Appendix

Appendix A1.1 Study Characteristics: Baenen, Bernhold, Dulaney, and Banks, 1997 (randomized controlled trial)

Characteristic	Description
Study citation	Baenen, N., Bernhold, A., Dulaney, C., & Bankes, K. (1997). *Reading Recovery*: Long-term progress after three cohorts. *Journal of Education for Students Placed at Risk, 2*(2), 161.
Participants	The study involved 772 first-grade students who were studied in four cohorts between 1990 and 1994. Of this total number, only one cohort (that received treatment in 1990–91 at 10 schools[1]) with 168 students meets WWC criteria for inclusion. After 1990–91, the comparison group was made up of students who were not comparable to the intervention group in terms of their achievement levels—they were not the lowest-achieving students in participating schools, as the *Reading Recovery®* students were. Due to attrition, the authors' final analysis sample included 147 first-grade students in the 1990–91 cohort.[2] All 147 students were followed longitudinally into second grade; 127 of these students were followed and included in the third-grade analysis. No information was reported regarding ethnicity or gender, but all students were low achieving, and a high percentage of students qualified for free or reduced-price lunch (N. Baenen, personal communication, January 9, 2006).
Setting	The study took place in 10 elementary schools in Wake County public schools in North Carolina.
Intervention	The intervention group was originally composed of 84 students who qualified for *Reading Recovery®* on the basis of three subtests from the Observation Survey of Early Literacy Achievement—Text Reading Level (running record), Dictation, and Writing Vocabulary. Intervention students, among the lowest-achieving students at their schools, were randomly assigned to receive the *Reading Recovery®* intervention. They were taught by one of 12 teachers. Results were presented for 72 students for first- and second-grade analyses. Of these, 27% of the students received a "partial program" instead of the full set of more than 60 lessons in *Reading Recovery®*.
Comparison	The comparison group was composed of 84 students who qualified for *Reading Recovery®* on the basis of three subtests from the Observation Survey of Early Literacy Achievement—Text Reading Level (running record), Dictation, and Writing Vocabulary. These students, among the lowest-achieving students at their schools, were randomly assigned to a wait list for *Reading Recovery®* intervention. They did not receive *Reading Recovery®* during the time of the study but received the regular services available to them. Results were presented for 75 students for first- and second-grade analyses and 68 students for third-grade analysis.
Primary outcomes and measurement	Grade retention was measured at the end of first and second grade, and the North Carolina End-of-Grade test in reading was used at the end of third grade. The authors also measured referral to special education and Title I services and teacher perception of student achievement, but these outcomes are not included in this report because they are not specified by the WWC Beginning Reading Protocol. The authors used three subtests of the Observation Survey of Early Literacy Achievement, but these measures are not included in the review because collection of these data did not focus on the full sample of students. For a more detailed description of the outcome measures, see Appendix A2.4.
Staff/teacher training	*Reading Recovery®* teachers participated in a graduate-level course that included 33 training sessions from September to June of the intervention year. The 1990–91 school year was the first year of implementing *Reading Recovery®* for the district; therefore, it was the first year the teachers taught the intervention.

1. One of the manuscripts related to this study reported 11 schools.
2. Pretest comparability of the students was demonstrated for only 146 students. It is reasonable to assume that the pretest mean for the 146 students would be similar to that based on 147 students.

Appendix A1.2 Study Characteristics: Pinnell, DeFord, and Lyons, 1988 (randomized controlled trial)

Characteristic	Description
Study citation	Pinnell, G. S., DeFord, D. E., & Lyons, C. A. (1988). *Reading Recovery: Early intervention for at-risk first graders* (Educational Research Service Monograph). Arlington, VA: Educational Research Service.
Participants	The study involved 187 first-grade students from 14 schools. Although information about the specific schools included in the study was not presented, the district has a racial composition of 45% nonwhite students and a gender composition of 51% male students. The city has a mobility rate of approximately 20%, and 66% of students receive free or reduced-price lunch. Students in the study were low achieving, as defined by scoring in the lowest 20% of their class on reading measures. Due to attrition, results were reported for 184 students.
Setting	The study took place in 14 urban public schools in Columbus, Ohio.
Intervention	There were two intervention groups with a total of 134 students taught by 32 teachers. One group was randomly assigned to receive the standard *Reading Recovery®* pull-out program (n = 38). These students had regular classroom teachers who were not trained in *Reading Recovery®*. This group experienced attrition of one student. A second group of students determined to be eligible for *Reading Recovery®* received the standard *Reading Recovery®* pull-out program, with the addition of having regular classroom teachers trained in *Reading Recovery®* (n = 96). The second group was not randomly assigned to *Reading Recovery®* or to their classroom teacher, so this portion of the study is considered a quasi-experimental design. It is not included in the intervention rating because the second intervention group with a *Reading Recovery®*–trained teacher as its regular classroom teacher goes beyond the standard implementation of the program.[1]
Comparison	The comparison group was composed of low-achieving students who did not have regular classroom teachers trained in *Reading Recovery®* and who were randomly assigned to an alternative compensatory program (n = 53) (G. S. Pinnell, personal communication, September 9, 2006). Students in this group received a series of skill-oriented drill activities conducted in small groups or individual sessions of 30–45 minutes from a trained paraprofessional. The comparison group was not statistically different from either intervention group on pretest measures (G. S. Pinnell, personal communication, December 5, 2005). The final analysis sample included 51 students.
Primary outcomes and measurement	Five subtests of the Observation Survey of Early Literacy Achievement were included—Letter Identification, Word Recognition, Concepts About Print, Writing Vocabulary, and Dictation. Additional measures included a writing assessment, the Reading Vocabulary subtest of the Comprehensive Test of Basic Skills (CTBS), and the Reading Comprehension subtest of the CTBS. Results from the Observation Survey: Text Reading Level subtest were not reported because effect sizes that were comparable to other measures could not be calculated.[2] For a more detailed description of these outcome measures, see Appendices A2.1–A2.4.
Staff/teacher training	Teachers received a full year of special training, during which they practiced teaching and observed other teachers through a one-way mirror. Twelve teachers received training from a university program and were in their second year of teaching the intervention during the time of the study. The other 20 teachers received training from a local teacher leader and were in their first year of teaching the intervention during the time of the study.

