through communication with the author.

Appendix A1.6 Study characteristics: Vadasy, Jenkins, & Pool, 2000

Characteristic	Description
Study citation	Vadasy, P. F., Jenkins, J. R., & Pool, K. (2000). Effects of tutoring in phonological and early reading skills on students at risk for reading disabilities. *Journal of Learning Disabilities, 33*(6), 579–590.
Participants	Vadasy, Jenkins, and Pool (2000) is a randomized controlled trial in which 46 first-graders from four elementary schools were randomly assigned to either participate in *Sound Partners* or receive the schools' regular classroom instruction. Teachers in 11 classrooms identified up to 6 students each whose reading performance in the fall concerned them. The 64 students identified by the teachers were pretested on four assessments, and those with the 46 lowest scores were randomly assigned. The remaining 18 students were kept as replacement students. In the course of the study, the researchers replaced two treatment and two comparison students on the basis of convenience and scheduling considerations.[1] The groups were balanced on gender (9 girls and 14 boys in each group). The study also examined second-year follow-up scores for a subsample of 37 students. This analysis is not included in this review, however, because the authors did not demonstrate that the intervention and comparison students included in the follow-up results were equivalent at baseline.
Setting	Participants were from four elementary schools in an urban school district. At the schools, nearly half of the students were eligible for free or reduced-price lunch, Title I services were available to students, and two-thirds of students were racial or ethnic minorities.
Intervention	In the study, tutoring took place for 27 weeks. Students attended from 54 to 89 sessions over this period, with an average of 72 sessions per child. The version of *Sound Partners* used for the study included additional, revised, or expanded components of a preceding version.
Comparison	Students in the counterfactual condition participated in the schools' regular classroom and Title I reading instruction activities.
Primary outcomes and measurement	For both pre- and posttests, the authors administered the Wide Range Achievement Test–Revised Reading subtest. For additional posttests, the authors used the Dolch Word Recognition, the Woodcock-Johnson Psycho-Educational Battery–Revised Word Attack subtest, the Bryant Pseudoword Test, the Yopp-Singer Segmentation Task, and the primary and first-grade passages of the Analytical Reading Inventory. The authors also used two spelling assessments, but they were not included in this review because they are outside the scope of the Beginning Reading review protocol. For a more detailed description of the included outcome measures, see Appendices A2.1 and A2.2.
Staff/teacher training	The researchers recruited tutors through the school newsletters. Nine tutors participated in the study (mainly parents of children in the schools). Tutors received $5 per hour for their tutoring and training time, which included eight hours of training before the program began and six additional hours of training during the school year. Training for tutors consisted of explanations, modeling, role-playing of each lesson component, guidelines for behavior management, record keeping, and error correction strategies. Follow-up training occurred during the year by tutor request or when researchers identified a need. Researchers replaced two tutors in the middle of the year with one new tutor.

1. Information on replacement procedures was received by the WWC through communication with the authors. Because the replacement was made based on convenience rather than random assignment, this procedure could have compromised the random assignment process. For this reason, the WWC determined that this study meets evidence standards with reservations.

Appendix A1.7 Study characteristics: Vadasy, Sanders, & Peyton, 2005

Characteristic	Description
Study citation	Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2005). Relative effectiveness of reading practice or word-level instruction in supplemental tutoring: How text matters. *Journal of Learning Disabilities, 38*(4), 364–380.
Participants	This sample was drawn from 12 participating schools, six of which were assigned as treatment sites, five as control sites, and one that included both treatment and control students. During the first month of first grade, 22 teachers referred students they judged to be at risk for reading; in all, 99 first-graders met the criteria for participation, which included (1) parental consent, (2) not repeating first grade, and (3) scoring below the 25th percentile on the WRAT-R. Students at treatment sites were assigned to tutors based on schedules and availability. Of the 78 students completing all phases of the study, the authors chose 57 to be included in the analyses based on the comparability of their pretest scores. The authors selected students to analyze for two treatment groups and a control group by matching triads of students as closely as possible on a pretest composite score calculated by averaging the *z*-scores of all pretest scores. Both treatment groups received 30 minutes of tutoring, but one of the treatment groups spent 10 of the minutes in oral reading practice and the other did not. The WWC considers the two treatment groups to be variants of the *Sound Partners* intervention and so combines them into a single treatment group.
Setting	The study includes 12 schools from a large, urban school district in the northwestern United States.
Intervention	In addition to regular classroom reading instruction, both intervention groups received supplementary individual tutoring using *Sound Partners*. Tutoring occurred for 30-minute sessions during the school day, four days a week, from October to May. One treatment group used *Sound Partners* phonics-based instruction for 15 to 20 minutes, followed by oral text reading practice in Bob Books® for the remaining 10 to 15 minutes. The other treatment group spent all 30 minutes using *Sound Partners*.
Comparison	The comparison students received regular classroom reading instruction only.
Primary outcomes and measurement	Students were tested on a variety of measures, most of which are standardized tests. They included the WRAT-R Reading subtest; the WRMT-R/NU Word Attack, Word Identification, and Passage Comprehension subtests; the TOWRE Phonemic Decoding and Sight Word subtests; and a passage reading fluency test devised by the authors to measure the rate and accuracy at which students could read grade-appropriate texts. The authors also assessed spelling, but it is not included in this report because it is outside the scope of the Beginning Reading review protocol. For a more detailed description of the included outcome measures, see Appendices A2.1–A2.3.
Staff/teacher training	Nineteen paraprofessional tutors were hired and paid by the schools in which they worked. More than half of the tutors had at least one year of *Sound Partners* tutoring experience. Experienced tutors received about two hours of initial training, and new tutors received about four hours of training.

