

Appendix

Appendix A1 Study characteristics: Torgesen et al. (2006)

Characteristic	Description
Study citation	Torgesen, J., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., Mansfield, W., et al. (2006). <i>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</i> . Washington, DC: National Center for Education Evaluation and Regional Assistance. Additional source: Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., et al. (2007). <i>National assessment of Title I. Final report. Volume II: Closing the reading gap: Findings from a randomized trial of four reading interventions for striving readers</i> (NCEE 2008-4013). Washington, DC: National Center for Education Evaluation and Regional Assistance.
Participants	The study design was based on random assignment of 32 school units, ¹ formed from a pool of 52 schools, to one of four interventions (<i>Corrective Reading</i> , <i>Kaplan SpellRead</i> , <i>Failure Free Reading</i> , and <i>Wilson Reading</i>). Within each school, students were randomly assigned to the treatment group that would receive the intervention assigned to its school or to the control group that would receive the standard reading curriculum. This report focuses on schools assigned to <i>Corrective Reading</i> and on findings for 5th graders (as specified by the Adolescent Literacy review protocol). At the time of the analysis, the sample relevant to this review included 86 fifth-grade students (55 in <i>Corrective Reading</i> and 31 in the control group) in seven school units. The number of 5th-grade students at baseline was not reported. ² Students were eligible for participation if their teacher identified them as a struggling reader 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. On average, at baseline, students scored about one-half to one standard deviation below national norms on measures used to assess their ability to decode words. About 51% of the intervention group students were females, compared to 36% in the control group. About 41% of the intervention group students were eligible for free or reduced-price lunch programs, equal to 41% of the students in the control group.
Setting	The analysis sample included seven school units in the Allegheny Intermediate Unit (AIU), outside Pittsburgh, Pennsylvania. The AIU consisted of 42 school districts.
Intervention	The decoding component of <i>Corrective Reading</i> was implemented by nine teachers beginning in the first week in November 2003 through the first week in May 2004. The comprehension component was not implemented. The intervention was administered to students in groups of three that were heterogeneous with regard to students' basic reading skills. The average skills of the students in each of the instructional groups determined the pace of instruction. Implementation fidelity was determined by reading program trainers who observed the teachers and coached them over a period of months, project coordinators who observed a sample of instructional sessions, and ratings based on a sample of videotaped sessions. Implementation was rated as acceptable. The decoding component used in the study included four levels—A, B1, B2, and C. Placement testing was used to start each group at the appropriate level. The lessons provided during the study clustered in levels B1 and B2. For those groups that progressed to level C, explicit vocabulary instruction was not provided. Over a six-month period, students received a total of about 90 hours of instruction. Students received <i>Corrective Reading</i> instruction five days a week in sessions that were approximately 55 minutes long. The study reported student outcomes after six months of program implementation. Additional findings reflecting students' outcomes one year after the end of the implementation of the intervention can be found in Appendices A4.1–A4.3.
Comparison	The control group students received their regular reading instruction, which included typical classroom instruction and, in many cases, other services (such as another pull-out program). Across four interventions, the control group students had fewer small-group instructional hours and average weekly hours of total reading instruction than the intervention group students.

(continued)

Appendix A1 Study characteristics: Torgesen et al. (2006) (continued)

Characteristic	Description
Primary outcomes and measurement	The primary outcome measures in the alphabets domain were the Word Identification and Word Attack subtests of the Woodcock Reading Mastery Test–Revised (WRMT-R) and the Phonetic Decoding Efficiency and Sight Word Efficiency subtests of the Test of Word Reading Efficiency (TOWRE). The primary measure in the reading fluency domain was the Oral Reading Fluency test (also referred to as AIMSweb). The primary measures in the comprehension domain were the WRMT-R: Comprehension subtest and the Group Reading Assessment and Diagnostic Evaluation (GRADE): Passage Comprehension subtest. For a more detailed description of these outcome measures, see Appendices A2.1–A2.3.
Staff/teacher training	Professional development on how to use <i>Corrective Reading</i> included training and coaching by <i>Corrective Reading</i> program staff, teachers' independent study of program materials, and telephone conferences between teachers and <i>Corrective Reading</i> staff. On average, throughout the course of the study, the <i>Corrective Reading</i> intervention group teachers participated in 70.8 professional development hours specifically related to using <i>Corrective Reading</i> (32.8 hours were initial training in use of the program, 26.4 hours were spent in a practice phase, and 11.6 hours occurred during the six-month period in which teachers were using <i>Corrective Reading</i>).

