

Appendix

Appendix A1.1 Study characteristics: Beattie, 2000 (randomized controlled trial with severe attrition)

Characteristic	Description
Study citation	Beattie, K. K. (2000). The effects of intensive computer-based language intervention on language functioning and reading achievement in language-impaired adolescents (Doctoral dissertation, George Mason University, 2000). <i>Dissertation Abstracts International</i> , 61(08A), 194–3116.
Participants	Eighty-one students with language deficits, ranging in age from 11 to 16 years, were randomly assigned by computer-generated procedures to one of four intervention groups ¹ or to a control group in a two-step process. The researchers first assigned 18 students to the two intervention groups (that received a phase of <i>SuccessMaker</i> [®] and <i>Fast ForWord</i> [®] and also concomitantly participated in a functional resonance imaging research project). Then, the remaining participants were randomly assigned across the five groups. To ensure an equal distribution among groups, fewer students were placed in the first two groups at the second step of randomization. For this review, the WWC reported results from 14 students in the <i>SuccessMaker</i> [®] group who were compared to 12 students in the comparison group. ² Although the differential attrition rate was higher than 7%, the post-attrition intervention and comparison groups were equivalent on the pretest achievement measures.
Setting	Two middle schools and one middle-high school located in the suburbs of a large metropolitan area in northern Virginia.
Intervention	Students worked on <i>SuccessMaker</i> [®] for 90–94 minutes a day, five days a week. The intervention ended after each student completed 64–80 hours on the program. The study reported student outcomes after two months of program implementation.
Comparison	The control group received the standard instruction provided in the regular school curriculum.
Primary outcomes and measurement	For both pre- and posttests, the author administered the Gray Oral Reading Test, four subtests of the Woodcock-Johnson Psycho-Educational Battery (Letter-Word Identification, Word Attack, Passage Comprehension, and Auditory Processing), the Spelling subtest of the Wide Range Achievement Test, and the Receptive Language subtest of the Clinical Evaluation of Language Fundamentals. For a more detailed description of these outcome measures, see Appendices A2.1–A2.4.
Staff/teacher training	No information on training for the teachers and staff in this study was provided. To facilitate the use of <i>SuccessMaker</i> [®] , computers were procured or updated to meet criteria for running <i>SuccessMaker</i> [®] software.

1. The first intervention group received two phases of *Fast ForWord*[®]; the second intervention group received two phases of *SuccessMaker*[®]; the third and fourth intervention groups received a phase of *Fast ForWord*[®] and a phase of *SuccessMaker*[®].
2. The analysis samples for *SuccessMaker*[®] and *Fast ForWord*[®] groups were not shown to be equivalent at baseline. Two other groups, which combined *SuccessMaker*[®] and *Fast ForWord*[®], are not appropriate counterfactuals, because the measures of effects cannot be attributed solely to the *SuccessMaker*[®] program.

Appendix A1.2 Study characteristics: Campbell, 2000 (quasi-experimental design)

Characteristic	Description
Study citation	Campbell, J. P. (2000). A comparison of computerized and traditional instruction in the area of elementary reading (Doctoral dissertation, University of Alabama, 2000). <i>Dissertation Abstracts International</i> , 61(03A), 77–952.
Participants	Based on the School Ability Index score, five elementary schools that used both <i>SuccessMaker</i> ® and traditional instruction were matched to five elementary schools that used only traditional instruction. ¹ Poverty level and gender were similar across intervention and comparison schools. Although the overall and differential student attrition rates were high (58% and 37%, respectively), the post-attrition intervention and comparison samples of fourth-graders were equivalent on both subtests of the Stanford Achievement Test at baseline. ² After one year, 143 students in four <i>SuccessMaker</i> ® schools and 186 students in four comparison schools remained in the sample.
Setting	The analysis sample included eight elementary schools in Etowah County, Alabama.
Intervention	Students in the intervention group received 10 to 20 minutes of <i>SuccessMaker</i> ® instruction daily. They were also given traditional instruction that included the <i>Accelerated Reader</i> program in conjunction with a basal reader. The study was conducted during the first year of <i>SuccessMaker</i> ® program implementation.
Comparison	Comparison classrooms implemented the standard district curriculum, which used the <i>Accelerated Reader</i> program in conjunction with a basal reader.
Primary outcomes and measurement	For both pre- and posttests, the author used two subtests of the Stanford Achievement Test administered by schools. The Vocabulary and Reading Comprehension Otis Lennon School Ability test was also used in the study, but was not included in this report because it was outside the scope of the Adolescent Literacy review. For a more detailed description of the outcome measures included in this report, see Appendix A2.3.
Staff/teacher training	In order to maintain consistency in the administration of the outcome measure (SAT-9), all test administrators and proctors were trained in the areas of test security and proper administration techniques.