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

Outcome measure	Description
Phonemic awareness	
Yopp-Singer Segmentation Task	This task asks students to segment sounds of 22 orally given words with corrective feedback. Testing continues until students miss 10 consecutive items. The score is the total number of words segmented correctly (as cited in Vadasy, Jenkins, & Pool, 2000).
Phonological awareness	
Comprehensive Test of Phonological Processes (CTOPP)—Phonological Awareness	This norm-referenced assessment provides an overall measure of a child's phonological awareness. The composite score is based on three subtests: Blending Words, Elision, and Sound Matching. The Blending Words subtest measures skill in blending separately presented sounds together to form words. The Sound Matching subtest measures skill at matching words that begin and end with the same sounds as a spoken word. The Elision subtest measures students' ability to manipulate components of a word. The student listens to words and is asked to repeat the word with one of the sounds missing (as cited in Vadasy & Sanders, 2008 and Vadasy, Sanders, & Peyton, 2006).
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)—Phoneme Segmentation Fluency subtest	This standardized test measures a child's ability to segment three- and four-phoneme words into their individual phonemes fluently. The child is presented with words orally and asked to produce verbally the individual phonemes for each word (as cited in Mooney, 2003 and Vadasy, Sanders, & Peyton, 2006).
Letter knowledge	
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)—Letter Naming Fluency subtest	This task presents students with a page of lower- and uppercase letters arranged randomly and asks them to name as many of the letters as they can. The score is the number of letters named correctly in one minute (as cited in Vadasy & Sanders, 2008 and Vadasy, Sanders, and Peyton, 2006).
Phonics	
Bryant Pseudoword Test	For this test, a student reads a list of 50 pseudowords until five consecutive items are missed. One point is assigned to each correct response (as cited in Vadasy et al., 1997a; Jenkins et al., 2004; and Vadasy, Jenkins, and Pool, 2000).
Dolch Word Recognition	In this test, a student reads from a list of 220 short, frequently used words arranged in groups according to basal reading levels, until 10 consecutive items are missed. The score is the total number of words correctly identified (as cited in Vadasy et al., 1997a and Vadasy, Jenkins, and Pool, 2000).
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)—Nonsense Word Fluency subtest	This subtest measures a child's word reading ability, including letter-sound correspondence, and the ability to blend letter sounds into words (as cited in Mooney, 2003 and Vadasy, Sanders, & Peyton, 2006).
Pseudoword List	This test asks students to read a list of 45 nonwords. The list includes only one-syllable items with few similar words (to decrease the chance of reading from analogy) and items with many consonant clusters, which are not featured until the last half of the Bryant list (as cited in Vadasy et al., 1997a).
Test of Word Reading Efficiency (TOWRE)	The TOWRE is a standardized, nationally normed measure consisting of two subtests: Phonemic Decoding and Sight Word Efficiency. The composite score on the TOWRE is the mean of the two subtest scores (as cited in Vadasy & Sanders, 2008 and Vadasy, Sanders, & Peyton, 2006).

