What Works Clearinghouse

Kaplan SpellRead

Program description[1]

Kaplan SpellRead (formerly known as *SpellRead Phonological Auditory Training*®) is a literacy program for struggling readers in grades 2 or above, including special education students, English language learners, and students more than two years below grade level in reading. *Kaplan SpellRead* integrates the auditory and visual aspects of the reading process and emphasizes specific skill mastery through systematic and explicit instruction. The program takes five to nine months to complete and consists of 140 lessons divided into three phases.

Research

Two studies of *Kaplan SpellRead* met the What Works Clearinghouse (WWC) evidence standards. The two studies included 208 students from first to third grades in Pennsylvania and in Newfoundland, Canada.[2] The WWC considers the extent of evidence for *Kaplan SpellRead* to be small for alphabetics, fluency, and comprehension. No studies that met WWC evidence standards with or without reservations addressed general reading achievement.

Effectiveness

Kaplan SpellRead was found to have positive effects on alphabetics and potentially positive effects on fluency and comprehension.

	Alphabetics	*Fluency*	*Comprehension*	*General reading achievement*
Rating of effectiveness	Positive	Potentially positive	Potentially positive	na
Improvement index[3]	Average: +18 percentile points	Average: +9 percentile points	Average: +20 percentile points	na
	Range: +2 to +44 percentile points	Range: +1 to +20 percentile points	Range: +1 to +37 percentile points	na

na = not applicable

1. The descriptive information for this program was obtained from publicly avialable sources: the program's web site (www.spellread.com, downloaded April, 2007) and the research literature (Torgesen et al., 2006). The WWC requests developers to review the program description sections for accuracy from their perspective. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review.
2. The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.
3. These numbers show the average and range of student-level improvement indices for all findings across the studies.

Additional program information[1]

Developer and contact

Kaplan SpellRead is distributed through Kaplan, Inc. and its Kaplan K12 Learning Services Division. Address: 1 Liberty Plaza, 22nd Floor, New York, NY 10006. Email: info@KaplanK12.com. Web: http://kaplank12.com/. Telephone: (888) 527-5268.

Scope of use

The program is currently being used in schools in Florida, Maryland, Virginia, and Pennsylvania. According to the current distributor, most of the students receiving the program are two or more years below grade level in reading, are receiving special education services, or are English language learners. The number of schools and students using the program is not available. In 2006 Kaplan K12 acquired SpellRead, the developer and distributor of *SpellRead Phonological Auditory Training*®.

Teaching

Kaplan SpellRead consists of 140 lessons implemented in three distinct phases that interweave phonemics, phonetics, and instruction in language-based reading and writing. Phase A (50 lessons) is designed to train the auditory process function of the brain to hear and manipulate the 44 sounds of the English language. Phase B (30 lessons) focuses on secondary spelling of vowel sounds, consonant blends, and syllabication of two-syllable words. Phase C (25 lessons) focuses on how to decode words of three or more syllables, as well as clusters and verb forms. The *Kaplan SpellRead* program is used with small groups of five students and one instructor in 60–90 minute classes. Each lesson includes activities to develop phonemic, phonetic, semantic, syntactic, comprehension/vocabulary, and fluency skills.

Kaplan SpellRead includes comprehensive professional development and ongoing expert support to educators as they implement the program. Kaplan K12 staff provides five days of initial workshops, two follow-up workshops, and regular on-site coaching visits. A web-based Instructor Support System allows educators to closely monitor student progress.

Cost

The cost of implementation in a school or a school district varies based on the number of participating students and their grade level (elementary, middle, or high) and on the number of teachers or schools participating in the program. Cost information is available from the distributor.

Research

Two studies reviewed by the WWC investigated the effects of *Kaplan SpellRead*. Both studies (Torgesen et al., 2006; Rashotte, MacPhee, & Torgesen, 2001) were randomized controlled trials that met WWC evidence standards.

Met evidence standards

Torgesen et al. (2006) examined the effects of *Kaplan SpellRead* on 203 third-grade students in eight school units[4] in Pennsylvania. Students in the comparison group participated in the regular reading program at their schools.

