

Appendix

Appendix A1 Study characteristics: Hecht & Close, 2002 (quasi-experimental design)

Characteristic	Description
Study citation	<p>Hecht, S. A., & Close, L. (2002). Emergent Literacy Skills and Training Time Uniquely Predict Variability in Responses to Phonemic Awareness Training in Disadvantaged Kindergartners. <i>Journal of Experimental Child Psychology</i>, 82(2), 93–115.</p> <p><i>Additional source:</i> Hecht, S. A. (2000). <i>Research Compendium: The Waterford Early Reading Program</i>. (Available from Waterford Institute, Inc., 55 West 900 South, Salt Lake City, UT 84101). (Study: Waterford Early Reading program in Ohio.)</p>
Participants	<p>The study began with 140 full-day, at-risk Kindergarten students who were randomly selected from six schools. Students from four schools who received the <i>Waterford Early Reading Program™</i> were matched to students in two schools who did not receive the program. Students were pretested in the fall and posttested in the spring of the same school year. Because of mobility and absences, 64 students attrited from the study. The final analysis sample included 76 students. The mean age of students was five years and seven months. The majority of students were eligible to receive free/reduced lunch.¹ The majority of students in the schools came from low socio-economic status and African-American families.</p>
Setting	<p>The study took place in six inner city or rural public schools in Ohio.</p>
Intervention	<p>Students received the computer-assisted instruction of <i>Waterford Early Reading Program™</i>—Level One (WERP–1) during their normal classroom lessons for six months. The program focused on phonological awareness skills, letter knowledge, print concepts, and oral language skills. Students worked on the Waterford multimedia computer on their own for 15 minutes each session. A teacher management system was used to track daily time use.</p>
Comparison	<p>Students in the comparison group received their regular reading curriculum and were not exposed to the <i>Waterford Early Reading Program™</i>.</p>
Primary outcomes and measurement	<p>Nine outcomes were assessed in the alphabetic domain including the Comprehensive Test of Phonological Process (Phonemic Segmenting, Phonemic Blending, Elision, and Sound Matching subtests), the Woodcock-Johnson Tests of Achievement (the Letter Word Identification subtest), the Wide Range Achievement Test (the Spelling subtest with Phonemic Representation scoring), the Concepts About Print Test, the Letter Name Knowledge and Letter Sound Knowledge measures, and the Stanford-Binet: Fourth Edition Vocabulary subtest. The study also used a letter writing task from the Spelling subtest of the Wide Range Achievement Test, but this test was outside the domains specified by the Beginning Reading protocol (see Appendices A2.1–2.2 for more detailed descriptions of outcome measures).</p>
Teacher training	<p>Information about teacher training was not provided in the study.</p>

1. The WWC received additional information on the analytic sample from the study authors. Baseline equivalence of the intervention and comparison group students remaining in the study was demonstrated by the authors.

Appendix A2.1 Outcome measures in the alphabets domain

Outcome measure	Description
Phonological awareness	
Comprehensive Test of Phonological Processing (CTOPP): Elision subtest	A standardized measure of children's phonological awareness skills. Children were asked to say a word. Then children were asked what the word would be if a specific phoneme in the word were deleted. The remaining phonemes were used to form a word (as cited in Hecht, 2000).
CTOPP: Phonemic Blending subtest	A standardized measure of children's phonemic synthesis skills. It includes four practice items and 15 test items consisting of two- to four-phoneme, one- and two-syllable words. This test measures the total number of words correctly spoken (as cited in Hecht & Close, 2002).
CTOPP: Phonemic Segmenting subtest	A standardized measure of children's phonemic analysis skills. It includes three practice items and 15 test items consisting of two- to five-phoneme single-syllable words. This test measures the total number of words correctly pronounced one phoneme at a time (as cited in Hecht & Close, 2002).
CTOPP: Sound Matching subtest	A standardized measure of children's sound matching skills. Children were asked to pick which of three pictured words began with the same first sound as a target word (as cited in Hecht, 2000).
Letter identification	
Letter Name Knowledge	A researcher-developed measure designed to measure the total number of letter names correctly pronounced (as cited in Hecht & Close, 2002).
Print awareness	
Concepts About Print Test	This 18-question test (Stones version) yielded one score reflecting students' knowledge about print. The score is measured by the total number of correct items (as cited in Hecht & Close, 2002).
Phonics	
Letter Sound Knowledge	A researcher-developed measure designed to measure the total number of letter sounds correctly pronounced (as cited in Hecht & Close, 2002).
Wide Range Achievement Test: Spelling subtest with phonemic representation scoring	Students wrote 15 words as dictated by the test administrator. Scoring was based on Wilkinson's method of giving partial credit for accuracy of phonemic representation (as cited in Hecht & Close, 2002). Students received between 0 and 6 points depending on how many and the placement of phonemes were represented by phonemically related or conventional letters in each written word.
Woodcock-Johnson Tests of Achievement-Revised: Letter Word Identification subtest	A standardized measure of children's word reading. Children identified various letters of the alphabet as well as words, ranging from commonly used words to less familiar words of the English language (as cited in Hecht, 2000).