(continued)

Appendix A2.1 Outcome measures for the alphabetics domain by construct *(continued)*

Outcome measure	Description
Test of Word Reading Efficiency (TOWRE)—Phonemic Decoding Efficiency subtest	This subtest measures the number of pronounceable nonprinted words that students can accurately decode within 45 seconds (as cited in Jenkins et al., 2004 and Vadasy, Sanders, & Peyton, 2005).
Test of Word Reading Efficiency (TOWRE)—Sight Word Efficiency subtest	This subtest assesses the number of real printed words that students can accurately identify within 45 seconds (as cited in Jenkins et al., 2004 and Vadasy, Sanders, & Peyton, 2005).
Wide Range Achievement Test–Revised (WRAT-R)—Reading	This norm-referenced achievement test asks students to name letters and words. The number of words and letters correctly identified is transformed to an age-based standard score (as cited in Vadasy et al., 1997a; Jenkins et al., 2004; Vadasy, Jenkins, and Pool, 2000; and Vadasy, Sanders, & Peyton, 2005).
Woodcock-Johnson Psycho-Educational Battery–Revised (WJ-R)—Word Attack subtest	For this test, the examinee pronounces pseudowords that increase in difficulty. One point is awarded for each correct response, and the number of correct items is transformed into age-based standard scores (as cited in Vadasy et al., 1997a and Vadasy, Jenkins, and Pool, 2000).
Woodcock Reading Mastery Test–Revised (WRMT-R)—Word Attack subtest	The Word Attack subtest of the WRMT-R measures the student's ability to apply phonic and structural analysis skills to pronounce unfamiliar words. Subjects cannot read the pseudowords by sight and must rely on phonological processes to decode them (as cited in Jenkins et al., 2004 and Vadasy, Sanders, & Peyton, 2005).
Woodcock Reading Mastery Test–Revised (WRMT-R)—Word Identification subtest	This is a test of decoding skills. The standardized test requires children to read aloud isolated real words that range in frequency and difficulty (as cited in Vadasy, Sanders, & Peyton, 2006; Jenkins et al., 2004; and Vadasy, Sanders, & Peyton, 2005).
Woodcock Reading Mastery Test–Revised/Normative Update (WRMT-R/NU)—Word Reading Accuracy	The WRMT-R/NU Word Reading Accuracy score averages the scores from the Word Attack and Word Identification subtests (as cited in Vadasy & Sanders, 2008 and Vadasy, Sanders, & Peyton, 2006).

Appendix A2.2 Outcome measures for the fluency domain

Outcome measure	Description
Analytical Reading Inventory	This test asks students to read grade-appropriate passages aloud and measures their reading fluency (time and accuracy). The score is the number of words correctly read per minute (as cited in Vadasy et al., 1997a and Vadasy, Jenkins, & Pool, 2000).
Dynamic Indicator of Basic Early Literacy Skills (DIBELS)—Oral Reading Fluency subtest	This is an individually administered assessment in which students read aloud from a passage for one minute. Scorers record the total number of words read correctly during that time (as cited in Mooney, 2003).
Nonphonetically Controlled Passage Accuracy	This task requires a student to read aloud a passage from a book that was judged to have fewer decodable high-frequency words. The score is the percentage of words read correctly in one minute (as cited in Jenkins et al., 2004).
Nonphonetically Controlled Passage Rate	This test requires a student to read aloud a passage from a book that was judged to have fewer decodable high-frequency words. The score is the number of words read correctly in one minute (as cited in Jenkins et al., 2004).
Passage Reading Accuracy	This test requires a student to read aloud from three grade-level passages. The score is the average percentage of words read correctly across the three passages (as cited in Vadasy, Sanders, & Peyton, 2005).
Passage Reading Rate	This test requires a student to read aloud from grade-level passages for one minute per passage. The score is the average number of words read correctly across the passages (as cited in Vadasy & Sanders, 2008; Vadasy, Sanders, & Peyton, 2006; and Vadasy, Sanders, & Peyton, 2005).
Phonetically Controlled Passage Rate	In this test, a student reads aloud passages from two books that were judged to include highly decodable words. The score is the number of words read correctly in one minute (as cited in Jenkins et al., 2004).
Phonetically Controlled Passage Accuracy	In this test, a student reads aloud passages from two books that were judged to include highly decodable words. The score is the percentage of words read correctly in one minute (as cited in Jenkins et al., 2004).

Appendix A2.3 Outcome measure for the comprehension domain

Outcome measure	Description
Woodcock Reading Mastery Test–Revised (WRMT-R)—Passage Comprehension subtest	This standardized test measures comprehension by asking students to fill in missing words in a short paragraph. The normative update (NU) of the WRMT-R (WRMT-R/NU) scales the tests based on revised norms (as cited in Vadasy & Sanders, 2008; Vadasy, Sanders, & Peyton, 2006; Jenkins et al., 2004; and Vadasy, Sanders, & Peyton, 2005).

