Appendix

Appendix A1.1 Study characteristics: Kemp, 2006

Characteristic	Description
Study citation	Kemp, S. C. (2006). Teaching to *Read Naturally*: Examination of a fluency training program for third grade students (Doctorial dissertation, University of California, Irvine and University of California, Los Angeles, 2006). *Dissertation Abstracts International, 67*(07A), 95–2447.
Participants	A randomized controlled trial was used to examine the effects of *Read Naturally*® on third-grade reading performance. A total of 42 English language learners, from three elementary schools across 13 classrooms, initially participated in the study. Students in each participating classroom were ranked by standardized tests of reading and then randomly assigned to either the *Read Naturally*® intervention group or the scaffolded sustained silent reading (SSSR) comparison group. Of the 42 original students, 21 were assigned to the *Read Naturally*® group and 21 were assigned to the SSSR group. Three students were excluded from the study because they were receiving special education services. The analysis sample consisted of 39 English language learners; 20 students in the intervention group, and 19 students in the comparison group.
Setting	The study was conducted in three schools in a suburban school district located in western Orange County, California. The intervention (*Read Naturally*®) and comparison (SSSR) conditions were implemented in each classroom.
Intervention	The *Read Naturally*® program was implemented four days per week for 20 minutes a day during the months of October through January. *Read Naturally*® consists of teaching modeling, repeated reading, and progress monitoring for the purpose of promoting fluency. Students are assigned to instructional level reading materials. When participating in the program, students (1) practice a "cold reading" of a self-selected passage from their assigned reading level, (2) practice reading the same passage three or four times with an audio recorded model, (3) practice reading independently until they reach their timed goal, and (4) meet with the classroom teacher so a timed reading sample can be documented.
Comparison	Students in the comparison condition participated in scaffolded sustained silent reading (SSSR), which involved teaching students to select materials at their individual reading level. Students then engaged in independent, silent reading. Teachers did not provide significant feedback; they walked around the room and monitored whether or not students were documenting the number of pages they read. As in the case of *Read Naturally*®, use of SSSR occurred from October through January, four days a week, for 20 minutes each day.
Primary outcomes and measurement[1]	Study measures in the reading achievement domain included the Test of Oral Word Reading Efficiency (TOWRE) Sight Word and Phonemic Decoding Efficiency subtests; the Dynamic Indicator of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency subtest; the Stanford Diagnostic Reading test, 4th Edition, Vocabulary and Comprehension subtests; the Orthographic Choice test; and the Morphological Relatedness Test (MRT), Written and Oral/Written versions. All measures were administered at pre- and post-test. (For a more detailed description of these outcome measures, see Appendices A2.1–A2.2.)
Staff/teacher training	Thirteen general education teachers received training on both the *Read Naturally*® program and the use of SSSR. No additional details were provided.

1. The Bear Spelling Inventory, the Title Recognition test, and the Rosner Auditory Analysis test were also administered but not included in the review because they were not eligible outcomes as determined by the English Language Learners protocol.