1. Results are reported in Appendices A4.1–4.3
2. Findings based on the Observation Survey: Text Reading Level subtest are not included in the effectiveness ratings because effect sizes and the statistical significance of the findings could not be calculated given the information provided in the study. The Observation Survey: Text Reading Level subtest is reported as reading levels based on ordinal, rather than equal-interval, scales. For example, the increase in fluency measured by scoring at level 3 compared with level 2 on the scale may not be equal to the increase in fluency as measured by scoring at level 24 compared with level 23. The author no longer had information on the number of students scoring at each level. For more detail, see Denton, C.A., Ciancio, D. J., & Fletcher, J. M. (2006). Validity, reliability, and utility of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly, 41*(1) 8–34.

Appendix A1.3 Study Characteristics: Pinnell, Lyons, DeFord, Bryk, and Seltzer, 1994 (randomized controlled trial)

Characteristic	Description
Study citation	Pinnell, G. S., Lyons, C. A., DeFord, D. E., Bryk, A. S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly, 29*(1), 8–39.
Participants	The study was designed to examine outcomes of 403 first-grade students distributed across 43 schools. Percentages of children receiving Aid to Dependent Children ranged between 9% and 42% among districts. For this report, the WWC looked at results for students in 10 districts at schools that were using *Reading Recovery®*. In those schools, eligible students were randomly assigned to either the *Reading Recovery®* group or the comparison group, which did not receive any special instruction. For this report, the WWC did not look at the portions of the study that examined results for students who received three additional reading interventions—Reading Success, Direct Instruction Skills Plan, and Reading and Writing Group. These were delivered at schools other than the *Reading Recovery®* schools and were compared with comparison students in those schools. In the original study design, 100 students at 10 schools were randomly assigned to receive either *Reading Recovery®* or the comparison condition. However, random assignment was not successfully implemented at two schools, and there was minor attrition at the remaining schools, resulting in a final analytic sample of 79 students from eight schools. Treatment and comparison students attended the same schools. All students were low achieving.
Setting	The study took place in 10 school districts (two rural, two suburban, and six urban) in Ohio.
Intervention	The intervention group was composed of 31 low-achieving students from one school in each of the districts that already had a *Reading Recovery®* program in place. Intervention students participated in the *Reading Recovery®* curriculum.
Comparison	The comparison group included 48 students who were randomly assigned to receive no special instruction but continued to participate in their regular reading program. These students attended the same schools as did the intervention students.
Primary outcomes and measurement	The Dictation subtest of the Observation Survey of Early Literacy Achievement was administered in addition to the Woodcock Reading Mastery Test–Revised and the Gates-MacGinitie Reading Test. Results from the Observation Survey: Text Reading Level subtest were not reported because effect sizes that were comparable to other measures could not be calculated.[1] For a more detailed description of these outcome measures, see Appendices A2.1–A2.4.
Staff/teacher training	At least two years prior to the study, *Reading Recovery®* teachers received specialized training through weekly 2.5-hour sessions for one year, during which they practiced teaching and observed other teachers through a one-way mirror.

1. Findings based on the Observation Survey: Text Reading Level subtest are not included in the effectiveness ratings because effect sizes and the statistical significance of the findings could not be calculated given the information provided in the study. The Observation Survey: Text Reading Level subtest is reported as reading levels based on ordinal, rather than equal-interval, scales. For example, the increase in fluency measured by scoring at level 3 compared with level 2 on the scale may not be equal to the increase in fluency as measured by scoring at level 24 compared with level 23. The author no longer had information on the number of students scoring at each level. For more detail, see Denton, C. A., Ciancio, D. J., & Fletcher, J. M. (2006). Validity, reliability, and utility of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly, 41*(1), 8–34.

Appendix A1.4 Study Characteristics: Schwartz, 2005 (randomized controlled trial)

Characteristic	Description
Study citation	Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the *Reading Recovery®* early intervention. *Journal of Educational Psychology, 97*(2), 257–267.
Participants	The study was designed to examine the outcomes of first-grade students distributed across 47 *Reading Recovery®* teachers. Each teacher identified for random assignment two students eligible for *Reading Recovery®* based on their low scores on the Observation Survey of Early Literacy Achievement. These 94 students were randomly assigned to enter the *Reading Recovery®* program during either the first or second half of the school year. The study included two additional comparison groups of 47 low-average and 47 high-average readers from the same classrooms as the *Reading Recovery®* students who were not expected to participate in the *Reading Recovery®* program. Analysis involving these comparison groups was not eligible for WWC review because the WWC considers only comparisons of students with similar achievement backgrounds in assessing the effectiveness of *Reading Recovery®*. Because of missing testing data, the author's final analysis of the randomly assigned groups included 74 students distributed across 37 teachers.
Setting	The study took place in elementary schools in 14 states.
Intervention	The intervention group was composed of the lowest 20%–30% of students who qualified for *Reading Recovery®* based on the six subtests of the Observation Survey and on judgments by teachers. All teachers agreed to dedicate one of their four *Reading Recovery®* slots to the study, allowing random assignment of one of two eligible students to receive intervention delivery in the first half of the year. Thus, students in the intervention group were randomly assigned to receive *Reading Recovery®* during the first half of the year. Originally, participants were taught by 47 *Reading Recovery®* teachers who had volunteered to be part of the study, but because of missing information, data from only 37 teachers and 37 students were included in the author's final analysis. The intervention group was 61% male, 38% white, 47% black, 12% Hispanic, and 3% Asian. About 60% of the group received free or reduced-price lunch.
Comparison	The comparison group of interest to the WWC was composed of the lowest 20%–30% of students who qualified for *Reading Recovery®* based on six subtests of the Observation Survey and on judgments by teachers. Participants in the comparison group were randomly assigned to receive *Reading Recovery®* during the second half of the year. Thus, these participants served as comparisons only during the first part of the year when they received instruction in their regular classroom. Data from 37 teachers and 37 students were included in the author's final analysis. The comparison group was 41% male, 47% white, 38% black, and 15% Hispanic. Approximately 57% of the group received free or reduced-price lunch.
Primary outcomes and measurement	Six subtests of the Observation Survey of Early Literacy Achievement were administered—Letter Identification, Word Recognition, Concepts About Print, Writing Vocabulary, Dictation, and Text Reading Level.[1] Other assessments included the Yopp-Singer Phoneme Segmentation Task, the deletion task, the Slosson Oral Reading Test–Revised, and the Degrees of Reading Power Test. For a more detailed description of these outcome measures, see Appendices A2.1–A2.4.
Staff/teacher training	No information on training for the specific teachers participating in this study was provided.