Appendix A2.4 Outcome measure for the general reading achievement domain

Outcome measure	Description
Woodcock Reading Mastery Test–Revised (WRMT-R)—Total Reading	The Total Reading score for the WRMT-R consists of the scores from four subtests: Word Identification, Word Attack, Word Comprehension, and Passage Comprehension, which are all described above (as cited in Mooney, 2003).

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain by construct[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Phonemic awareness construct								
Vadasy et al., 1997a[7]								
Yopp-Singer Segmentation	Grade 1	13/40	16.75 (3.67)	14.65 (6.03)	2.10	0.41	ns	+16
Vadasy, Jenkins, & Pool, 2000[7]								
Yopp-Singer Segmentation	Grade 1	11/46	15.51 (3.79)	11.15 (5.53)	4.36	0.90	Statistically significant	+32
Phonological awareness construct								
Mooney, 2003[7]								
DIBELS Phoneme Segmentation subtest	Grade 1	47 students	30.90 (10.30)	30.10 (14.50)	0.80	0.06	ns	+3
Vadasy & Sanders, 2008[7,8]								
CTOPP: Phonological Awareness	Kindergarten	30/86	97.82 (12.39)	90.69 (12.97)	7.13	0.59	Statistically significant	+22
Vadasy, Sanders, & Peyton, 2006[7]								
CTOPP: Phonological Awareness	Kindergarten	19/67	88.00 (11.90)	85.00 (10.20)	3.00	0.27	ns	+10
DIBELS Phoneme Segmentation subtest	Kindergarten	19/67	8.58 (10.62)	4.65 (5.83)	3.93	0.44	ns	+17
Letter knowledge construct								
Vadasy & Sanders, 2008[7,8]								
DIBELS Letter Naming Fluency subtest	Kindergarten	30/86	25.72 (12.74)	27.72 (17.46)	–2.00	–0.14	ns	–6
Vadasy, Sanders, & Peyton, 2006[7]								
DIBELS Letter Naming Fluency subtest	Kindergarten	19/67	21.00 (14.20)	20.00 (10.40)	1.00	0.08	ns	+3

(continued)

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain by construct[1] *(continued)*

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Phonics construct								
Mooney, 2003[7]								
DIBELS Nonsense Word Fluency subtest	Grade 1	47 students	68.50 (36.20)	55.30 (25.20)	13.20	0.40	ns	+16
Vadasy et al., 1997a[7]								
Bryant Pseudoword Test	Grade 1	13/40	19.47 (11.86)	13.29 (10.74)	6.18	0.54	ns	+20
Dolch Word Recognition	Grade 1	13/40	131.93 (52.31)	123.57 (57.10)	8.36	0.15	ns	+6
Pseudoword List	Grade 1	13/40	12.75 (12.31)	9.65 (8.80)	3.10	0.28	ns	+11
WJ-R Word Attack subtest	Grade 1	13/40	8.58 (5.22)	7.42 (5.51)	1.16	0.21	ns	+8
WRAT-R: Reading	Grade 1	13/40	46.08 (8.62)	43.37 (8.91)	2.71	0.30	ns	+12
Vadasy & Sanders, 2008[7,8]								
TOWRE	Kindergarten	30/86	96.14 (6.28)	94.50 (5.64)	1.64	0.29	ns	+11
WRMT-R/NU Word Reading Accuracy	Kindergarten	30/86	105.02 (9.33)	99.38 (9.26)	5.64	0.63	Statistically significant	+24
Vadasy, Sanders, & Peyton, 2006[7]								
DIBELS Nonsense Word Fluency subtest	Kindergarten	19/67	5.94 (5.22)	3.35 (5.19)	2.59	0.49	ns	+19
TOWRE	Kindergarten	19/67	93.00 (5.80)	90.00 (6.30)	3.00	0.49	ns	+19
WRMT-R/NU Word Reading Accuracy	Kindergarten	19/67	98.00 (9.50)	90.00 (6.90)	8.00	0.94	Statistically significant	+33

(continued)

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain by construct[1] *(continued)*