Rashotte, MacPhee, & Torgesen (2001) randomly assigned 47 first-grade and second-grade students from one school in Newfoundland, Canada, to the intervention and comparison groups. Students in the intervention group received the *Kaplan SpellRead* program. Students in the comparison group received the regular literacy-based reading program at their school.

Extent of evidence

The WWC categorizes the extent of evidence in each domain as small or moderate to large (see the What Works Clearinghouse Extent of Evidence Categorization Scheme). The extent of

4. A school unit consists of several partnered schools so that the cluster included two third-grade and two fifth-grade instructional groups. Because of the age range of the Beginning Reading review, only the data on the third-grade students were included in this review.

Research *(continued)*

evidence takes into account the number of studies and the total sample size across the studies that met WWC evidence standards with or without reservations.[5] The WWC considers the extent of evidence for *Kaplan SpellRead* to be small for alphabetics, fluency, and comprehension. No studies that met WWC evidence standards with or without reservations addressed general reading achievement.

Effectiveness

Findings

The WWC review of interventions for beginning reading addresses student outcomes in four domains: alphabetics, fluency, comprehension, and general reading achievement.[6] The studies included in this report cover three domains: alphabetics, fluency, and comprehension.

Alphabetics. Torgesen et al. (2006) examined four outcomes in the phonics construct of the alphabetics domain—the Woodcock Reading Mastery Test–Revised (WRMT–R) word identification and word attack subtests and the Test of Word Reading Efficiency (TOWRE) phonemic decoding efficiency and sight word efficiency subtests. The authors reported statistically significant effects of *Kaplan SpellRead* on two of these outcomes (the WRMT–R word attack subtest and the TOWRE decoding efficiency subtest). The statistical significance of these findings was consistent with the WWC calculation. The average effect size across the four outcomes was large enough to be considered substantively important according to WWC criteria (that is, an effect size of at least 0.25).

Rashotte, MacPhee, & Torgesen (2001) examined seven outcomes in the alphabetics domain—WRMT–R word identification and word attack subtests; the TOWRE phonetic decoding efficiency and sight word efficiency subtests; and the Comprehensive Test of Phonological Processing (CTOPP) elision, blending words, and segmenting words subtests. The authors reported statistically significant positive effects on six of the outcomes. However, the WWC analysis confirmed statistically significant differences for only four of the outcomes (WRMT–R word attack subtest, the TOWRE phonetic decoding efficiency subtest, and the CTOPP blending words and segmenting words subtests). The average effect size across all seven outcomes was statistically significant and positive.

Fluency. Torgesen et al. (2006) examined one outcome in this domain (the Oral Reading Fluency test) and reported no statistically significant effect. The effect size was not large enough to be considered substantively important.

Rashotte, MacPhee, & Torgesen (2001) examined two outcomes in the fluency domain and reported statistically significant positive effects for the outcomes (the Gray Oral Reading Tests (GORT–3) accuracy and rate subtests). However, none of those effects were statistically significant according to WWC analysis. The average effect size across the two outcomes was large enough to be considered substantively important.

Comprehension. Torgesen et al. (2006) examined two outcomes in this domain—the WRMT–R passage comprehension subtest and the GRADE passage comprehension subtest—and reported no statistically significant effects. The average effect size across the two outcomes was neither statistically significant nor large enough to be considered substantively important.

Rashotte, MacPhee, & Torgesen (2001) examined two outcomes in the comprehension domain—the Woodcock Diagnostic Reading Battery (WDRB) passage comprehension subtest and the GORT-3 comprehension subtest—and reported statistically significant effects for both outcomes. The statistical significance of these findings was consistent with the WWC calculation. The average effect size across the two outcomes was also statistically significant and positive.

5. The Extent of Evidence Categorization was developed to tell readers how much evidence was used to determine the intervention rating, focusing on the number and size of studies. Additional factors associated with a related concept, external validity, such as the students' demographics and the types of settings in which studies took place, are not taken into account for the categorization.
6. For definitions of the domains, see the Beginning Reading Protocol.

Effectiveness *(continued)*

Rating of effectiveness

The WWC rates the effects of an intervention in a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative. The rating of effectiveness takes into account four factors: the quality of the research design, the statistical significance of the findings,[7] the size of the difference between participants in the intervention and the comparison conditions, and the consistency in findings across studies (see the WWC Intervention Rating Scheme).