1. A school unit consists of several schools partnering so that the cluster included two 3rd-grade and two 5th-grade instructional groups.
2. The study reported that 10 students in the intervention group and no students in the control group were lost to analysis. However, it is not clear if those students were in 5th grade or were part of the sample of 3rd-grade students that was also examined in this study. The 3rd-grade sample that was included in this study is not reviewed in this report because it is outside the scope of the review.

Appendix A2.1 Outcome measures for the alphabetic domain

Outcome measure	Description
<i>Phonics construct</i>	
Woodcock Reading Mastery Test–Revised (WRMT-R): Word Attack subtest	This standardized test measures phonemic decoding skills by asking students to pronounce printed pseudo-words. Students are aware that the words are not real (as cited in Torgesen et al., 2006).
Woodcock Reading Mastery Test–Revised (WRMT-R): Word Identification subtest	The word identification subtest is a test of decoding skills. The standardized test requires children to pronounce real words from a list of increasing difficulty (as cited in Torgesen et al., 2006).
Test of Word Reading Efficiency (TOWRE): Phonemic Decoding Efficiency (PDE) subtest	The TOWRE is a standardized, nationally normed measure. The PDE subtest measures the number of nonwords of increasing difficulty that students can pronounce within 45 seconds (as cited in Torgesen et al., 2006).
Test of Word Reading Efficiency (TOWRE): Sight Word Efficiency (SWE) subtest	The TOWRE is a standardized, nationally normed measure. The SWE subtest measures the number of real words of increasing difficulty that students can pronounce within 45 seconds (as cited in Torgesen et al., 2006).

Appendix A2.2 Outcome measures for the reading fluency domain

Outcome measure	Description
Oral Reading Fluency assessment	This test (also referred to as AIMSweb) measures the number of words correct per minute (WCPM) that students read using three brief grade-level-appropriate passages. These passages contain both fiction and nonfiction text. The norms for this test are updated by Edformation each school year (as cited in Torgesen et al., 2006).

Appendix A2.3 Outcome measures for the comprehension domain

Outcome measure	Description
<i>Reading comprehension construct</i>	
Woodcock Reading Mastery Test–Revised (WRMT-R): Passage Comprehension subtest	In this standardized test, comprehension is measured by having students read silently and fill in missing words in a short paragraph (as cited in Torgesen et al., 2006).
Group Reading Assessment and Diagnostic Evaluation (GRADE): Passage Comprehension subtest	The GRADE is a norm-referenced reading assessment that can be used with students at any level. The GRADE has four subtests: (1) Vocabulary, (2) Sentence Comprehension, (3) Passage Comprehension, and (4) Listening Comprehension. The Passage Comprehension subtest includes a passage of text and corresponding multiple-choice comprehension questions (as cited in Torgesen et al., 2006).

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain¹

Outcome measure	Study sample	Authors' findings from the study						
		Sample size (school units/ students)	Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>Corrective Reading</i> – comparison)	WWC calculations		
			<i>Corrective Reading</i> group	Comparison group		Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
Torgesen et al., 2006⁸								
WRMT-R: Word Attack subtest	Grade 5	7/86	97.40 (15.00)	95.50 (15.00)	1.90	0.13	ns	+5
WRMT-R: Word Identification subtest	Grade 5	7/86	92.90 (15.00)	92.60 (15.00)	0.30	0.02	ns	+1
TOWRE: Phonemic Decoding Efficiency (PDE) subtest	Grade 5	7/86	87.30 (15.00)	85.40 (15.00)	1.90	0.13	ns	+5
TOWRE: Sight Word Efficiency (SWE) subtest	Grade 5	7/86	88.70 (15.00)	86.50 (15.00)	2.20	0.15	ns	+6
Domain average for alphabetics (Torgesen et al., 2006)⁹						0.10	ns	+4