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Jenkins et al., 2004[7,9]								
Bryant Pseudoword Test	Grade 1	25/99	20.82 (10.81)	9.40 (6.05)	11.42	1.13	Statistically significant	+37
TOWRE Phonemic Decoding subtest	Grade 1	25/99	10.73 (7.58)	8.05 (4.93)	2.68	0.37	ns	+15
TOWRE Sight Word Efficiency subtest	Grade 1	25/99	27.18 (11.60)	21.10 (9.62)	6.08	0.54	ns	+20
WRAT-R Reading	Grade 1	25/99	46.77 (8.93)	40.40 (6.34)	6.37	0.75	Statistically significant	+27
WRMT-R Word Attack subtest	Grade 1	25/99	14.70 (8.64)	8.25 (6.96)	6.45	0.77	Statistically significant	+28
WRMT-R Word Identification subtest	Grade 1	25/99	32.84 (13.46)	26.20 (9.87)	6.64	0.51	ns	+20
Vadasy, Jenkins, & Pool, 2000[7]								
Bryant Pseudoword	Grade 1	11/46	19.45 (11.65)	8.94 (7.79)	10.51	1.04	Statistically significant	+35
Dolch Word Recognition	Grade 1	11/46	144.74 (54.95)	102.67 (47.37)	42.07	0.81	Statistically significant	+29
W-J Word Attack subtest	Grade 1	11/46	109.27 (13.66)	94.12 (10.71)	15.15	1.21	Statistically significant	+39
WRAT-R: Reading	Grade 1	11/46	102.45 (18.81)	88.77 (11.38)	13.68	0.86	Statistically significant	+31
Vadasy, Sanders, & Peyton, 2005[7]								
TOWRE Phonemic Decoding subtest	Grade 1	57 students	93.60 (9.27)	88.40 (9.43)	5.20	0.55	ns	+21
TOWRE Sight Word Efficiency subtest	Grade 1	57 students	91.60 (9.47)	85.80 (10.48)	5.80	0.58	ns	+22

(continued)

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain by construct[1] *(continued)*

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
WRAT-R Reading subtest	Grade 1	57 students	99.40 (12.70)	86.30 (13.13)	13.10	1.01	Statistically significant	+34
WRMT-R Word Attack subtest	Grade 1	57 students	110.10 (10.53)	96.60 (12.92)	13.50	1.17	Statistically significant	+38
WRMT-R Word Identification subtest	Grade 1	57 students	104.20 (9.83)	93.90 (12.16)	10.30	0.95	Statistically significant	+33
Average for alphabetics (Mooney, 2003)[10]						**0.23**	**ns**	**+9**
Average for alphabetics (Vadasy et al., 1997a)[10]						**0.32**	**ns**	**+13**
Average for alphabetics (Vadasy & Sanders, 2008)[10]						**0.34**	**ns**	**+13**
Average for alphabetics (Vadasy, Sanders, & Peyton, 2006)[10]						**0.45**	**ns**	**+17**
Average for alphabetics (Jenkins et al., 2004)[10]						**0.68**	**Statistically significant**	**+25**
Average for alphabetics (Vadasy, Jenkins, & Pool, 2000)[10]						**0.97**	**Statistically significant**	**+33**
Average for alphabetics (Vadasy, Sanders, & Peyton, 2005)[10]						**0.85**	**Statistically significant**	**+30**
Domain average for alphabetics across all studies[10]						**0.55**	**na**	**+21**

ns = not statistically significant
na = not applicable
CTOPP = Comprehensive Test of Phonological Processes
DIBELS = Dynamic Indicators of Basic Early Literacy Skills
TOWRE = Test of Word Reading Efficiency
W-J = Woodcock-Johnson Psycho-Educational Battery
WJ-R = Woodcock-Johnson Psycho-Educational Battery–Revised
WRAT-R = Wide Range Achievement Test–Revised
WRMT-R = Woodcock Reading Mastery Test–Revised
WRMT-R/NU = Woodcock Reading Mastery Test–Revised/Normative Update

(continued)

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain by construct[1] *(continued)*

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the alphabetics domain.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of Vadasy and Sanders (2008), the mean difference represents the tutoring effect from the hierarchical linear model (HLM).
4. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of Vadasy and Sanders (2008), the effect sizes were reported by the authors and the WWC could not verify the calculation.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the cases of Vadasy and Sanders (2008), Vadasy et al. (1997a), Vadasy, Sanders, and Peyton (2006), Jenkins et al. (2004), and Vadasy, Jenkins, and Pool (2000), corrections for multiple comparisons were needed, and in the case of Vadasy, Sanders, and Peyton (2005), corrections for clustering and multiple comparisons were needed, so the significance levels may differ from those reported in the original studies. Mooney (2003) did not require adjustment for clustering or multiple comparisons. However, it is a randomized controlled trial that did not adjust for pretest differences. Thus, the means, effect sizes, improvement index, and statistical significance have been adjusted for pretest values using the difference-in-differences method. For an explanation of the difference-in-differences adjustment, see the WWC Procedures and Standards Handbook, Appendix B.
8. Vadasy and Sanders (2008) reported HLM-adjusted results. In this table, the treatment mean equals the comparison mean plus the intervention coefficient from the HLM analysis. The standard deviations were calculated by the WWC by combining the unadjusted posttest standard deviations from the two treatment groups. The statistical significance represents the statistical significance of the HLM coefficient as reported by the study authors.
9. Means and standard deviations for the combined treatment group were obtained by the WWC through communication with the author. The author provided unadjusted means and standard deviations.
10. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.