The WWC found *Kaplan SpellRead* to have positive effects on alphabetics and potentially positive effects on fluency and comprehension

Improvement index

The WWC computes an improvement index for each individual finding. In addition, within each outcome domain, the WWC computes an average improvement index for each study and an average improvement index across studies (see Technical Details of WWC-Conducted Computations). The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. Unlike the rating of effectiveness, the improvement index is based entirely on the size of the effect, regardless of the statistical significance of the effect, the study design, or the analyses. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.

The average improvement index for alphabetics is +18 percentile points across the two studies, with a range of +2 to +44 percentile points across findings. The average improvement index for fluency is +9 percentile points across the two studies, with a range of +1 to +20 percentile points across findings. The average improvement index for comprehension is +20 percentile points across the two studies, with a range of +1 to +37 percentile points across findings.

Summary

The WWC reviewed two studies on *Kaplan SpellRead*. Both studies met WWC evidence standards. Based on these two studies, the WWC found positive effects in alphabetics and potentially positive effects in fluency and comprehension. The evidence presented in this report may change as new research emerges.

References

Met WWC evidence standards

Rashotte, C. A., MacPhee, K., & Torgesen, J. K. (2001). The effectiveness of a group reading instruction program with poor readers in multiple grades. *Learning Disability Quarterly, 24*(2), 119–134.

Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., et al. (2006). *National assessment of Title I interim report—Volume II: Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers.* Retrieved from Institute of Education Sciences, U.S. Department of Education Web site: http://www.ed.gov/rschstat/eval/disadv/title1interimreport/index.html

For more information about specific studies and WWC calculations, please see the WWC *Kaplan SpellRead* Technical Appendices.

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, see the WWC Tutorial on Mismatch. See the Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate the statistical significance. In the case of *Kaplan SpellRead,* a correction for multiple comparisons was needed.

Appendix

Appendix A1.1 Study Characteristics: Rashotte, MacPhee, & Torgesen, 2001 (randomized controlled trial)

Characteristic	Description
Study citation	Rashotte, C. A., MacPhee, K., & Torgesen, J. K. (2001). The effectiveness of a group reading instruction program with poor readers in multiple grades. *Learning Disability Quarterly, 24*(2), 119–134.
Participants	The study included 116 students from grades 1–6 with below-average phonetic decoding and word-level reading skills (as measured by the word attack and word identification subtests of the Woodcock Reading Mastery Test–Revised (WRMT–R)). This WWC report focuses on 47 first-grade and second-grade students.[1] Students were matched on phonemic decoding and word-level skills at each grade level with one of each pair randomly assigned to *Kaplan SpellRead* and the other assigned to the comparison condition. Most of the students in the sample were from low-income families and all were Caucasian.
Setting	One elementary school in Newfoundland, Canada.
Intervention	*Kaplan SpellRead* was implemented in small groups of three to five students during language arts time outside the regular classroom. The students received 31–35 hours of the program over eight weeks. Each lesson consisted of 30 minutes of phonemic activities, 15 minutes of share reading, and five to six minutes of free reading. The phonemic activities used unscripted lessons with sound cards. New phonemic and phonetic skills were practiced during share reading, followed by free writing where students wrote down what was read.
Comparison	The comparison group children participated in the school's regular literacy-based reading program. The regular classroom teachers did not have training in phonetics. After the first posttest assessment, the comparison group was given the *Kaplan SpellRead* program while the intervention group was given no further *Kaplan SpellRead* instruction.
Primary outcomes and measurement	The primary outcomes in the alphabetics domain were the word identification and word attack subtests of the WRMT–R, the phonemic decoding efficiency subtest of the Test of Word Reading Efficiency (TOWRE), and elision, blending words, and segmenting words subtests of the Comprehensive Test of Phonological Processing (CTOPP). The primary outcomes in the fluency domain were the sight words efficiency subtest of the TOWRE and the Gray Oral Reading Test (GORT-3) word accuracy subtest. The main outcomes in the comprehension domain were the passage comprehension subtest of the Woodcock Diagnostic Reading Battery (WDRB) and the comprehension subtest of the GORT-3. (See Appendices A2.1–2.3 for more detailed descriptions of outcome measures.)
Teacher training	Three teachers and one supervisor implemented the *Kaplan SpellRead* program. The supervisor had previously taught the program for two years and one of the three teachers was certified. All instructors had been screened to insure that they had strong phonological skills. The four instructors participated in an intensive six-day training program provided by experienced *SpellRead* staff.