ns = not statistically significant

WRMT-R = Woodcock Reading Mastery Test–Revised

TOWRE = Test of Word Reading Efficiency

- This appendix reports findings considered for the effectiveness rating and the average improvement indices for the alphabetics domain. Follow-up findings from the same studies are not included in these ratings but are reported in Appendix A4.1. Torgesen et al. (2006) also included subgroup analyses by initial skill level (WRMT-R Word Attack subtest and Peabody Picture Vocabulary Test [PPVT]) and socioeconomic status. The study reported that *Corrective Reading* had statistically significant positive effects on TOWRE Sight Word Efficiency for students with low initial Word Attack scores and low initial PPVT scores, and for students who were eligible for free or reduced-price lunch programs. The study also found statistically significant positive effects on WRMT-R Word Attack scores for students with low initial PPVT scores. No other differences were found between subgroups of students for outcomes in the alphabetics domain.
- For Torgesen et al. (2006), the mean outcomes were computed using information reported in the paper. For the control group, the mean outcome is the control group baseline mean standard score plus the control group gain. For the intervention group, the mean outcome is the control group baseline mean standard score plus the control group gain plus the impact of the intervention.
- 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. The standard deviations in the Torgesen et al. (2006) study were the population standard deviations for these standardized outcomes.
- Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the control group.
- For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
- Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
- 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 control condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
- The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Torgesen et al. (2006) and the alphabetics domain, no corrections for clustering were needed because the authors adjusted for clustering, and no correction for multiple comparisons was needed because there were no statistically significant findings in this domain.
- This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A3.2 Summary of study findings included in the rating for the reading fluency domain¹

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study			WWC calculations		
			Corrective Reading group	Comparison group	Mean outcome ² (standard deviation) ³	Mean difference ⁴ (Corrective Reading – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)
Torgesen et al., 2006⁸								
Oral Reading Fluency test	Grade 5	7/86	96.80 (47.00)	91.90 (47.00)	4.90	0.10	ns	+4
Domain average for reading fluency (Torgesen et al., 2006)⁹						0.10	ns	+4

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the reading fluency domain. Follow-up findings from the same studies are not included in these ratings but are reported in Appendix A4.2. The 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 reading fluency outcome.
2. For Torgesen et al. (2006), the mean outcomes were computed using information reported in the paper. For the control group, the mean outcome is the control group baseline mean standard score plus the control group gain. For the intervention group, the mean outcome is the control group baseline mean standard score plus the control group gain plus the impact of the intervention.
3. 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. The standard deviations in the Torgesen et al. (2006) study were the population standard deviations for these standardized outcomes.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the control group.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the control condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
8. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Torgesen et al. (2006) and the reading fluency domain, no corrections for clustering were needed because the authors adjusted for clustering, and no correction for multiple comparisons was needed because there is only one outcome in this domain.
9. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A3.3 Summary of study findings included in the rating for the comprehension domain¹

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study		WWC calculations			
			Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>Corrective Reading</i> – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			<i>Corrective Reading</i> group	Comparison group				
Torgesen et al., 2006⁸								
WRMT-R: Passage Comprehension subtest	Grade 5	7/86	93.80 (15.00)	92.00 (15.00)	1.80	0.12	ns	+5
GRADE: Passage Comprehension subtest	Grade 5	7/86	96.30 (15.00)	96.00 (15.00)	0.30	0.02	ns	+1
Domain average for comprehension (Torgesen et al., 2006)⁹						0.07	ns	+3

ns = not statistically significant

WRMT-R = Woodcock Reading Mastery Test–Revised

GRADE = Group Reading Assessment and Diagnostic Evaluation

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the comprehension domain. Follow-up findings from the same studies are not included in these ratings but are reported in Appendix A4.3. The 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. For Torgesen et al. (2006), the mean outcomes were computed using information reported in the paper. For the control group, the mean outcome is the control group baseline mean standard score plus the control group gain. For the intervention group, the mean outcome is the control group baseline mean standard score plus the control group gain plus the impact of the intervention.
3. 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. The standard deviations in the Torgesen et al. (2006) study were the population standard deviations for these standardized outcomes.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the control group.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the control condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results for the intervention group.
8. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Torgesen et al. (2006) and the comprehension domain, no corrections for clustering were needed because the authors adjusted for clustering. No correction for multiple comparisons was needed because there were no statistically significant findings in this domain.
9. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A4.1 Summary of follow-up findings for the alphabetic domain¹

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study		WWC calculations			
			Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>Corrective Reading</i> – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			<i>Corrective Reading</i> group	Comparison group				
Torgesen, et al., 2007⁸								
WRMT-R: Word Attack subtest	Grade 5	7/84	98.60 (15.00)	98.40 (15.00)	0.20	0.01	ns	+1
WRMT-R: Word Identification subtest	Grade 5	7/84	92.10 (15.00)	94.00 (15.00)	-1.90	-0.13	ns	-5
TOWRE: Phonemic Decoding Efficiency (PDE) subtest	Grade 5	7/84	90.00 (15.00)	88.60 (15.00)	1.40	0.09	ns	+4
TOWRE: Sight Word Efficiency (SWE) subtest	Grade 5	7/84	87.10 (15.00)	87.50 (15.00)	-0.40	-0.03	ns	-1