Appendix A3.2 Summary of study findings included in the rating for the fluency domain[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Mooney, 2003[7]								
DIBELS Oral Reading Fluency subtest	Grade 1	47 students	57.60 (38.20)	44.90 (32.50)	12.70	0.35	ns	+14
Vadasy et al., 1997a[7]								
Analytical Reading Inventory	Grade 1	13/40	33.16 (22.62)	29.55 (23.79)	3.61	0.15	ns	+6
Vadasy & Sanders, 2008[7,8]								
Passage Reading Rate	Kindergarten	30/86	10.32 (7.98)	6.84 (6.82)	3.48	0.48	Statistically significant	+18
Vadasy, Sanders, & Peyton, 2006[7]								
Passage Reading Rate	Kindergarten	19/67	6.00 (6.10)	2.00 (3.10)	4.00	0.80	Statistically significant	+29
Jenkins et al., 2004[7,9]								
Nonphonetically Controlled Passage Accuracy	Grade 1	25/99	0.81 (0.17)	0.73 (0.17)	0.08	0.47	ns	+17
Nonphonetically Controlled Passage Rate	Grade 1	25/99	36.13 (24.00)	26.35 (17.70)	9.78	0.42	ns	+16
Phonetically Controlled Passage Accuracy	Grade 1	25/99	0.81 (0.16)	0.71 (0.14)	0.10	0.63	ns	+24
Phonetically Controlled Passage Rate	Grade 1	25/99	41.30 (27.41)	27.70 (22.03)	13.60	0.51	ns	+20
Vadasy, Jenkins, & Pool, 2000[7]								
Analytical Reading Inventory: Primary	Grade 1	11/46	45.36 (34.77)	29.42 (18.19)	15.94	0.56	ns	+21
Analytical Reading Inventory: First Grade	Grade 1	11/46	36.57 (33.38)	25.43 (19.69)	11.14	0.40	ns	+16

(continued)

Appendix A3.2 Summary of study findings included in the rating for the fluency domain[1] *(continued)*

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Vadasy, Sanders, & Peyton, 2005[7]								
Passage Reading Accuracy	Grade 1	57 students	0.78 (0.13)	0.61 (0.25)	0.17	0.94	Statistically significant	+33
Passage Reading Rate	Grade 1	57 students	31.10 (17.49)	23.40 (22.73)	7.70	0.39	ns	+15
Average for fluency (Mooney, 2003)[10]						**0.35**	**ns**	**+14**
Average for fluency (Vadasy et al., 1997a)[10]						**0.15**	**ns**	**+6**
Average for fluency (Vadasy & Sanders, 2008)[10]						**0.48**	**Statistically significant**	**+18**
Average for fluency (Vadasy, Sanders, & Peyton, 2006)[10]						**0.80**	**Statistically significant**	**+29**
Average for fluency (Jenkins et al., 2004)[10]						**0.51**	**Statistically significant**	**+20**
Average for fluency (Vadasy, Jenkins, & Pool, 2000)[10]						**0.48**	**ns**	**+19**
Average for fluency (Vadasy, Sanders, & Peyton, 2005)[10]						**0.67**	**ns**	**+25**
Domain average for fluency across all studies[10]						**0.49**	**na**	**+19**

ns = not statistically significant
na = not applicable
DIBELS = Dynamic Indicators of Basic Early Literacy Skills

(continued)

Appendix A3.2 Summary of study findings included in the rating for the fluency domain[1] *(continued)*

