

Appendix

Appendix A1 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). <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> . 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, <i>Corrective Reading</i> , <i>Kaplan SpellRead</i> , <i>Failure Free Reading</i> , or <i>Wilson Reading</i> . Within each school, students were randomly assigned to the intervention or to the comparison condition. This report focuses on eight school units assigned to <i>Corrective Reading</i> . ² At the time of analysis, the sample included 79 third-grade students (44 in intervention and 35 in comparison groups). The number of students at baseline was not reported. ³ About 44% of the intervention group students were female, compared with 52% in the comparison group. About 42% of the intervention group students were eligible for free/reduced lunch program and 48% in the comparison group.
Setting	Eight school units in Pennsylvania.
Intervention	<i>Corrective Reading</i> was implemented by nine teachers beginning in the first week in November 2003 through the first weeks in May 2004. The intervention was administered to groups of three students with heterogeneous basic reading skills. The average skills of each three-student instructional group 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, by project coordinators who observed a sample of instructional sessions, and by ratings based on a sample of videotaped sessions. Implementation was rated as acceptable. For purposes of this study, only the word-level skills components were implemented. However, the complete program of <i>Corrective Reading</i> also contains instructional routines and materials that focus on comprehension and vocabulary.
Comparison	The comparison group students received their regular reading instruction, which included typical classroom instruction 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 alphabets domain were the word identification and word attack subtests of the Woodcock Reading Mastery Test—Revised (WRMT—R) and the phonetic decoding efficiency subtest of the Test of Word Reading Efficiency (TOWRE). The primary measures in the fluency domain were the TOWRE sight word efficiency subtest and the Oral Reading Fluency test. The primary measures in the comprehension domain were the passage comprehension subtests of WRMT—R and the Group Reading Assessment and Diagnostic Evaluation (GRADE). (See Appendix 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 70.8 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 *Kaplan SpellRead*, *Failure Free Reading*, and *Wilson Reading* are included in other WWC beginning reading reports.
3. The study reported that 10 students in the intervention 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. The fifth-grade sample that was included in this study is not reviewed in this report because it is outside the scope of the review. For sample relevancy criteria please see the [Beginning Reading Protocol](#).

Appendix A2.1 Outcome measures in the alphabetic domain

Outcome measure	Description
<i>Phonics</i>	
Test of Word Reading Efficiency (TOWRE): Phonetic Decoding Efficiency subtest	The TOWRE is a standardized 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).
TOWRE: Sight Word Efficiency subtest	The TOWRE is a standardized 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).
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).
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).

Appendix A2.2 Outcome measure 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.

Appendix A2.3 Outcome measures in the comprehension domain

Outcome measure	Description
<i>Reading comprehension</i>	
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).
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).

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

Outcome measure	Study sample	Sample size (school units/ students)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation ²)		Mean difference ³ (Corrective Reading – comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
			Corrective Reading group	Comparison group				
Torgesen et al., 2006 (randomized controlled trial)⁷								
TOWRE: Phonetic Decoding Efficiency subtest	Grade 3	8/79	89.86 (15.00)	89.48 (15.00)	0.38	0.02	ns	+1
TOWRE: Sight Word Efficiency subtest	Grade 3	8/79	90.98 (15.00)	86.41 (15.00)	4.57	0.30	Statistically significant	+12
WRMT–R: Word Identification subtest	Grade 3	8/79	91.06 (15.00)	87.77 (15.00)	3.30	0.22	Statistically significant	+9
WRMT–R: Word Attack subtest	Grade 3	8/79	100.34 (15.00)	95.15 (15.00)	5.18	0.34	ns	+13
Domain average⁸ for alphabetics						0.22	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 study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socioeconomic status. The study found no differences for alphabetics outcomes as a function of students' initial level of word attack scores. The study reported that Corrective Reading had a statistically significant positive effect on alphabetics for students with high pretest PPVT scores and no effects on students with low pretest PPVT scores. An additional subgroup analysis found statistically significant positive effects on the TOWRE sight word efficiency subtest for students not eligible for the free/reduced lunch program, but not for students eligible for the free/reduced lunch program.
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. The standard deviations in Torgesen et al. (2006) were the population standard deviations for these standardized outcomes.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The intervention group mean is 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) and the alphabetics domain, corrections for multiple comparisons were needed because the study's reported corrections for multiple comparisons were based on a grouping of outcomes that differed from the groupings of domains for this review.
8. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

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

Outcome measure	Study sample	Sample size (school units/ students ³)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation ²)		Mean difference ⁴ (Corrective Reading – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			Corrective Reading group	Comparison group				
Torgesen et al., 2006 (randomized controlled trial)⁷								
Oral Reading Fluency	Grade 3	8/79	66.04 (39.20)	55.33 (39.20)	10.71	0.27	Statistically significant	+11
Domain average⁹ for fluency						0.27	Statistically significant	+11