Appendix A2.2 Outcome measure in the comprehension domain

Outcome measure	Description
<i>Vocabulary</i>	
Stanford-Binet (4th ed.): Vocabulary subtest	A standardized measure to assess general cognitive ability and estimate general verbal IQ. The score is measured by the total number of correctly defined words (as cited in Hecht & Close, 2002).

Appendix A3.1 Summary of study findings included in the rating for the alphabetics domain by construct¹

Outcome measure	Study sample	Sample size (schools/students)	Authors' findings from the study			WWC calculations		
			Mean outcome (standard deviation ²)		Mean difference ³ (Waterford – comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
			Waterford group	Comparison group				
Hecht & Close, 2002 (quasi-experimental design)⁷								
<i>Phonological awareness</i>								
CTOPP: Elision subtest	Kindergarten	6/76	4.71 (3.45)	2.82 (2.39)	1.89	0.62	ns	+23
CTOPP: Phonemic Blending subtest	Kindergarten	6/76	9.53 (5.55)	4.24 (5.08)	5.29	0.98	ns	+34
CTOPP: Phonemic Segmenting subtest	Kindergarten	6/76	7.58 (7.05)	1.53 (2.84)	6.05	1.07	ns	+36
CTOPP: Sound Matching subtest	Kindergarten	6/76	10.91 (4.71)	6.27 (4.89)	4.64	0.96	ns	+33
<i>Letter identification</i>								
Letter Name Knowledge	Kindergarten	6/76	21.58 (4.43)	24.65 (4.14)	-3.07	-0.71	ns	-26
<i>Print awareness</i>								
Concepts About Print Test	Kindergarten	6/76	8.58 (3.05)	9.01 (4.57)	-0.43	-0.11	ns	-4
<i>Phonics</i>								
Letter Sound Knowledge	Kindergarten	6/76	19.09 (8.87)	22.55 (9.33)	-3.46	-0.38	ns	-15
WRAT: Spelling subtest with phonemic representation scoring	Kindergarten	6/76	25.57 (19.67)	8.09 (7.79)	17.48	1.11	ns	+37
Woodcock-Johnson Tests of Achievement-Revised: Letter Word Identification subtest	Kindergarten	6/76	3.54 (3.43)	0.77 (1.16)	2.67	0.99	ns	+34
Domain average⁸ for alphabetics						0.50	ns	+19

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for alphabetics.

(continued)

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

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 for CTOPP Elision and Sound Matching subtests were received from the first author.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The intervention group mean in this table equals the comparison group mean plus the mean difference. The mean difference is calculated as the difference between gain scores and takes into account the pretest difference between the study groups.
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 Hecht & Close (2002), corrections for clustering and multiple comparisons were needed, so the significance levels differ from those reported in the original study.
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 comprehension domain¹

Outcome measure	Study sample	Sample size (schools/students)	Authors' findings from the study					
			Mean outcome (standard deviation ²)		WWC calculations			
			Waterford group	Comparison group	Mean difference ³ (Waterford – comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
Hecht & Close, 2002 (quasi-experimental design)								
Stanford-Binet (4th ed.): Vocabulary subtest	Kindergarten	6/76	16.91 (3.66)	16.58 (3.35)	0.33	0.09	ns	+4

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for comprehension.
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 intervention group mean in this table equals the comparison group mean plus the mean difference. The mean difference is calculated as difference between gain scores and takes into account the pretest difference between the study groups.
3. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see *Technical Details of WWC-Conducted Computations*.
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition 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.

Appendix A4.1 Waterford Early Reading Program™ 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 *Waterford Early Reading Program™* as having potentially positive effects. It did not meet the criteria for positive effects because no studies showed statistically significant positive effects. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, negative effects) were not considered because the intervention 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. The single study of *Waterford Early Reading Program™* 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 of *Waterford Early Reading Program™* showed statistically significant or substantively important negative effects, and no studies showed indeterminate 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 of *Waterford Early Reading Program™* showed statistically significant positive effects.

AND

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

Met. The single study of *Waterford Early Reading Program™* did not show 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 Waterford Early Reading Program™ 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 *Waterford Early Reading Program™* as having no discernible effects. It did not meet the criteria for other ratings (positive effects, potentially positive effects, mixed effects, potentially negative effects, and negative effects) because the one study that met WWC standards with reservations 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. The one study of *Waterford Early Reading Program™* did not show 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. The one study of *Waterford Early Reading Program™* did not show a statistically significant positive effect.

AND

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

Met. The one study of *Waterford Early Reading Program™* did not show 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. The one study of *Waterford Early Reading Program™* did not show a statistically significant or substantively important positive effect.

AND

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

Met. The one study of *Waterford Early Reading Program™* did not show a statistically significant or substantively important negative effect. This study showed an indeterminate effect.

(continued)

Appendix A4.2 Waterford Early Reading Program™ rating for the comprehension domain (continued)

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. The one study of *Waterford Early Reading Program™* did not show statistically significant or substantively important effects.

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.

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.

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 negative effects.

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
		Schools	Students	
Alphabetics	1	6	76	Small
Fluency	0	0	0	na
Comprehension	1	6	76	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.”