1. The study conducted statistical analyses of three groups of students: grades 1 and 2, grades 3 and 4, and grade 5 and 6. Results for third-grade students were not reviewed because they were not disaggregated from the results of fourth-grade students in this study, and the WWC Beginning Reading topic focuses only on the impact of interventions on students in grades K–3, as defined in the Beginning Reading protocol.

Appendix A1.2 Study Characteristics: Torgesen et al. 2006 (randomized controlled trial)

Characteristic	Description
Study citation	Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., et al. (2006). *National assessment of Title I interim report—Volume II: Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers.* Retrieved from Institute of Education Sciences, U.S. Department of Education web site: http://www.ed.gov/rschstat/eval/disadv/title1interimreport/index.html
Participants	The study design was based on random assignment of 37 school units[1] to one of four interventions: *Corrective Reading, Kaplan SpellRead, Failure Free Reading,* or *Wilson Reading*. Within each school, students were randomly assigned to the comparison condition or to the intervention randomly assigned to their school. This report focuses on eight school units assigned to *Kaplan SpellRead*.[2] At the time of analysis, the study included 92 third-grade[3] students (56 in the intervention and 36 in the comparison groups). The number of students at baseline was not reported.[4] Students were eligible to participate in the study if they were identified as struggling readers by their teachers and if they scored at or below the 30th percentile on a word-level reading test and at or above the 5th percentile on a vocabulary test. Thirty-five percent of students in the intervention groups were African-American and 32% in the comparison groups. The other students were Caucasian. Forty-six percent of students in the intervention groups and 36% in the comparison groups were eligible for free/reduced lunch.
Setting	Eight school units in Pennsylvania.
Intervention	The intervention was implemented from the first week of November 2003 through the first weeks in May 2004. During this time students received, on average, about 90 hours of instruction, which was delivered in 50-minute sessions five days a week to groups of three students. The three-student groups were heterogeneous with regard to students' basic reading skills. The average skills of each group determined the pace of learning. Many of the sessions took place during the student's regular classroom reading instruction, but outside their regular classrooms. Therefore, intervention group students received less reading instruction in the classroom than did students in the comparison group. Implementation fidelity was examined by trainers who observed the teachers and coached them over a period of months and by project coordinators who observed a sample of instructional sessions. In addition, ratings of a sample of videotaped sessions were used. Implementation was rated as acceptable.
Comparison	The comparison group students received their typical reading instruction, which included the regular classroom curriculum and, in many cases, other services (such as another pull-out program). The comparison group students had fewer small-group instructional hours than the intervention group students, but more one-on-one instructional hours.
Primary outcomes and measurement	The primary outcome measures in the alphabetics domain were the word identification and word attack subtests of the WRMT–R and the phonemic decoding efficiency and the sight words efficiency subtests of the TOWRE. The primary measure in the fluency domain was the Oral Reading Fluency test. The primary measures in the comprehension domain were the WRMT-R passage comprehension subtest and the Group Reading Assessment and Diagnostic Evaluation (GRADE) passage comprehension subtest. (See Appendices A2.1–2.3 for more detailed descriptions of outcome measures.)
Teacher training	Professional development included training and coaching by reading program staff, independent study of program materials, and telephone conferences. On average, intervention group teachers participated in 63.5 professional development hours across all phases of the study (initial training phase, practice phase, and implementation phase).

1. A school unit consists of several partnered schools so that the cluster included two third-grade and two fifth-grade instructional groups.
2. Findings on *Corrective Reading, Failure Free Reading,* and *Wilson Reading* are included in other WWC Beginning Reading reports.
3. The study also included analysis of impact on fifth-grade students. However, this WWC intervention report focuses on impact of beginning reading interventions for students in grades K-3. For further details please, see the Beginning Reading Protocol.
4. The study reported that six students in the intervention group and two students in the comparison group were lost to analysis. However, it is not clear whether those students were in third grade or were part of an additional sample of fifth-grade students that was also examined in this study.