1. For many studies, findings based on the Observation Survey: Text Reading Level subtest were not included in the effectiveness ratings because effect sizes and the statistical significance of the findings could not be calculated given the information provided in the study. In this case, however, the author was able to provide the number of students scoring at each level (R. M. Schwartz, personal communication, December 5, 2006). The outcome was calculated as a logged-odds ratio, with the number of students reading at or above the first-grade level compared with students reading below the first-grade level.

Appendix A1.5 Study Characteristics: Iverson and Tunmer, 1993 (quasi-experimental design)

Characteristic	Description
Study citation	Iverson, S., & Tunmer, W. E. (1993). Phonological processing skills and the *Reading Recovery®* program. *Journal of Educational Psychology, 85*(1), 112–126.
Participants	The study included 160 first-grade students distributed across 47 classrooms in 30 schools. Ninety-six students who performed at the lowest levels on tests administered at the beginning of first grade were placed into one of three matched groups: a standard *Reading Recovery®*group, a modified *Reading Recovery®* group, and a comparison group that did not participate in *Reading Recovery®*. Students across the three groups were matched based on pretest scores from the Letter Identification and Dictation subtests of the Observation Survey. The study also compared the *Reading Recovery®* groups with comparison groups consisting of 64 children from the same classrooms as the *Reading Recovery®* participants who were deemed by the classroom teachers to be performing at average reading level. But analyses involving these comparison groups were not eligible for this WWC review because the WWC considers only comparisons of students with similar achievement backgrounds to provide information on the effectiveness of *Reading Recovery®*.
Setting	Thirty schools from 13 school districts in Rhode Island.
Intervention	The study authors used two intervention groups drawn from the 23 schools operating *Reading Recovery®* programs. Sixty-four low-scoring students from 34 classrooms were assigned to receive standard or modified *Reading Recovery®*. The 32 students in the standard *Reading Recovery®* group received one-on-one lessons that followed the procedures described by the program developer, which included a letter identification segment as needed. Once a child mastered letter identification, time allocated to this segment could be used for incidental word analysis in other segments. For the 32 students in the modified *Reading Recovery®* program, explicit instruction in letter-phoneme patterns replaced the letter identification segment beginning in the fourth week of the program. The WWC did not include the second comparison group in its intervention ratings because it was a modified version of the standard program.[1]
Comparison	The matched comparison group of interest to the WWC was drawn from the seven schools that did not operate *Reading Recovery®* programs. Thirty-two low-scoring students from 13 classrooms were matched to the intervention groups based on pretest scores. Students in the comparison group received the support services normally available to at-risk readers, which consisted of small group instruction provided at least four days a week through either the federally funded Chapter One program or the state-supported literacy program.
Primary outcomes and measurement	Five subtests of the Observation Survey of Early Literacy Achievement were included—Letter Identification, Word Recognition, Concepts About Print, Writing Vocabulary, and Dictation. Additional measures included the Dolch Word Recognition Test, Yopp-Singer Phoneme Segmentation Test, a phoneme deletion task, and a pseudoword decoding task. Results from the Observation Survey: Text Reading Level subtest were not reported because effect sizes that were comparable to other measures could not be calculated.[2] (See Appendices A2.1, A2.2, and A2.4 for more detailed descriptions of outcome measures.)
Staff/teacher training	*Reading Recovery®* teachers in the study were reading specialists with master's degrees. To learn how to administer the assessments, teachers participated in 4 two-hour training sessions. To learn how to deliver the intervention, teachers attended weekly two-hour classes for the duration of the study. During the session, teachers were able to practice teaching and observe other teachers through a one-way mirror. They also received feedback from the teacher leader during individual visits to the classes.

1. Results are reported in Appendices A4.1 and A4.3.
2. Findings based on the Observation Survey: Text Reading Level subtest are not included in the effectiveness ratings because effect sizes and the statistical significance of the findings could not be calculated given the information provided in the study. The Observation Survey: Text Reading Level subtest is reported as reading levels based on ordinal, rather than equal-interval, scales. For example, the increase in fluency measured by scoring at level 3 compared with level 2 on the scale may not be equal to the increase in fluency as measured by scoring at level 24 compared with level 23. The author no longer had information on the number of students scoring at each level. For more detail, see Denton, C. A., Ciancio, D. J., & Fletcher, J. M. (2006). Validity, reliability, and utility of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly, 41*(1), 8–34.

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

Outcome measure	Description
Phonemic awareness	
Deletion task	A 10-item version of the Rosner task that requires students to repeat a word and then say it again after omitting a given syllable or sound. The assessment is not standardized (as cited in Schwartz, 2005).
Phoneme deletion task	Developed by Calfee, this measure requires students to delete the first consonant segment in a word and read the word aloud. It includes four training lists and six transfer lists that increase in difficulty as students move through each list. The most difficult lists consist of both real words and pseudowords. Student scores are based on the number of correct words read from the transfer lists (as cited in Iverson & Tunmer, 1993).
Yopp-Singer Phoneme Segmentation Test	Developed by Yopp, the test is an orally administered assessment. A teacher works with each student individually and introduces the test as a word game. The teacher has a list of 22 words that the student is not allowed to see. After the teacher reads each word, the student must repeat all of the sounds in the word separately and slowly (as cited in Iverson & Tunmer, 1993; Schwartz, 2005).
Print awareness	
Observation Survey of Early Literacy Achievement: Concepts About Print subtest[1]	Students perform tasks related to printed language concepts (for example, directionality and word concepts) while reading a book. This assessment, developed by Clay, is not standardized (as cited in Iverson & Tunmer, 1993; Pinnell, DeFord, & Lyons, 1988; Schwartz, 2005).
Letter knowledge	
Observation Survey of Early Literacy Achievement: Letter Identification subtest[1]	Students identify upper- and lowercase letters. This assessment, developed by Clay, is not standardized (as cited in Iverson & Tunmer, 1993; Pinnell et al., 1988; Schwartz, 2005).
Phonics	
Dolch Word Recognition Test	A standardized, individually administered word recognition test that contains a list of 220 short, frequently occurring words. Students are asked to read aloud from the list of words to measure their automatic word recognition and word retrieval skills, not their ability to sound out words (as cited in Iverson & Tunmer, 1993).
Observation Survey of Early Literacy Achievement: Word Recognition subtest (also known as Ready to Read or Ohio Word Test)[1]	Students read 20 common sight words from basic reading texts, and their accuracy is scored. This assessment, developed by Clay, is not standardized (as cited in Iverson & Tunmer, 1993; Pinnell et al., 1988; Schwartz, 2005).
Pseudoword decoding task	A 40-item test that requires students to name single-syllable synthetic words that conform to the rule of English orthography. The assessment is not standardized (as cited in Iverson & Tunmer, 1993).