1. For the overall grade 5 analysis sample, the intervention and comparison groups were not shown to be equivalent at baseline and are, therefore, excluded from review. As a result, two schools were dropped from the analysis.
2. WWC aggregated reading achievement data across schools to conduct the analyses.

Appendix A1.3 Study characteristics: Gallagher, 1996 (quasi-experimental design)

Characteristic	Description
Study citation	Gallagher, E. M. (1996). Utilization of an ILS to increase reading comprehension (integrated learning systems, CAI) (Doctoral dissertation, Northern Illinois University, 1996). <i>Dissertation Abstracts International</i> , 58(05A), 79–159.
Participants	Students in grades 4–7 were pretested using the Iowa Test of Basic Skills (ITBS), sorted by either the ITBS score or student identification number (ID), and then alternately assigned to treatment or control groups within classrooms. ¹ All of the students were African-American and were eligible for the federal free lunch program. Students who scored below 3.0 on the reading comprehension subtest or who were part of the school's special education program were eliminated from the study's sample prior to the assignment. Although the overall attrition rate at posttest was 38%, the post-attrition intervention and comparison groups were equivalent on the reading achievement pretest measure (ITBS). In all, 48 students in the <i>SuccessMaker</i> ® group and 47 students in the comparison group were included in the analysis sample. Additional findings reflecting student outcomes by grade can be found in Appendix A4.
Setting	The study took place in an inner city elementary school in Chicago, Illinois.
Intervention	The intervention group spent a minimum of 40 minutes a day on the two reading components of the <i>SuccessMaker</i> ® program. The Readers Workshop component is an individualized basic skill building program. In the first 100 minutes a student participates in the program, the computer analyzes their skills development and assigns specific segments of the program appropriate to further develop the students' skills, introducing new skills as they become appropriate. The Reading Adventures component places each student at a reading level and provides stories and comprehension questions at that level. The student progresses through a semi-linear program where the only choice is among stories at the assigned level. Outside of the <i>SuccessMaker</i> ® instruction, the intervention group also received the regular reading curriculum. The study reported students' outcomes after six weeks of program implementation.
Comparison	The comparison group spent a minimum of 40 minutes a day on the math components of the <i>SuccessMaker</i> ® program (Math Concepts and Skills and Problem Solving). Comparison students also received the regular reading curriculum.
Primary outcomes and measurement	For both pre- and posttests, the author used the reading comprehension subtest of the Iowa Test of Basic Skills. For a more detailed description of this outcome measure, see Appendix A2.3.
Staff/teacher training	No information on training for the teachers and staff in this study was provided.

1. The authors *either* sorted the students by student identification numbers (ID) or Iowa Test of Basic Skills (ITBS) reading comprehension scores, and *then* assigned students to groups in an alternating fashion, but it is not clear which method was used from the text. If they sorted by student ID and then assigned students to groups, the assignment might be functionally random, but if they sorted by ITBS score, and always assigned students in an alternating fashion (starting with the treatment group, for example), the groups would be imbalanced, because they were always assigning the lower (or higher) scores to the treatment group. The WWC could not confirm that the assignment was truly random, as the authors had not responded to the WWC query at the time of publication of this review.