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. The study also included subgroup analyses by initial skill level (WRMT–R word attack subtest and Peabody Picture Vocabulary Test (PPVT)) and socioeconomic status. The study found statistically significant positive effects on the Oral Fluency Test for students not eligible for the free/reduced lunch program, but not for students eligible for the free/reduced lunch program. No differences were found between initial skill level subgroups of students for the fluency outcome.
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. The standard deviations in Torgesen et al. (2006) were the population standard deviations for these standardized outcomes.
3. The sample size for the analysis was not reported in the study. The sample size reported is the total number of third-grade students in the intervention and control conditions at baseline, which may differ from the actual number of students used in the analysis in the report.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The intervention group mean is the comparison group mean plus the mean difference.
5. For an explanation of the effect size calculation, see [Technical Details of WWC-Conducted Computations](#).
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the [WWC Tutorial on Mismatch](#). See [Technical Details of WWC-Conducted Computations](#) for the formulas the WWC used to calculate statistical significance. In the case of Torgesen et al. (2006) and the fluency domain, no corrections for clustering were needed because students were assigned to conditions. No corrections for multiple comparisons were needed because there is only one outcome in this domain.
9. This row provides the domain average, which in this instance is also the single study finding.

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

Outcome measure	Study sample	Sample size (school units/ students ³)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation ²)		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., 2006 (randomized controlled trial)⁷								
WRMT: Passage Comprehension subtest	Grade 3	8/79	93.16 (15.00)	92.30 (15.00)	0.86	0.06	ns	+2
GRADE: Passage Comprehension subtest	Grade 3	8/79	87.39 (15.00)	83.22 (15.00)	4.18	0.28	ns	+11
Domain average⁹ for comprehension						0.17	ns	+7

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices. 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. 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.
3. The sample size for the analysis was not reported in the study. The sample size reported is the total number of third-grade students in the intervention and control conditions at baseline, which may differ from the actual number of students used in the various analysis in the report.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The intervention group mean is the comparison group mean plus the mean difference.
5. For an explanation of the effect size calculation, see [Technical Details of WWC-Conducted Computations](#).
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the [WWC Tutorial on Mismatch](#). See [Technical Details of WWC-Conducted Computations](#) for the formulas the WWC used to calculate statistical significance. In the case of Torgesen et al. (2006) and the comprehension domain, no corrections for clustering were needed. No corrections for multiple comparisons were needed because the study's reported corrections for multiple comparisons were based on the same group of outcomes as the domain for this review.
9. This row provides the domain average, which in this instance is also the study 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 Corrective Reading 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 *Corrective Reading* as potentially positive. It did not meet the criteria for positive effects because only one study showed a statistically significant positive effect. The remaining ratings (mixed, no discernible effects, potentially negative, or negative) were not considered because *Corrective Reading* 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 a statistically significant positive effect.

AND

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

Met. No studies showed a statistically significant or substantively important negative effect. The single study that met the WWC standards showed a statistically significant positive 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. Only one study showed a statistically significant positive effect.

AND

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

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

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

Appendix A4.2 Corrective Reading 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 *Corrective Reading* as potentially positive. It did not meet the criteria for positive effects because only one study showed statistically significant positive effects. The remaining ratings (mixed, no discernible effects, potentially negative, or negative) were not considered because *Corrective Reading* 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 a statistically significant positive effect.

AND

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

Met. No studies showed a statistically significant or substantively important negative effect. The single study that met the WWC standards showed a statistically significant positive 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. Only one study showed a statistically significant positive effect.

AND

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

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

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

Appendix A4.3 Corrective Reading 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 *Corrective Reading* as having no discernible effects. It did not meet the criteria for other ratings (positive, potentially positive, mixed, potentially negative, and negative) because the single study that met WWC standards did not show statistically significant or substantively important effects.

Rating received

No discernible effects: No affirmative evidence of effects.

- Criterion 1: None of the studies shows a statistically significant or substantively important effect, either *positive* or *negative*.

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

Other ratings considered

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

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

Not met. No studies showed statistically significant positive effects.

AND

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

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

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 statistically significant or 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.

Not met. The single study that met WWC standards 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 statistically significant or substantively important effects, either positive or negative.

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 statistically significant or substantively important effects, either positive or negative.

(continued)

Appendix A4.3 *Corrective Reading* rating for the comprehension domain *(continued)*

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

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

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

AND

- Criterion 2: No studies showing a statistically significant or substantively important *positive* effect, or more studies showing statistically significant or substantively important *negative* effects than showing statistically significant or substantively important *positive* effects.

Met. No studies showed statistically significant or substantively important positive effects. In addition, no studies showed a statistically significant or substantively important negative effect.

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

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

Not met. No studies showed statistically significant negative effects.

AND

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

Met. No studies showed statistically significant or substantively important positive 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
		School units	Students	
Alphabetics	1	8	More than 70	Small
Fluency	1	8	More than 70	Small
Comprehension	1	8	More than 70	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.”