1. The Clay Observation Survey was developed by Dr. Marie Clay, who also developed *Reading Recovery®*. Although there is no evidence of obvious overalignment between the measure and the intervention (intervention student receiving exposure to the measure during the course of treatment), it should be noted that the same person developed the intervention and the measure.

Appendix A2.2 Outcome measures for the fluency domain

Outcome measure	Description
Observation Survey of Early Literacy Achievement: Text Reading Level subtest [1, 2]	The percentage of students scoring at the first-grade reading level or higher compared with those scoring lower than first grade. To determine this, students read from passages of increasing difficulty, and student's error rate and self-correcting behavior are recorded using the running record technique. Students read from leveled texts drawn from a basal reading series until their accuracy rate falls below 90%. Results are translated to a numerical reading level from level 1 to level 30, which in turn matches up to grade-level equivalency. This assessment method is not standardized (as cited in Schwartz, 2005).
Slosson Oral Reading Test–Revised (SORT–R3)	Developed by Nicholson, the measure consists of 200 words arranged in order of difficulty, with 20 words per list. Each list represents an approximate reading grade level (for example, list one is equivalent to first grade). Administration ends after all the words on one list are missed. The measure is standardized and norm-referenced (as cited in Schwartz, 2005).

1. The Clay Observation Survey was developed by Dr. Marie Clay, who also developed *Reading Recovery®*. Although there is no evidence of obvious overalignment between the measure and the intervention (intervention student receiving exposure to the measure during the course of treatment), it should be noted that the same person developed the intervention and the measure.
2. Findings based on the Observation Survey: Text Reading Level subtest are not included in the effectiveness ratings for most studies because effect sizes and the statistical significance of the findings could not be calculated given the information provided in the studies. The author of one study (Schwartz, 2005) was able to provide the percentage of students scoring at each level. The outcome was changed to a logged-odds ratio for students scoring at the first-grade level or higher compared with those scoring lower than first grade.

Appendix A2.3 Outcome measures for the comprehension domain by construct

Outcome measure	Description
Reading comprehension	
Comprehensive Test of Basic Skills (CTBS): Reading Comprehension subtest	A group-administered, standardized assessment of reading comprehension (as cited in Pinnell, DeFord, & Lyons, 1988).
Degrees of Reading Power Test	An untimed standardized test requiring students to read a nonfiction passage with a word or set of words missing. Students select an appropriate answer to complete the sentence from a set of four or five alternatives (as cited in Schwartz, 2005).
Vocabulary development	
Comprehensive Test of Basic Skills (CTBS): Reading Vocabulary subtest	A group-administered, standardized assessment of vocabulary (as cited in Pinnell et al., 1988).

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

Outcome measure	Description
Gates-MacGinitie (1978)	A standardized test that covers vocabulary and comprehension aspects of reading. It evaluates students' abilities to decode initial consonants, consonant clusters, final consonants, and vowels in real English words and also measures their ability to recognize commonly used words without decoding. For reading comprehension, answer choices are given as pictures and words (as cited in Pinnell, Lyons, DeFord, Bryk, & Seltzer, 1994).
North Carolina End-of-Grade reading test	A standardized state assessment designed to match the North Carolina curriculum. It uses multiple-choice questions with reading passages and is designed to measure reading and thinking skills (as cited in Baenen, Bernhole, Dulaney, & Banks, 1997).
Observation Survey of Early Literacy Achievement: Hearing and Recording Sounds in Words (Dictation) subtest[1]	Students write the words that are dictated to them in sentence form. This assessment, developed by Clay, is not standardized (as cited in Iverson & Tunmer, 1993; Pinnell, DeFord, and Lyons, 1988; Pinnell et al., 1994; Schwartz, 2005).
Observation Survey of Early Literacy Achievement: Writing Vocabulary subtest[1]	Students are given 10 minutes to write as many words as they can on a blank sheet of paper. If needed, a standard set of prompts is used to encourage additional attempts to write. The measure is scored by counting the number of correctly spelled words (as cited in Iverson & Tunmer, 1993; Pinnell et al., 1988; Schwartz, 2005).
Retention	The percentage of students who were retained in their current grade (as cited in Baenen et al., 1997).
Woodcock Reading Mastery Test–Revised	A standardized test composed of six subtests. The subtests measure ability to form associations between visual stimuli and oral responses; ability to recognize upper- and lowercase letters in a variety of fonts; ability to read words aloud; ability to read aloud nonsense words or uncommon words to test phonic and structural analysis skills for pronouncing unfamiliar words; vocabulary ability through the use of antonyms, synonyms, and analogies; and passage comprehension by filling in missing words in a short paragraph (as cited in Pinnell et al., 1994).

1. The Clay Observation Survey was developed by Dr. Marie Clay, who also developed *Reading Recovery®*. Although there is no evidence of obvious overalignment between the measure and the intervention (intervention student receiving exposure to the measure during the course of treatment), it should be noted that the same person developed the intervention and the measure.