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the fluency domain.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of Vadasy and Sanders (2008), the effect sizes were reported by the authors and the WWC could not verify the calculation.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the cases of Vadasy and Sanders (2008), Vadasy et al. (1997a), and Vadasy, Sanders, and Peyton (2006), no corrections for clustering or multiple comparisons were needed. In the cases of Jenkins et al. (2004) and Vadasy, Jenkins, and Pool (2000), a correction for multiple comparisons was needed, and in the case of Vadasy, Sanders, and Peyton (2005), a correction for clustering and multiple comparisons was needed, so the significance levels may differ from those reported in the original studies. Mooney (2003), did not require corrections for clustering or multiple comparisons. However, it is a randomized controlled trial that does not adjust for pretest differences. Thus, the means, effect sizes, improvement index, and statistical significance have been adjusted for pretest values using the difference-in-differences method. For an explanation of the difference-in-differences adjustment, see the WWC Procedures and Standards Handbook, Appendix B.
8. Vadasy and Sanders (2008) reported HLM-adjusted results. In this table, the treatment mean equals the comparison mean plus the intervention coefficient from the HLM analysis. The standard deviations were calculated by the WWC by combining the unadjusted posttest standard deviations from the two treatment groups. The statistical significance represents the statistical significance of the HLM coefficient as reported by the study authors.
9. Means and standard deviations for the combined treatment effect were obtained by the WWC through communication with the authors.
10. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.

Appendix A3.3 Summary of study findings included in the rating for the comprehension domain[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Vadasy & Sanders, 2008[7,8]								
WRMT-R/NU Passage Comprehension subtest	Kindergarten	30/86	96.26 (10.35)	92.38 (9.21)	3.88	0.41	Statistically significant	+16
Vadasy, Sanders, & Peyton, 2006[7]								
WRMT-R/NU Passage Comprehension subtest	Kindergarten	19/67	89.00 (7.40)	87.00 (6.80)	2.00	0.28	ns	+11
Jenkins et al., 2004[7,9]								
WRMT-R Passage Comprehension subtest	Grade 1	25/99	14.66 (6.58)	9.75 (6.66)	4.91	0.74	Statistically significant	+27
Vadasy, Sanders, & Peyton, 2005[7]								
WRMT-R/NU Passage Comprehension subtest	Grade 1	57 students	98.80 (8.00)	92.10 (10.30)	6.70	0.75	ns	+27
Domain average for comprehension across all studies[10]						**0.55**	**na**	**+21**

ns = not statistically significant
na = not applicable
WRMT-R = Woodcock Reading Mastery Test–Revised
WRMT-R/NU = Woodcock Reading Mastery Test–Revised/Normative Update

(continued)

Appendix A3.3 Summary of study findings included in the rating for the comprehension domain[1] *(continued)*

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the comprehension domain.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of Vadasy and Sanders (2008), the effect sizes were reported by the authors and the WWC could not verify the calculation.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the cases of Vadasy and Sanders (2008), Vadasy, Sanders, and Peyton (2006), and Jenkins et al. (2004), no corrections for clustering or multiple comparisons were needed. In the case of Vadasy, Sanders, and Peyton (2005), a correction for clustering was needed, so the significance levels may differ from those reported in the original study.
8. Vadasy and Sanders (2008) reported HLM-adjusted results. In this table, the treatment mean equals the comparison mean plus the intervention coefficient from the HLM analysis. The standard deviations were calculated by the WWC by combining the unadjusted posttest standard deviations from the two treatment groups. The statistical significance represents the statistical significance of the HLM coefficient as reported by the study authors.
9. Means and standard deviations for the combined treatment effect were obtained by the WWC through communication with the authors.
10. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.

Appendix A3.4 Summary of study findings included in the rating for the general reading achievement domain[1]

Outcome measure	Study sample	Sample size (students)	Author's findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Author's findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Mooney, 2003[7]								
WRMT-R/NU Total Reading	Grade 1	47	95.70 (14.90)	92.40 (14.00)	3.30	0.22	ns	+9
Domain average for general reading achievement[8]						**0.22**	**na**	**+9**

ns = not statistically significant

na = not applicable

WRMT-R/NU = Woodcock Reading Mastery Test–Revised/Normative Update

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the general reading achievement domain. Subscale findings from the same study are not included in these ratings, but are reported in Appendices A4.1 and A4.2.
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. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Mooney (2003), no corrections for clustering or multiple comparisons were needed. However, Mooney (2003) is a randomized controlled trial that does not adjust for pretest differences. Thus, the means, effect sizes, improvement index, and statistical significance have been adjusted for pretest values using the difference-in-differences method. For an explanation of the difference-in-differences adjustment, see the WWC Procedures and Standards Handbook, Appendix B.
8. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A4.1 Summary of subscale findings for the alphabetics domain[1]

Outcome measure	Study sample	Sample size (students)	Author's findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Author's findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Mooney, 2003[7]								
WRMT-R/NU Basic Reading Skills subtest	Grade 1	47	99.60 (14.00)	95.60 (15.10)	4.00	0.27	ns	+11

ns = not statistically significant

WRMT-R/NU = Woodcock Reading Mastery Test–Revised/Normative Update

1. This appendix presents subscale findings for measures that fall in the alphabetics domain. Aggregated scale 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. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools (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. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C. In the case of Mooney (2003), no correction for clustering was needed. However, Mooney (2003) is a randomized controlled trial that did not adjust for pretest differences. Thus, the means, effect sizes, improvement index, and statistical significance have been adjusted for pretest values using the difference-in-differences method. For an explanation of the difference-in-differences adjustment, see the WWC Procedures and Standards Handbook, Appendix B.