Appendix A2.1 Outcome measures in the alphabetics domain

Outcome measure	Description
Phonological awareness	
Comprehensive Test of Phonological Processing (CTOPP): Blending Words subtest	A norm-referenced assessment that provides an overall measure of the child's phonological awareness skills. The blending words subtest includes 20 items that measure the extent to which the child can combine sounds to form words (as cited in Rashotte, MacPhee, & Torgesen, 2001).
CTOPP: Elision subtest	A norm-referenced assessment that provides an overall measure of the child's phonological awareness skills. The elision subtest includes 20 items that measure the extent to which the child can say a word and then say what is left after dropping out designated sounds (as cited in Rashotte, MacPhee, & Torgesen, 2001).
CTOPP: Segmenting Words subtest	A norm-referenced assessment that provides an overall measure of the child's phonological awareness skills. The 20-item segmenting words subtest was administered only in grade 2 and has the student repeat words and then say them one sound at a time (as cited in Rashotte, MacPhee, & Torgesen, 2001).
Phonics	
Test of Word Reading Efficiency (TOWRE): Phonetic Decoding Efficiency subtest	The TOWRE is a standardized, nationally normed measure. The phonetic decoding efficiency subtest measures the number of pronounceable printed nonwords that can be accurately decoded within 45 seconds (as cited in Torgesen et al., 2006, and Rashotte, MacPhee, & Torgesen, 2001).
TOWRE: Sight Word Efficiency subtest	The TOWRE is a standardized, nationally normed measure. The sight word efficiency subtest assesses the number of real printed words that can be accurately identified within 45 seconds (as cited in Torgesen et al., 2006, and Rashotte, MacPhee, & Torgesen, 2001).
Woodcock Reading Mastery Test-Revised (WRMT-R): Word Identification subtest	The word identification subtest is a test of decoding skills. The standardized test requires the child to read aloud isolated real words that range in frequency and difficulty (as cited in Torgesen et al., 2006, and Rashotte, MacPhee, & Torgesen, 2001).
WRMT-R: Word Attack subtest	This standardized test measures phonemic decoding skills by asking students to read pseudowords. Students are aware that the words are not real (as cited in Torgesen et al., 2006, and Rashotte, MacPhee, & Torgesen, 2001).

Appendix A2.2 Outcome measures in the fluency domain

Outcome measure	Description
Edformation Oral Fluency Assessment	This test measures the number of words correct per minute (WCPM) that students read using three brief grade-level passages (AIMSweb, as cited in Torgesen et al., 2006). These passages include both fiction and nonfiction text. The norms for this test are updated by Edformation each school year.
The Gray Oral Reading Test (GORT-3): Word Accuracy subtest	The word accuracy subtest of the GORT-3 is a standardized reading test that measures the number of word reading errors that occurred while reading a series of short paragraphs that increased in difficulty (as cited in Rashotte, MacPhee, & Torgesen, 2001).
GORT-3: Text Reading Rate subtest	The text reading rate subtest of the GORT-3 is a standardized reading test that measures the amount of time taken to read short paragraphs that increase in difficulty (as cited in Rashotte, MacPhee, & Torgesen, 2001).

Appendix A2.3 Outcome measures in the comprehension domain

Outcome measure	Description
Reading comprehension	
Group Reading Assessment and Diagnostic Evaluation (GRADE): Passage Comprehension subtest	The GRADE is an untimed, norm-referenced standardized test. The passage comprehension subtest includes a passage of text and corresponding multiple-choice comprehension questions (as cited in Torgesen et al., 2006).
GORT-3: Comprehension subtest	In this standardized test, students read paragraphs and answer five comprehension questions for each paragraph. The questions are read to students by the tester (as cited in Rashotte, MacPhee, & Torgesen, 2001).
WRMT–R: Passage Comprehension subtest	In this standardized test, comprehension is measured by having students fill in missing words in a short paragraph (as cited in Torgesen et al., 2006; Rashotte, MacPhee, & Torgesen, 2001).
Woodcock Diagnostic Reading Battery (WDRB): Passage Comprehension subtest	The passage comprehension subtest of the WDRB asks students to read silently a series of paragraphs and complete the missing words in each paragraph (as cited in Rashotte, MacPhee, & Torgesen, 2001).