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

	Domain							
	Alphabetics				Fluency	Comprehension		General reading achievement
	Phonemic awareness	Print awareness	Letter knowledge	Phonics		Reading comprehension	Vocabulary development	
Met evidence standards								
Baenen et al., 1997	nr	nr	nr	nr	nr	nr	nr	ind
Pinnell et al., 1988	nr	+	ind	(+)	nr	+	+	+
Pinnell et al., 1994	nr	nr	nr	nr	nr	nr	nr	+
Schwartz, 2005	(+)	+	(+)	+	+	ind	nr	+
Met evidence standards with reservations								
Iverson and Tunmer, 1993	+	+	+	+	nr	nr	nr	+
Rating of Effectiveness	Positive				Potentially Positive	Potentially Positive		Positive

nr = no reported outcomes under this domain or 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 average improvement index in each domain. More detailed information on findings for the measures within the domains and the constructs that factor into the domains are reported in Appendices A3.2–A3.5.

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

Outcome measure	Study sample	Sample size	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Phonemic awareness construct								
Schwartz, 2005 (randomized controlled trial)[7]								
Deletion task[8]	Grade 1	74 students	6.64 (2.56)	5.58 (2.50)	1.06	0.41	ns	+16
Yopp-Singer Phoneme Segmentation Test[8]	Grade 1	74 students	17.70 (4.93)	15.27 (5.43)	2.43	0.46	ns	+18
Iverson and Tunmer, 1993 (quasi-experimental design)—standard *RR* group vs. comparison group[7]								
Phoneme deletion task[8]	Grade 1	64 students 14 schools	14.00 (6.16)	3.91 (5.31)	10.09	1.73	Statistically significant	+46
Yopp-Singer Phoneme Segmentation Test[8]	Grade 1	64 students 14 schools	17.63 (4.46)	6.21 (5.18)	11.42	2.33	Statistically significant	+49
Print awareness construct								
Pinnell, DeFord, and Lyons, 1988 (randomized controlled trial)—*RR* group without *RR* teachers vs. comparison group[7]								
Observation Survey: Concepts About Print subtest[9]	Grade 1	88 students	16.80 (2.91)	13.98 (3.31)	2.82	0.89	Statistically significant	+31
Schwartz, 2005 (randomized controlled trial)[7]								
Observation Survey: Concepts About Print subtest[9]	Grade 1	74 students	19.24 (2.55)	16.68 (2.30)	2.56	1.04	Statistically significant	+31
Iverson and Tunmer, 1993 (quasi-experimental design)—standard *RR* group vs. comparison group[7]								
Observation Survey: Concepts About Print subtest[9,10]	Grade 1	64 students 14 schools	18.12 (1.74)	13.72 (3.49)	4.40	1.58	Statistically significant	+44
Letter knowledge construct								
Pinnell, DeFord, and Lyons, 1988 (randomized controlled trial)—*RR* group without *RR* teachers vs. comparison group[7]								
Observation Survey: Letter Identification subtest[9]	Grade 1	88 students	48.05 (1.41)	49.61 (8.33)	–1.56	–0.24	ns	–10
Schwartz, 2005 (randomized controlled trial)[7]								
Observation Survey: Letter Identification subtest[9,10]	Grade 1	74 students	52.28 (1.27)	51.68 (2.78)	0.60	0.23	ns	+9

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

Outcome measure	Study sample	Sample size	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Iverson and Tunmer, 1993 (quasi-experimental design)—standard *RR* group vs. comparison group[7]								
Observation Survey: Letter Identification subtest[9]	Grade 1	64 students 14 schools	53.53 (1.01)	49.00 (6.63)	4.53	0.94	Statistically significant	+33
Phonics construct								
Pinnell, DeFord, and Lyons, 1988 (randomized controlled trial)—*RR* group without *RR* teachers vs. comparison group[7]								
Observation Survey: Word Recognition subtest[9]	Grade 1	88 students	13.29 (1.63)	11.98 (3.92)	1.31	0.41	ns	+16
Schwartz, 2005 (randomized controlled trial)[7]								
Observation Survey: Word Recognition subtest[9,10]	Grade 1	74 students	14.96 (3.99)	8.87 (4.75)	6.09	1.37	Statistically significant	+41
Iverson and Tunmer, 1993 (quasi-experimental design)—standard *RR* group vs. comparison group[7]								
Dolch Word Recognition Test[9]	Grade 1	64 students 14 schools	94.87 (29.07)	20.22 (13.93)	74.65	3.24	Statistically significant	+50
Observation Survey: Word Recognition subtest[9]	Grade 1	64 students 14 schools	12.66 (1.85)	5.56 (3.83)	7.10	2.33	Statistically significant	+49
Pseudoword decoding task[8]	Grade 1	64 students 14 schools	9.00 (7.07)	1.41 (2.31)	7.59	1.43	Statistically significant	+42
Average for alphabetics (Pinnell, DeFord, & Lyons, 1988)[11]						**0.35**	**ns**	**+14**
Average for alphabetics (Schwartz, 2005)[11]						**0.70**	**Statistically significant**	**+26**
Average for alphabetics (Iverson & Tunmer, 1993)[11]						**1.94**	**Statistically significant**	**+47**
Domain average for alphabetics across all studies[11]						**1.00**	**na**	**+34**

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the alphabetics domain. Additional findings from the same studies are not included in these ratings but are reported in Appendix A4.1.
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 Technical Details of WWC-Conducted Computations.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.

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

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.
7. 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. For the formulas the WWC used to calculate statistical significance, see Technical Details of WWC-Conducted Computations. In the case of Schwartz (2005), the author applied a correction for multiple comparisons, so the WWC did not apply a correction. In the case of Pinnell, DeFord, and Lyons (1988), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study. In the case of Iverson and Tunmer (1993), corrections for clustering and multiple comparisons were needed, so the significance levels may differ from those reported in the original study.
8. Means presented for these measures are the posttest means for each group (pretest data were not available in the study report).
9. The *RR* group mean equals the comparison group mean plus the mean difference. The computation of the mean difference took into account the pretest difference between the study groups.
10. The pretest data were based on 36 of the 37 pairs of students in this group. It is reasonable to assume that the pretest mean for the 36 pairs of students would be similar to that based on 37 pairs.
11. 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 fluency domain[1]