Appendix A2.1 Outcome measures for the alphabetic domain

Outcome measure	Description
<i>Phonemic awareness</i>	
Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R), Tests of Cognitive Abilities: Auditory Processing subtest	This composite is a standardized measure of a student's ability to appreciate patterns among speech-based auditory stimuli. The score is derived from scores on three subtests: (1) the Sound Blending subtest measures the ability to synthesize sequences of sounds into whole words; (2) the Incomplete Words subtest measures the ability to identify a word with missing sounds; and (3) the Sound Patterns subtest measures ability to indicate whether pairs of computer-generated sound sequences are the same or different (as cited in Beattie, 2000).
<i>Phonics</i>	
WJ-R, Tests of Achievement: Word Attack subtest	This standardized subtest measures phonemic decoding skills by asking students to read pseudowords (e.g., plurp, fronkett). Students are aware that the words are not real (as cited in Beattie, 2000 and http://www.concordspedpac.org/WJ-III-subtests.htm#Achievement).
WJ-R, Tests of Achievement: Letter-Word Identification subtest	This standardized subtest requires the student to read aloud isolated letters and real words that range in frequency and difficulty (as cited in Beattie, 2000).
Wide Range Achievement Test-Third Edition (WRAT-3): Spelling subtest	This standardized subtest is a paper-and-pencil task that tests students' ability to write their names, as well as letters and words from dictation. Dictated letters and words followed either phonetically regular or irregular patterns (as cited in Beattie, 2000).

Appendix A2.2 Outcome measures for the reading fluency domain

Outcome measure	Description
Gray Oral Reading Test-Third edition (GORT-3)	In this standardized test, students are required to read orally a variety of graded passages to measure reading rate, word identification, and comprehension skills. The Passage subtest assesses a combination of rate and accuracy. The Comprehension subtest requires a student to respond to five multiple choice questions following each story. The Oral Reading Quotient is reflective of a total measure of one's oral reading performance and is calculated by combining the Passage and Comprehension scores (as cited in Beattie, 2000).

Appendix A2.3 Outcome measures for the comprehension domain

Outcome measure	Description
Vocabulary development	
The Stanford Achievement Test (SAT–9): Reading Vocabulary subtest	This standardized subtest is composed of multiple-choice and open-ended assessment questions that measure word reading and achievement. The open-ended reading section includes a narrative reading selection followed by nine questions. There are three types of reading selections: (1) recreational (material read for enjoyment or literary merit, including folk tales, historical fiction, contemporary fiction, humor, and poetry), (2) textual (expository material with content from the natural, physical, and social sciences, as well as other nonfiction general information materials) and (3) functional (material encountered in everyday life both inside and outside of school, including directions, forms, labels, schedules, and advertisements) (as cited in Campbell, 2000 and http://brighted.funeducation.com/Prepare/StateTests/?state=SAT-9).
Reading comprehension	
SAT–9: Reading Comprehension subtest	This standardized subtest is based on questions that range from interpreting simple sentences to understanding more complex paragraphs. The questions on complex paragraphs ask the student to recognize directly stated details or relationships, as well as implicit information and relationships that demand integration of what is provided in the text (as cited in Campbell, 2000).
WJ–R, Tests of Achievement: Passage Comprehension subtest	In this standardized test, comprehension is measured by having students fill in missing words in a short paragraph (e.g., “Woof,” said the _____, biting the hand that fed it.) (as cited in Beattie, 2000 and http://www.concordspedpac.org/WJ-III-subtests.htm#Achievement).
The Iowa Test of Basic Skills: Reading Comprehension subtest	This standardized test consists of reading passages of varying length and difficulty and assesses three types of understanding: (1) factual questions tap students’ literal understanding of what is stated in the text; (2) inferential/interpretive questions require students to “read between the lines” to demonstrate their understanding of what is implied; and (3) analysis and generalization questions require students to “step back from” the text to generalize about a passage’s main points or ideas or to analyze aspects of the author’s viewpoint or use of language (as cited in http://www.riverpub.com/products/itbs/details.html).

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

Outcome measure	Description
Clinical Evaluation of Language Fundamentals–Third Edition (CELF–3): Receptive Language Score	This standardized assessment measures a student’s ability to interpret and execute commands of increasing complexity and understand relationships between words and categories. It addresses sentence structure, concepts and directions, and word classes (as cited in Beattie, 2000).