ns = not statistically significant

WRMT-R = Woodcock Reading Mastery Test–Revised

TOWRE = Test of Word Reading Efficiency

1. This appendix presents findings from data collected one year after the end of the implementation of the intervention for measures that fall in the alphabetic domain. Data that reflected students' exposure to six months of the intervention were used for rating purposes and are presented in Appendix A3.1. Torgesen et al. (2007) also included subgroup analyses by initial skill level (WRMT-R Word Attack subtest and Peabody Picture Vocabulary Test [PPVT]) and socioeconomic status. The study reported that *Corrective Reading* had a statistically significant positive effect on the TOWRE PDE and TOWRE SWE for students eligible for free or reduced-price lunch programs. The study also reported that *Corrective Reading* had a statistically significant positive effect on the TOWRE SWE for students with low initial PPVT scores and on the TOWRE PDE for students with low initial Word Attack scores but high initial PPVT scores. Additional subgroup analyses found statistically significant positive effects on Word Attack for students with low initial Word Attack scores and for students with low scores on both baseline assessments (Word Attack and PPVT). No other differences were reported between subgroups of students for outcomes in the alphabetic domain.
2. For Torgesen et al. (2007), the mean outcomes were computed using information reported in the paper. For the control group, the mean outcome is the control group baseline mean standard score plus the control group gain. For the intervention group, the mean outcome is the control group baseline mean standard score plus the control group gain plus the impact of the intervention.
3. 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. The standard deviations in the Torgesen et al. (2007) study were the population standard deviations for these standardized outcomes.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the control group.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the control condition. The improvement index can take on values between -50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Torgesen et al. (2007) and the alphabetic domain, no corrections for clustering were needed because the authors adjusted for clustering, and no correction for multiple comparisons was needed because there were no statistically significant findings in this domain.

Appendix A4.2 Summary of follow-up findings for the reading fluency domain¹

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study		WWC calculations			
			Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>Corrective Reading</i> – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			<i>Corrective Reading</i> group	Comparison group				
Torgesen et al., 2007⁸								
Oral Reading Fluency test	Grade 5	7/84	102.10 (47.00)	107.40 (47.00)	–5.30	–0.11	ns	–4

ns = not statistically significant

1. This appendix presents findings from data collected one year after the end of the implementation of the intervention for measures that fall in the reading fluency domain. Data that reflected students' exposure to six months of the intervention were used for rating purposes and are presented in Appendix A3.2. Torgesen et al. (2007) also included subgroup analyses by initial skill level (WRMT-R Word Attack subtest and Peabody Picture Vocabulary Test [PPVT]) and socioeconomic status. The study reported that *Corrective Reading* had a statistically significant negative effect on the Oral Reading Fluency test for students with high initial PPVT scores, for students who had both high initial word attack scores and high initial PPVT scores, and for students who were not eligible for free or reduced-price lunch programs. No other differences were found between subgroups of students for the reading fluency outcome.
2. For Torgesen et al. (2007), the mean outcomes were computed using information reported in the paper. For the control group, the mean outcome is the control group baseline mean standard score plus the control group gain. For the intervention group, the mean outcome is the control group baseline mean standard score plus the control group gain plus the impact of the intervention.
3. 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. The standard deviations in the Torgesen et al. (2007) study were the population standard deviations for these standardized outcomes.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the control group.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the control condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Torgesen et al. (2007) and the reading fluency domain, no corrections for clustering were needed because the authors adjusted for clustering, and no correction for multiple comparisons was needed because there is only one outcome in this domain.

Appendix A4.3 Summary of follow-up findings for the comprehension domain¹

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study		WWC calculations			
			Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>Corrective Reading</i> – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			<i>Corrective Reading</i> group	Comparison group				
Torgesen et al., 2007⁸								
WRMT-R: Passage Comprehension	Grade 5	7/84	93.50 (15.00)	95.60 (15.00)	–2.10	–0.14	ns	–6
GRADE: Passage Comprehension	Grade 5	7/84	91.60 (15.00)	91.60 (15.00)	0	0	ns	0