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Schwartz, 2005 (randomized controlled trial)[7]								
SORT–R3[8]	Grade 1	74	30.58 (14.41)	18.12 (11.87)	12.46	0.93	Statistically significant	+32
Observation Survey: Text Reading subtest[9]	Grade 1	74	0.78	0.05	0.73	2.49	Statistically significant	+49
Average for fluency (Schwartz, 2005)[10]						**1.71**	**Statistically significant**	**+46**

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 Technical Details of WWC-Conducted Computations.
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.
7. 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. For the formulas the WWC used to calculate statistical significance, see Technical Details of WWC-Conducted Computations. In the case of Schwartz (2005), the author applied corrections for multiple comparisons, so the WWC did not apply an additional correction.
8. Means presented for these measures are the posttest means for each group (pretest data were not available in the study report).
9. Means presented for this measure are the posttest proportions for each group scoring at or above a first-grade reading level. Effect size is computed as a Cox Index: logged-odds ratio transformation divided by 1.65. See Technical Details of WWC-Conducted Computations for the computation of effect sizes for binary outcomes.
10. This row provides the study average, which, in this instance, is also the domain average. The WWC-computed average effect sizes 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 comprehension domain by construct[1]

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Reading comprehension construct								
Pinnell, DeFord, and Lyons, 1988 (randomized controlled trial)—*RR* group without *RR* teachers vs. comparison group[7]								
CTBS: Reading Comprehension subtest[8]	Grade 1	82	36.67 (19.27)	27.33 (13.94)	9.34	0.56	Statistically significant	+21
Schwartz, 2005 (randomized controlled trial)[7]								
Degrees of Reading Power Test[8]	Grade 1	74	4.82 (3.88)	4.27 (3.88)	0.55	0.14	ns	+6
Vocabulary development construct								
Pinnell, DeFord, and Lyons, 1988 (randomized controlled trial)—*RR* group without *RR* teachers vs. comparison group[7]								
CTBS: Reading Vocabulary subtest[8]	Grade 1	81	36.64 (11.93)	28.07 (17.00)	8.57	0.57	Statistically significant	+21
Average for comprehension (Pinnell, DeFord, & Lyons, 1998)[9]						**0.56**	**Statistically significant**	**+21**
Average for comprehension (Schwartz, 2005)[8]						**0.14**	**ns**	**+6**
Domain average for comprehension across all studies[8]						**0.35**	**na**	**+14**

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the comprehension domain. Additional findings from the same studies are not included in these ratings but are reported in Appendix 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 Technical Details of WWC-Conducted Computations.
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.
7. 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. For the formulas the WWC used to calculate statistical significance, see Technical Details of WWC-Conducted Computations. In the case of Schwartz (2005), the author applied corrections for multiple comparisons, so the WWC did not apply an additional correction. In the case of Pinnell, DeFord, and Lyons (1988), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study.
8. Means presented for these measures are the posttest means for each group (pretest data were not available in the study report).
9. 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 A3.5 Summary of study findings included in the rating for the general reading achievement domain[1]

Outcome measure	Study sample	Sample size	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Baenen, Bernhole, Dulaney, and Banks, 1997 (randomized controlled trial)[7]								
Outcome time 1—end of first grade								
Retention[8]	Grade 1	147 students	0.06	0.05	0.01	–0.12	ns	–5
Pinnell, DeFord, and Lyons, 1988 (randomized controlled trial)—*RR* group without *RR* teachers vs. comparison group[7]								
Observation Survey: Dictation subtest[9]	Grade 1	88 students	30.52 (6.13)	23.80 (7.99)	6.72	0.92	Statistically significant	+32
Observation Survey: Writing Vocabulary subtest[9]	Grade 1	88 students	33.21 (13.49)	25.37 (14.33)	7.84	0.56	Statistically significant	+21
Pinnell, Lyons, DeFord, Bryk, and Seltzer, 1994 (randomized controlled trial)[11]								
Gates-MacGinitie[10]	Grade 1	79 students	36.19 (13.12)	31.00 (na)	5.19 (10.16)	0.51	Statistically significant	+19
Observation Survey: Dictation subtest[10]	Grade 1	79 students	31.74 (6.18)	26.75 (na)	4.99 (7.67)	0.65	Statistically significant	+24
Woodcock Reading Mastery Test–Revised[10]	Grade 1	79 students	39.81 (na)	39.49 (na)	0.32 (0.70)	0.49	Statistically significant	+19
Schwartz, 2005 (randomized controlled trial)[11]								
Observation Survey: Dictation subtest[9]	Grade 1	74 students	35.58 (2.70)	29.08 (7.37)	6.50	1.16	Statistically significant	+38
Observation Survey: Writing Vocabulary subtest[9]	Grade 1	74 students	42.67 (11.42)	31.00 (12.94)	11.67	0.95	Statistically significant	+33
Iverson and Tunmer, 1993 (quasi-experimental design)—standard *RR* group vs. comparison group[11]								
Observation Survey: Dictation subtest[9]	Grade 1	64 students 14 schools	34.63 (2.53)	23.34 (8.68)	11.29	1.74	Statistically significant	+46
Observation Survey: Writing Vocabulary subtest[9]	Grade 1	64 students 14 schools	39.66 (6.42)	15.69 (9.15)	23.97	3.00	Statistically significant	+50
Average for general reading achievement (Baenen et al., 1997)[11]						**–0.12**	**ns**	**–5**
Average for general reading achievement (Pinnell et al., 1988)[11]						**0.74**	**Statistically significant**	**+27**

Appendix A3.5 Summary of study findings included in the rating for the general reading achievement domain *(continued)*

Outcome measure	Study sample	Sample size	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Average for general reading achievement (Pinnell et al., 1994)[11]						0.55	Statistically significant	+21
Average for general reading achievement (Schwartz, 2005)[11]						1.05	Statistically significant	+35
Average for general reading achievement (Iverson & Tunmer, 1993)[11]						2.37	Statistically significant	+49
Domain average for general reading achievement across all studies[11]						0.92	na	+32