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

Outcome measure	Study sample	Sample size (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Kaplan SpellRead* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Kaplan SpellRead* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Torgesen et al., 2006 (randomized controlled trial)[7]								
TOWRE: Phonetic Decoding Efficiency subtest	Grade 3	8 school units/92	95.84 (15.00)	88.74 (15.00)	7.10	0.47	Statistically significant	+18
TOWRE: Sight Word Efficiency subtest	Grade 3	8 school units/92	92.16 (15.00)	91.46 (15.00)	0.70	0.05	ns	+2
WRMT–R: Word Identification subtest	Grade 3	8 school units/92	89.61 (15.00)	87.61 (15.00)	2.00	0.13	ns	+5
WRMT–R: Word Attack subtest	Grade 3	8 school units/92	100.41 15.00)	93.91 (15.00)	6.50	0.43	Statistically significant	+17
Average[8] for alphabetics (Torgesen et al., 2006)						0.27	ns	+11
Rashotte, MacPhee, & Torgesen, 2001 (randomized controlled trial)[7]								
CTOPP: Elision subtest	Grades 1-2	1/47	98.90 (11.90)	95.20 (11.70)	3.70	0.31	ns	+12
CTOPP: Blending Words subtest	Grades 1-2	1/47	102.80 (11.40)	95.00 (10.90)	7.80	0.69	Statistically significant	+25
CTOPP: Segmenting Words subtest	Grade 2	1/20	98.50 (3.20)	89.00 (7.40)	9.50	1.60	Statistically significant	+44
TOWRE: Phonetic Decoding Efficiency subtest	Grades 1-2	1/47	90.70 (10.60)	82.10 (10.10)	8.60	0.82	Statistically significant	+29
TOWRE: Sight Word Efficiency subtest	Grades 1-2	1/47	88.00 (13.40)	86.90 (16.90)	1.10	0.07	ns	+3
WRMT–R: Word Identification subtest	Grades 1-2	1/47	93.90 (13.90)	91.70 (15.60)	2.20	0.15	ns	+6
WRMT–R: Word Attack subtest	Grades 1-2	1/47	101.40 (12.60)	88.8 (10.10)	12.60	1.08	Statistically significant	+36
Average[8] for alphabetics (Rashotte, MacPhee, & Torgesen, 2001)						0.67	Statistically significant	+25
Domain average[8] for alphabetics across all studies						0.47	na	+18

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

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. The Torgesen et al. (2006) study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socioeconomic status. No differences were found between subgroups of students for outcomes in 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 both studies reported here, the intervention group mean equals the comparison group mean plus the mean difference.
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 versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
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. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Torgesen et al. (2006), no corrections for clustering were needed in the alphabetics domain because students were assigned to conditions. Corrections for multiple comparisons were needed because the study's reported corrections for multiple comparisons are based on grouping of outcomes, which is different than the grouping of domains for this review. In the case of Rashotte, MacPhee, & Torgesen (2001), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study.
8. 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 (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Kaplan SpellRead* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Kaplan SpellRead* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Torgesen et al., 2006 (randomized controlled trial)[7]								
Oral Reading Fluency	Grade 3	8 school units/92	65.02 (39.20)	64.02 (39.20)	1.00	0.03	ns	+1
Average[8] for fluency (Torgesen et al., 2006)						0.03	ns	+1
Rashotte, MacPhee, & Torgesen, 2001 (randomized controlled trial)[7]								
GORT-3: Accuracy subtest	Grade 2	1/20	94.50 (20.60)	87.50 (16.70)	7.00	0.36	ns	+14
GORT-3: Rate subtest	Grade 2	1/20	92.50 (10.90)	87.50 (7.20)	5.00	0.52	ns	+20
Average[8] for fluency (Rashotte, MacPhee, & Torgesen, 2001)						0.44	ns	+17
Domain average[8] for fluency across all studies						0.23	na	+9