ns = not statistically significant

WRMT-R = Woodcock Reading Mastery Test–Revised

GRADE = Group Reading Assessment and Diagnostic Evaluation

1. This appendix presents findings from data collected one year after the end of the implementation of the intervention for measures that fall in the comprehension domain. Data that reflected students' exposure to six months of the intervention were used for rating purposes and are presented in Appendix A3.3. Torgesen et al. (2007) also included subgroup analyses by initial skill level (WRMT-R Word Attack subtest and Peabody Picture Vocabulary Test [PPVT]) and socioeconomic status. The study reported that *Corrective Reading* had a statistically significant negative effect on both comprehension outcomes for students with high initial level of Word Attack scores. The difference was also negative and statistically significant on the GRADE outcome for students with high initial level of Word Attack and PPVT scores, but positive and statistically significant for students with low initial PPVT scores. The study also found a statistically significant positive effect on the WRMT Passage Comprehension for students with low skill level on both baseline assessments: Word Attack and PPVT. No other differences were reported between subgroups of students for outcomes in the comprehension domain.
2. For Torgesen et al. (2007), the mean outcomes were computed using information reported in the paper. For the control group, the mean outcome is the control group baseline mean standard score plus the control group gain. For the intervention group, the mean outcome is the control group baseline mean standard score plus the control group gain plus the impact of the intervention.
3. 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. The standard deviations in the Torgesen et al. (2007) study were the population standard deviations for these standardized outcomes.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the control group.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the control condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Torgesen et al. (2007) and the comprehension domain, no corrections for clustering were needed because the authors adjusted for clustering. No correction for multiple comparisons was needed because there were no statistically significant findings in this domain.

Appendix A5.1 *Corrective Reading* rating for the alphabetics domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.¹ For the outcome domain of alphabetics, the WWC rated *Corrective Reading* as having no discernible effects for adolescent learners.

Rating received

No discernible effects: No affirmative evidence of effects.

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

Met. No studies showed a statistically significant or substantively important effect, either positive or negative.

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 a statistically significant or substantively important negative effect.

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

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

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

AND

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

Not met. No studies showed a statistically significant or substantively important negative effect. One study showed indeterminate effects.

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

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

Not met. No studies showed a statistically significant or substantively important positive effect, and no studies showed a statistically significant or substantively important negative effect.

OR

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

Not met. No studies showed a statistically significant or substantively important effect. One study showed indeterminate effects.

(continued)

Appendix A5.1 *Corrective Reading* rating for the alphabetics domain (continued)

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

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

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

OR

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

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

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

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

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

AND

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

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

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

Appendix A5.2 Corrective Reading rating for the reading fluency domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.¹ For the outcome domain of reading fluency, the WWC rated *Corrective Reading* as having no discernible effects for adolescent learners.

Rating received

No discernible effects: No affirmative evidence of effects.

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

Met. No studies showed a statistically significant or substantively important effect, either positive or negative.

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 a statistically significant or substantively important negative effect.

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

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

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

AND

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

Not met. No studies showed a statistically significant or substantively important negative effect. One study showed indeterminate effects.

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

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

Not met. No studies showed a statistically significant or substantively important positive effect, and no studies showed a statistically significant or substantively important negative effect.

OR

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

Not met. No studies showed a statistically significant or substantively important effect. One study showed indeterminate effects.

(continued)

Appendix A5.2 Corrective Reading rating for the reading fluency domain (continued)

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

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

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

OR

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

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

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

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

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

AND

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

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

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

Appendix A5.3 *Corrective Reading* rating for the comprehension domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.¹ For the outcome domain of comprehension, the WWC rated *Corrective Reading* as having no discernible effects for adolescent learners.

Rating received

No discernible effects: No affirmative evidence of effects.

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

Met. No studies showed a statistically significant or substantively important effect, either positive or negative.

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 a statistically significant or substantively important negative effect.

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

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

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

AND

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

Not met. No studies showed a statistically significant or substantively important negative effect. One study showed indeterminate effects.

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

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

Not met. No studies showed a statistically significant or substantively important positive effect, and no studies showed a statistically significant or substantively important negative effect.

OR

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

Not met. No studies showed a statistically significant or substantively important effect. One study showed indeterminate effects.

(continued)

Appendix A5.3 *Corrective Reading* rating for the comprehension domain (continued)

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

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

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

OR

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

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

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

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

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

AND

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

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

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

Appendix A6 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence ¹
		Schools	Students	
Alphabetics	1	7 school units ²	86	Small
Reading fluency	1	7 school units ²	86	Small
Comprehension	1	7 school units ²	86	Small
General literacy achievement	na	na	na	na

na = not applicable/not studied

1. A rating of “medium to large” requires at least two studies and two schools across studies in one domain and a total sample size across studies of at least 350 students or 14 classrooms. Otherwise, the rating is “small.” For more details on the extent of evidence categorization, see the WWC Procedures and Standards Handbook, Appendix G.
2. A school unit consists of several schools and includes two 3rd-grade and two 5th-grade instructional groups. The exact number of schools participating in *Corrective Reading* is unknown.