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

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study		WWC calculations			
			Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>SuccessMaker</i> [®] – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			<i>SuccessMaker</i> [®] group	Comparison group				
Beattie, 2000 (randomized controlled trial with attrition)⁸								
WJ–R Letter-Word Identification subtest	11–16 yrs old	26	89.69 (9.48)	92.08 (13.15)	–2.39	–0.20	ns	–8
WJ–R Word Attack subtest	11–16 yrs old	26	86.99 (17.65)	85.91 (12.87)	1.08	0.07	ns	+3
WJ–R Auditory Processing subtest	11–16 yrs old	26	87.44 (13.38)	85.66 (15.61)	1.78	0.12	ns	+5
WRAT–3 Spelling subtest	11–16 yrs old	26	87.02 (12.66)	85.66 (13.13)	1.36	0.10	ns	+4
Average for alphabetics (Beattie, 2000)⁹						0.02	ns	+1

ns = not statistically significant

WJ–R = Woodcock-Johnson Revised

WRAT-3 = Wide Range Achievement Test–Third Edition

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the alphabetics domain.
2. The intervention group values are the comparison group means plus the difference in mean gains between the intervention and comparison groups.
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.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison 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 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. For the formulas the WWC used to calculate 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 Beattie (2000), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study.
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 A3.2 Summary of study findings included in the rating for the reading fluency domain¹

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study		WWC calculations			
			Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>SuccessMaker</i> [®] – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			<i>SuccessMaker</i> [®] group	Comparison group				
Beattie, 2000 (randomized controlled trial with attrition)⁸								
Gray Oral Reading test (GORT–3)	11–16 yrs old	26	83.18 (12.72)	79.50 (17.76)	3.68	0.23	ns	+9
Average for reading fluency (Beattie, 2000)⁹						0.23	ns	+9

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the reading fluency domain.
2. The intervention group values are the comparison group means plus the difference in mean gains between the intervention and comparison groups.
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.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison 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 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, see the WWC Tutorial on Mismatch. 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 Beattie (2000), no corrections for clustering and multiple comparisons were needed.
9. This row provides the study average, which in this instance is also the domain average. 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 (clusters/students)	Authors' findings from the study ²			WWC calculations		
			Mean outcome ³ (standard deviation) ⁴		Mean difference ⁵ (<i>SuccessMaker</i> [®] – comparison)	Effect size ⁶	Statistical significance ⁷ (at $\alpha = 0.05$)	Improvement index ⁸
			<i>SuccessMaker</i> [®] group	Comparison group				
Beattie, 2000 (randomized controlled trial with attrition)⁹								
WJ–R Passage Comprehension subtest	11–16 yrs old	26	97.03 (8.08)	93.25 (11.30)	3.78	0.38	ns	+15
Average for comprehension (Beattie, 2000)¹⁰						0.38	ns	+15
Campbell, 2000 (quasi-experimental design)⁹								
SAT–9 Reading Vocabulary subtest	Grade 4	8/329	60.54 (23.36)	60.01 (24.12)	0.53	0.02	ns	+1
SAT–9 Reading Comprehension subtest	Grade 4	8/329	60.29 (23.14)	58.08 (24.76)	2.21	0.09	ns	+4
Average for comprehension (Campbell, 2000)¹⁰						0.06	ns	+2
Gallagher, 1996 (quasi-experimental design)⁹								
ITBS Reading Comprehension subtest	Grades 4–7	95	30.25 (10.78)	26.72 (8.32)	3.53	0.36	ns	+14
Average for comprehension (Gallagher, 1996)¹⁰						0.36	ns	+14
Domain average for comprehension across all studies¹⁰						0.27	na	+11

ns = not statistically significant

na = not applicable

WJ–R = Woodcock–Johnson–Revised

SAT–9 = Stanford Achievement Test

ITBS = Iowa Test of Basic Skills

- This appendix reports findings considered for the effectiveness rating and the average improvement indices for the comprehension domain.
- For Gallagher (1996), the WWC calculated groups' sample sizes, means, and standard deviations from the raw data presented in the study appendices.
- The intervention group values are the comparison group means plus the difference in mean gains between the intervention and comparison groups.
- 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. For Campbell (2000), the WWC aggregated means and standard deviations across four schools.
- Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison 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 comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
- The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation, see the WWC Tutorial on Mismatch. 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 Campbell (2000), corrections for clustering and multiple comparisons were needed, so the significance levels may differ from those reported in the original study. In the cases of Beattie (2000) and Gallagher (1996), no corrections for clustering or multiple comparisons were needed.
- The WWC-computed domain average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect size.