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the general reading achievement domain. Follow-up findings from the same 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. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of retention, the sign of the effect size was reversed so that a positive effect size favors the intervention group.
4. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
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.
7. 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. For the formulas the WWC used to calculate statistical significance, see Technical Details of WWC-Conducted Computations. In the case of Baenen et al. (1997), no corrections for clustering or multiple comparisons were needed. In the case of Schwartz (2005), the author applied a correction for multiple comparisons and so the WWC did not apply a correction. In the case of Pinnell et al. (1988), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study. In the case of Iverson and Tunmer (1993), corrections for clustering and multiple comparisons were needed, so the significance levels may differ from those reported in the original study.
8. Means presented for this measure are the proportions for each group retained (retention rates). Effect size is computed as a Cox Index: logged-odds ratio transformation divided by 1.65. See Technical Details of WWC-Conducted Computations for the computation of effect sizes for binary outcomes. The sign of the effect size was reversed so that a positive effect size favors the intervention group.
9. The *RR* group mean equals the comparison group mean plus the mean difference. The computation of the mean difference took into account the pretest difference between the study groups.
10. The authors stated that the exact sample varied from outcome to outcome, although they did not provide specific details. The mean difference is regression adjusted and the standard deviation is a pooled within-treatment group estimate. Study authors used HLM analysis to compute results. Effect size was calculated by dividing treatment effect estimate (regression-adjusted difference in the posttest measures) by the pooled within-treatment group standard deviation.
11. 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 findings for alternative intervention groups for the alphabetics domain by construct[1]

Outcome measure	Study sample	Sample size	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Phonemic awareness construct								
Iverson and Tunmer, 1993 (quasi-experimental design)—modified *RR* group vs. comparison group[7]								
Phoneme deletion task[8]	Grade 1	64 students 23 schools	11.00 (5.85)	3.91 (5.31)	7.09	1.25	Statistically significant	+40
Yopp-Singer Phoneme Segmentation Test[8]	Grade 1	64 students 23 schools	16.88 (4.53)	6.21 (5.18)	10.67	2.17	Statistically significant	+48
Print awareness construct								
Iverson and Tunmer, 1993 (quasi-experimental design)—modified *RR* group vs. comparison group[7]								
Observation Survey: Concepts About Print subtest[9]	Grade 1	64 students 23 schools	17.40 (2.40)	13.72 (3.49)	3.68	1.21	Statistically significant	+39
Pinnell, DeFord, and Lyons, 1988 (quasi-experimental design)—*RR* group with *RR* teachers vs. comparison group[7]								
Observation Survey: Concepts About Print subtest[9]	Grade 1	147 students[10] 14 schools	15.85 (2.77)	13.98 (3.31)	1.87	0.63	Statistically significant	+23
Letter knowledge construct								
Iverson and Tunmer, 1993 (quasi-experimental design)—modified *RR* group vs. comparison group[7]								
Observation Survey: Letter Identification subtest[9]	Grade 1	64 students 23 schools	52.68 (1.27)	49.00 (6.63)	3.68	0.75	Statistically significant	+27
Pinnell, DeFord, and Lyons, 1988 (quasi-experimental design)—*RR* group with *RR* teachers vs. comparison group[7]								
Observation Survey: Letter Identification subtest[9]	Grade 1	147 students[10] 14 schools	49.80 (3.83)	49.61 (8.33)	0.19	0.03	ns	+1
Phonics construct								
Iverson and Tunmer, 1993 (quasi-experimental design)—modified *RR* group vs. comparison group[7]								
Dolch Word Recognition Test[9]	Grade 1	64 students 23 schools	94.75 (23.74)	20.22 (13.93)	74.53	3.78	Statistically significant	+50
Observation Survey: Word Recognition subtest[9]	Grade 1	64 students 23 schools	11.91 (2.15)	5.56 (3.83)	6.35	2.02	Statistically significant	+48
Pseudoword decoding task[8]	Grade 1	64 students 23 schools	8.19 (6.69)	1.41 (2.31)	6.78	1.34	Statistically significant	+41

Appendix A4.1 Summary of findings for alternative intervention groups for the alphabetics domain by construct[1] *(continued)*

			Authors' findings from the study					
			Mean outcome (standard deviation)[2]		WWC calculations			
Outcome measure	Study sample	Sample size	*Reading Recovery® (RR)* group	Comparison group	Mean difference[3] (*RR*-comparison)	Effect size[4]	Statistical significance[5] (at $\alpha = 0.05$)	Improvement index[6]
Pinnell, DeFord, and Lyons, 1988 (quasi-experimental design)—*RR* group with *RR* teachers vs. comparison group[7]								
Observation Survey: Word Recognition subtest[9]	Grade 1	147 students[10] 14 schools	13.64 (1.85)	11.98 (3.92)	1.66	0.60	Statistically significant	+23

ns = not statistically significant

1. This appendix presents findings for alternative intervention groups on measures that fall in the alphabetics domain. Only primary 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. 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 Technical Details of WWC-Conducted Computations.
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.
7. 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. For the formulas the WWC used to calculate statistical significance, see Technical Details of WWC-Conducted Computations. In the cases of Iverson and Tunmer (1993) and Pinnell, DeFord, and Lyons (1988), a correction for clustering was needed, so the significance levels may differ from those reported in the original studies.
8. Means presented for this measure are the posttest means for each group (pretest data were not available in the study report).
9. The *RR* group mean equals the comparison group mean plus the mean difference. The computation of the mean difference took into account the pretest difference between the study groups.
10. Pretest data were available for only 94 of the 96 students in the *RR* group. It is reasonable to assume that the pretest mean for the 96 students is similar to that based on the 94 students.

Appendix A4.2 Summary of findings for alternative intervention groups for the comprehension domain by construct[1]

Outcome measure	Study sample	Sample size	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Reading comprehension construct								
Pinnell, DeFord, and Lyons, 1988 (quasi-experimental design)—*RR* group with *RR* teachers vs. comparison group[7]								
CTBS: Reading Comprehension subtest[8]	Grade 1	147 students[9] 14 schools	38.84 (15.31)	27.33 (13.94)	11.51	0.77	Statistically significant	+28
Vocabulary development construct								
Pinnell, DeFord, and Lyons, 1988 (quasi-experimental design)—*RR* group with *RR* teachers vs. comparison group[7]								
CTBS: Reading Vocabulary subtest[8]	Grade 1	147 students[9] 14 schools	38.16 (15.31)	28.07 (17.00)	10.09	0.63	Statistically significant	+24

1. This appendix presents findings for alternative intervention groups on measures that fall in the comprehension domain. Only primary 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. 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 Technical Details of WWC-Conducted Computations.
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.
7. 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. For the formulas the WWC used to calculate statistical significance, see Technical Details of WWC-Conducted Computations. In the case of Pinnell, DeFord, and Lyons (1988), a correction for clustering was needed, so the significance levels may differ from those reported in the original study. The clustering correction underadjusts clustering to the extent that the *RR* group mean and the comparison group mean were correlated within schools.
8. Means presented for these measures are the posttest means for each group (pretest data were not available in the study report).
9. Pretest data were available for only 94 of the 96 students in the *RR* group. It is reasonable to assume that the pretest mean for the 96 students is similar to that based on the 94 students.