Appendix A4.2 Summary of subscale findings for the comprehension domain[1]

Outcome measure	Study sample	Sample size (students)	Author's findings from the study: Mean outcome (standard deviation)[2] — *Sound Partners* group	Author's findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Sound Partners* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Mooney, 2003[7]								
WRMT-R/NU Reading Comprehension subtest	Grade 1	47	92.70 (15.10)	88.30 (12.50)	4.40	0.31	ns	+12

ns = not statistically significant

WRMT-R/NU = Woodcock Reading Mastery Test–Revised/Normative Update

1. This appendix presents subscale findings for measures that fall in the comprehension domain. Aggregated scale 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. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
7. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools (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. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C. In the case of Mooney (2003), no correction for clustering was needed. However, Mooney (2003) is a randomized controlled trial that did not adjust for pretest differences. Thus, the means, effect sizes, improvement index, and statistical significance have been adjusted for pretest values using the difference-in-differences method. For an explanation of the difference-in-differences adjustment, see the WWC Procedures and Standards Handbook, Appendix B.

Appendix A5.1 *Sound Partners* rating for the alphabetics domain

The WWC rates an intervention's effects for 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 *Sound Partners* as having positive effects for beginning readers. The remaining ratings (potentially positive effects, mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered, as *Sound Partners* was assigned the highest applicable rating.

Rating received
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.
Met. Five studies showed statistically significant positive effects, two of which had a strong design.
AND
• Criterion 2: No studies showing statistically significant or substantively important *negative* effects.
Met. None of the 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. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A5.2 *Sound Partners* rating for the fluency domain

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

For the outcome domain of fluency, the WWC rated *Sound Partners* as having positive effects for beginning readers. The remaining ratings (potentially positive effects, mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered, as *Sound Partners* was assigned the highest applicable rating.

Rating received
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.
Met. Three studies showed statistically significant positive effects, two of which had a strong design.
AND
• Criterion 2: No studies showing statistically significant or substantively important *negative* effects.
Met. None of the 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. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A5.3 *Sound Partners* rating for the comprehension domain

The WWC rates an intervention's effects for 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 *Sound Partners* as having positive effects for beginning readers. The remaining ratings (potentially positive effects, mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered, as *Sound Partners* was assigned the highest applicable rating.

Rating received
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.
Met. Two studies showed statistically significant positive effects, one of which had a strong design.
AND
• Criterion 2: No studies showing statistically significant or substantively important *negative* effects.
Met. None of the 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. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A5.4 *Sound Partners* rating for the general reading achievement domain

The WWC rates an intervention's effects for 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 *Sound Partners* as having no discernible effects for beginning readers.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: No studies showing a statistically significant or substantively important effect, either *positive* or *negative*.

  Met. Only one study examined an outcome in general reading achievement, and the effect was not statistically significant or substantively important.

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 study showed a statistically significant positive effect.

AND

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

  Met. No study showed a statistically significant or substantively important negative effect.

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

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

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

AND

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

  Met. No study showed a statistically significant or substantively important negative effect.

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 study showed a statistically significant or substantively important positive effect.

OR

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

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

(continued)

Appendix A5.4 *Sound Partners* rating for the general reading achievement domain *(continued)*

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

- Criterion 1: One study showing a statistically significant or substantively important *negative* effect and no studies showing a statistically significant or substantively important *positive* effect.

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

OR

- Criterion 2: Two or more studies showing statistically significant or substantively important *negative* effects, at least one study showing a statistically significant or substantively important *positive* effect, and more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

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

Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.

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

  Not met. No study showed a statistically significant negative effect.

AND

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

  Met. No study showed a statistically significant or substantively important positive effect.

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

Appendix A6 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Alphabetics	7	59	442	Medium to large
Fluency	7	59	442	Medium to large
Comprehension	4	44	309	Medium to large
General reading achievement	1	7	47	Small

1. A rating of "medium to large" requires at least two studies and two schools across studies in one domain and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small." For more details on the extent of evidence categorization, see the WWC Procedures and Standards Handbook, Appendix G.