Appendix A3.4 Summary of study findings included in the rating for the general literacy achievement domain¹

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study		WWC calculations			
			Mean outcome ² (standard deviation) ³		Mean difference ⁴ (<i>SuccessMaker</i> [®] – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			<i>SuccessMaker</i> [®] group	Comparison group				
Beattie, 2000 (randomized controlled trial with attrition)⁸								
Receptive Language subtest (CELF–3)	11–16 yrs old	26	92.81 (18.35)	86.63 (22.74)	5.98	0.28	ns	+11
Average for general literacy achievement (Beattie, 2000)⁹						0.28	ns	+11

ns = not statistically significant

CELF–3 = Clinical Evaluation of Language Fundamentals–Third Edition

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the general literacy achievement domain.
2. The intervention group values are the comparison group means plus the difference in mean gains between the intervention and comparison groups.
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.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison 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 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, see the WWC Tutorial on Mismatch. 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 Beattie (2000), no corrections for clustering and multiple comparisons were needed.
9. This row provides the study average, which in this instance is also the domain average. The domain improvement index is calculated from the average effect size.

Appendix A4 Summary of subgroup findings for the comprehension domain¹

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study ²					
			Mean outcome ³ (standard deviation) ⁴		WWC calculations			
			SuccessMaker [®] group	Comparison group	Mean difference ⁵ (SuccessMaker [®] – comparison)	Effect size ⁶	Statistical significance ⁷ (at $\alpha = 0.05$)	Improvement index ⁸
Gallagher, 1996 (quasi-experimental design)⁹								
ITBS Reading comprehension subtest	Grade 4	32	31.69 (8.01)	29.19 (5.53)	2.50	0.35	ns	+14
ITBS Reading comprehension subtest	Grade 5	32	27.31 (10.74)	25.00 (9.64)	2.31	0.22	ns	+9
ITBS Reading comprehension subtest	Grade 6	20	34.60 (12.43)	24.80 (6.12)	9.80	0.96	Statistically significant	+33
ITBS Reading comprehension subtest	Grade 7	11	27.47 (14.22)	28.20 (14.13)	–0.73	–0.05	ns	–2

ns = not statistically significant

ITBS = Iowa Test of Basic Skills

1. This appendix presents subgroup findings for measures that fall in the comprehension domain. Total group scores were used for rating purposes and are presented in Appendix A3.3.
2. For Gallagher (1996), the WWC calculated groups' sample sizes, means, and standard deviations from the raw data presented in the study appendices.
3. The intervention group values are the comparison group means plus the difference in mean gains between the intervention and comparison groups.
4. 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.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
6. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B.
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
8. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. For the formulas the WWC used to calculate statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. In the case of Gallagher (1996), no correction for clustering was needed.

Appendix A5.1 *SuccessMaker*[®] rating for the alphabets 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 alphabets, the WWC rated *SuccessMaker*[®] as having no discernible effects. It did not meet the criteria for positive, potentially positive, mixed, potentially negative, or negative effects because no studies showed statistically significant or substantively important effects, either positive or negative.

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 statistically significant or substantively important effects, 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 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 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, and 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 effect, 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 a statistically significant or substantively important effect, and one study showed indeterminate effects.

(continued)

Appendix A5.1 SuccessMaker® rating for the alphabetics 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 a statistically significant or substantively important negative effect.

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.

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 or substantively important 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. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A5.2 SuccessMaker® 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 *SuccessMaker*® as having no discernible effects. It did not meet the criteria for positive, potentially positive, mixed, potentially negative, or negative effects because no studies showed statistically significant or substantively important effects, either positive or negative.

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 statistically significant or substantively important effects, 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 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 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, and 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 effect, 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 a statistically significant or substantively important effect, and one study showed indeterminate effects.

(continued)

Appendix A5.2 SuccessMaker® rating for the reading fluency 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 a statistically significant or substantively important negative effect.

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.

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 or substantively important 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. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A5.3 SuccessMaker® 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 *SuccessMaker*® as having potentially positive effects.

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. Two studies showed substantively important positive effects.

AND

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

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

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 A5.4 SuccessMaker® rating for the general literacy achievement domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.¹ For the outcome domain of general literacy achievement, the WWC rated SuccessMaker® as having potentially positive effects.

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

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	3	26	Small
Reading fluency	1	3	26	Small
Comprehension	3	12	450	Medium to large
General literacy achievement	1	3	26	Small

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.