Appendix A4.3 Summary of findings for alternative intervention groups for the general reading achievement domain[1]

Outcome measure	Study sample	Sample size	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Iverson and Tunmer, 1993 (quasi-experimental design)—modified *RR* group vs. comparison group[7]								
Observation Survey: Dictation subtest[8]	Grade 1	64 students 23 schools	34.75 (3.32)	23.34 (8.68)	11.23	1.67	Statistically significant	+45
Observation Survey: Writing Vocabulary subtest[8]	Grade 1	64 students 23 schools	40.53 (10.24)	15.69 (9.15)	24.84	2.51	Statistically significant	+49
Pinnell, DeFord, and Lyons, 1988 (quasi-experimental design)—*RR* group with *RR* teachers vs. comparison group[7]								
Observation Survey: Dictation subtest[8]	Grade 1	147 students[9] 14 schools	32.67 (5.80)	23.80 (7.99)	8.87	1.33	Statistically significant	+41
Observation Survey: Writing Vocabulary subtest[8]	Grade 1	147 students[9] 14 schools	35.96 (12.56)	25.37 (14.33)	10.59	0.80	Statistically significant	+29

1. This appendix presents findings for alternative intervention groups on measures that fall in the general reading achievement domain. Only primary group scores were used for rating purposes and are presented in Appendix A3.5.
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. 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 Technical Details of WWC-Conducted Computations.
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.
7. 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. For the formulas the WWC used to calculate statistical significance, see Technical Details of WWC-Conducted Computations. In the cases of Iverson and Tunmer (1993) and Pinnell, DeFord, and Lyons (1988), a correction for clustering was needed, so the significance levels may differ from those reported in the original studies.
8. The *RR* group mean equals the comparison group mean plus the mean difference. The computation of the mean difference took into account the pretest difference between the study groups.
9. Pretest data were available for only 94 of the 96 students in the *RR* group. It is reasonable to assume that the pretest mean for the 96 students is similar to that based on the 94 students.

Appendix A4.4 Summary of follow-up findings for the general reading achievement domain[1]

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study: Mean outcome (standard deviation)[2]: *Reading Recovery® (RR)* group	Authors' findings from the study: Mean outcome (standard deviation)[2]: Comparison group	WWC calculations: Mean difference[3] (*RR*-comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Baenen, Bernhole, Dulaney, and Banks, 1997 (randomized controlled trial)[7]								
Outcome time 2—end of second grade								
Retention[8]	Grade 1	147	0.04	0.04	0	0	ns	0
Outcome time 3—end of third grade								
North Carolina End-of-Grade reading test[9]	Grade 1	127	135.50 (8.40)	136.20 (9.10)	–0.70	–0.08	ns	–3

ns = not statistically significant

1. This appendix presents follow-up findings for measures that fall in general reading achievement. Only first-grade scores were used for rating purposes and are presented in Appendix A3.5.
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. 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 Technical Details of WWC-Conducted Computations.
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.
7. 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 Baenen et al. (1997), no correction for clustering was needed.
8. Means presented for this measure are the posttest data and proportions (retention rates). Pretest data are not applicable for the measure. Effect size is computed as a logged-odds ratio transformation divided by 1.65. See Technical Details of WWC-Conducted Computations, specifically the computation of effect sizes for binary outcomes.
9. Means presented for this measure are the posttest means for each group (pretest data were not available in the study report).

Appendix A5.1 *Reading Recovery®* 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 *Reading Recovery®* as positive. The remaining ratings (potentially positive effects, mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered, as *Reading Recovery®* 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. *Reading Recovery®* had three studies showing statistically significant positive effects, two of which met WWC evidence standards for a strong design.
AND
• Criterion 2: No studies showing statistically significant or substantively important *negative* effects.
Met. No study 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 Intervention Rating Scheme.

Appendix A5.2 *Reading Recovery®* rating for the fluency 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 fluency, the WWC rated *Reading Recovery®* as potentially positive. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered, as *Reading Recovery®* 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. *Reading Recovery®* had one study showing statistically significant 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 *Reading Recovery®* studies showed statistically significant or substantively important negative 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. *Reading Recovery®* had only one study showing statistically significant positive effects that met WWC evidence standards for a strong design.

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

  Met. No *Reading Recovery®* 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 Intervention Rating Scheme.

Appendix A5.3 *Reading Recovery®* 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 *Reading Recovery®* as potentially positive. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered, as *Reading Recovery®* 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. *Reading Recovery®* had one study showing statistically significant positive effects and one study showing indeterminate 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 *Reading Recovery®* studies showed statistically significant or substantively important negative effects, and the number of studies showing indeterminate effects was equal to the number of studies showing statistically significant 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. *Reading Recovery®* had only one study showing statistically significant positive effects that met WWC evidence standards for a strong design.

AND

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

  Met. No *Reading Recovery®* 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 Intervention Rating Scheme.

Appendix A5.4 *Reading Recovery®* 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 *Reading Recovery®* as positive. The remaining ratings (potentially positive effects, mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered, as *Reading Recovery®* 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. *Reading Recovery®* had four studies showing statistically significant positive effects, three of which met WWC evidence standards for a strong design.
AND
• Criterion 2: No studies showing statistically significant or substantively important *negative* effects.
Met. No *Reading Recovery®* 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 Intervention Rating Scheme.

Appendix A6 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Alphabetics	3	16	226	Medium to large
Fluency	1	1	74	Small
Comprehension	2	2	155	Small
General reading achievement	5	18	452	Medium to large

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