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. The Torgesen et al. (2006) study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socioeconomic status. No differences were found between subgroups of students for the outcome in 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. In both studies reported here, the intervention group mean equals the comparison group mean plus the mean difference.
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 versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
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 of 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 Torgesen et al. (2006), no corrections for clustering were needed in the fluency domain. No corrections for multiple comparisons were needed because there is only one outcome in this domain. In the case of Rashotte, MacPhee, & Torgesen (2001), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study.
8. 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 (schools/ students)	Authors' findings from the study: Mean outcome (standard deviation[2]) — *Kaplan SpellRead* group	Authors' findings from the study: Mean outcome (standard deviation[2]) — Comparison group	WWC calculations: Mean difference[3] (*Kaplan SpellRead* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Torgesen et al., 2006 (randomized controlled trial)[7]								
GRADE: Passage Comprehension	Grade 3	8 school units/92	84.58 (15.00)	79.68 (15.00)	4.90	0.32	ns	+13
WRMT–R: Passage Comprehension	Grade 3	8 school units/92	92.54 (15.00)	92.34 (15.00)	0.20	0.01	ns	+1
Average[8] for comprehension (Torgesen et al., 2006)						0.17	ns	+7
Rashotte, MacPhee, & Torgesen, 2001 (randomized controlled trial)[7]								
GORT-3: Comprehension subtest	Grade 2	1/20	97.50 (13.80)	82.50 (12.10)	15.00	1.11	Statistically significant	+37
WDRB: Comprehension subtest	Grades 1-2	1/47	102.50 (15.70)	91.40 (16.70)	11.10	0.67	Statistically significant	+25
Average[8] for comprehension (Rashotte, MacPhee, & Torgesen, 2001)						0.89	Statistically significant	+31
Domain average[8] for comprehension across all studies						0.53	na	+20

ns = not statistically significant
na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. The Torgesen et al. (2006) study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socioeconomic status. No differences were found between subgroups of students for outcomes in 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. In both studies reported here, the intervention group mean equals the comparison group mean plus the mean difference.
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 versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
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. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Torgesen et al. (2006) , no corrections for clustering were needed in the comprehension domain. No corrections for multiple comparisons were needed because the study's reported corrections for multiple comparisons were based on the same grouping of outcomes as the domain for this review. In the case of Rashotte, MacPhee, & Torgesen (2001), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study.
8. 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 *Kaplan SpellRead* 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 *Kaplan SpellRead* as having positive effects. The remaining ratings (potentially positive effects, mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *Kaplan SpellRead* received 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 of *Kaplan SpellRead* showed statistically significant positive effects. Both studies met the WWC evidence standards for a strong design.
AND
• Criterion 2: No studies showing statistically significant or substantively important *negative* effects.
Met. No studies of *Kaplan SpellRead* showed statistically significant or substantively important negative effects.

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

Appendix A4.2 *Kaplan SpellRead* 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 *Kaplan SpellRead* as have potentially positive effects. It did not meet the criteria for positive effects because no studies showed a statistically significant positive effect. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, or negative effects) were not considered because *Kaplan SpellRead* was assigned the highest applicable rating.

Rating received

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

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

  Met. One study showed substantively important positive effects.

AND

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

  Met. No studies showed a statistically significant or substantively important negative effect, and one study showed an indeterminate effect.

Other ratings considered

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

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

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

AND

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

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

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

Appendix A4.3 *Kaplan SpellRead* 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 *Kaplan SpellRead* as having potentially positive effects. It did not meet the criteria for positive effects because only one study showed statistically significant positive effects. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, or negative effects) were not considered because *Kaplan SpellRead* was assigned the highest applicable rating.

Rating received

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

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

  Met. One study showed 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 studies showed a statistically significant or substantively important negative effect. One study showed an indeterminate effect.

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. One study showed a statistically significant positive effect and one study showed an indeterminate effect.

AND

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

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

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

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools[2]	Students	
Alphabetics	2	>9	139	Small
Fluency	2	>9	139	Small
Comprehension	2	>9	139	Small
General reading achievement	0	0	0	na

na = not applicable/not studied

1. A rating of "moderate to large" requires at least two studies and two schools across studies in one domain, and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is "small."
2. One of the two studies reviewed included students from eight schools units. A school unit consists of several partnered schools so that the cluster included two third-grade and two fifth-grade instructional groups.