Appendix A1.2 Study characteristics: Denton, Anthony, Parker, & Hasbrouck, 2004

Characteristic	Description
Study citation	Denton, C. A., Anthony, J. L., Parker, R., & Hasbrouck, J. E. (2004). Effects of two tutoring programs on the English reading development of Spanish-English bilingual students. *The Elementary School Journal, 104*(4), 289–305.
Participants	The study included a group of 93 students between second and fifth grade who were bilingual with Spanish as their native language, were recommended by their teacher for tutoring, and had standardized assessments suggesting they had adequate oral English proficiency and basic proficiency in reading Spanish. The students were enrolled in 17 bilingual classrooms in five schools. Students' ages ranged from 7 to 12 years with a mean age of 9 years; 48 were males and 45 were females. Students were assigned to one of two reading ability groups based on their scores on the Word Attack subtest of the Woodcock Reading Mastery Tests–Revised (WRMT-R). Students with scores below first grade equivalency were assigned to the "emergent" decoding group (*Read Well*), and those with scores above first grade equivalency were in the "established" decoding group (*Read Naturally®*). Within each of these groups, students were matched on pretest scores and randomly assigned to either the treatment or comparison group. A total of 63 students were initially assigned in the *Read Naturally®* study (32 in the treatment group and 31 in the control group). Three students originally assigned to the control group participated in the treatment and were ultimately dropped from the study. Additionally, three students originally assigned to the treatment group were moved to the comparison group one week after the study had begun (as requested by one of the participating schools). As a result of these changes, the study was treated as a quasi-experimental design that meets WWC evidence standards with reservations. The *Read Naturally®* analysis sample consisted of 60 students (32 treatment and 28 comparison).
Setting	The study took place as a pull-out tutoring program in five elementary schools in a central Texas district. During the school year, the district served a population of 43.1% Hispanic students; 56.2% of children in the district were identified as economically disadvantaged; 9% had limited English proficiency; and 7.3% were served by a bilingual or ESL program.
Intervention	The intervention occurred during pull-out tutoring sessions during the school day when the participants were not receiving their English instruction. Students in *Read Naturally®* were tutored three times per week for 40-minute periods over 10 weeks. The sessions consisted of repeated oral reading of connected text, vocabulary and comprehension instruction, and systematic monitoring of progress in the program. The standard *Read Naturally®* program was modified for use with English language learners by adding and extending activities related to vocabulary, decoding, and comprehension (such as oral discussions of vocabulary and comprehension and preteaching important or challenging vocabulary in reading passages).
Comparison	The comparison condition received the same regular education curriculum as the treatment group but did not receive any additional tutoring beyond what would have been part of the schools' business-as-usual approach.
Primary outcomes and measurement[1]	The study measures in the reading achievement domain included three subtests of the Woodcock Reading Mastery Test–Revised: Word Attack, Word Identification, and Passage Comprehension. (For a more detailed description of these outcome measures, see Appendices A2.1–A2.2.)
Staff/teacher training	Tutors were 23 undergraduate students enrolled in a class in teaching students with difficulties. Tutors received training in the implementation of both the *Read Naturally®* and *Read Well* programs as part of their course instruction. They were supervised by a graduate student experienced in *Read Naturally®*.

1. The measures in the reading achievement domain also included a researcher-developed oral reading assessment, but it was not included in the study because of unreliable administration and missing data (Denton et al., 2004).

Appendix A2.1 Outcome measures for the reading achievement domain

Outcome measure	Description
Test of Oral Word Reading Efficiency (TOWRE) Sight Word Efficiency (SWE) subtest	The TOWRE assessment is a nationally normed, age-based measure of word reading accuracy and fluency. The SWE subtest assessed the number of real printed words that could be accurately identified within 45 seconds (as cited by Kemp, 2006).
Test of Oral Word Reading Efficiency (TOWRE) Phonemic Decoding Efficiency (PDE) subtest	The TOWRE assessment is a nationally normed, age-based measure of word reading accuracy and fluency. The PDE subtest measured the number of pronounceable printed non-words that could be accurately decoded within 45 seconds (as cited by Kemp, 2006).
Dynamic Indicator of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency subtest	The DIBELS assessment is specifically designed to assess fluency with connected text. The Oral Reading Fluency subtest had students read an unfamiliar passage of grade-level material for one minute. Three passages were given. For each passage, the number of words read correctly in one minute was recorded. The final score was the middle score obtained from the three passages (as cited by Kemp, 2006).
The Stanford Diagnostic Reading Test, 4th Edition–Comprehension subtest	The Stanford Diagnostic Reading Test is a nationally norm-referenced test of reading comprehension. It provides criterion-referenced information to help teachers with instructional planning. The comprehension subtest was administered to the whole class, and raw scores and percentile scores were obtained (as cited by Kemp, 2006).
The Stanford Diagnostic Reading Test, 4th Edition–Vocabulary subtest	The Stanford Diagnostic Reading Test is a nationally norm-referenced test of reading comprehension. It provides criterion-referenced information to help teachers with instructional planning. The vocabulary subtest was administered to the whole class, and raw scores and percentile scores were obtained (as cited by Kemp, 2006).
Woodcock Reading Mastery Test–Revised (WRMT-R) Word Attack subtest	Word Attack assesses phonemic decoding and involves reading a list of nonwords (as cited by Denton, Anthony, Parker, & Hasbrouck, 2004).
Woodcock Reading Mastery Test–Revised (WRMT-R) Word Identification subtest	Word Identification is a measure of decoding. Students were asked to read words in a list format (as cited by Denton, Anthony, Parker, & Hasbrouck, 2004).
Woodcock Reading Mastery Test–Revised (WRMT-R) Passage Comprehension subtest	Passage Comprehension is a test of reading comprehension in which students are asked to read a brief passage that has a word omitted and to supply the target word or an acceptable alternative (as cited by Denton, Anthony, Parker, & Hasbrouck, 2004).

Appendix A2.2 Outcome measures for the English language development domain

Outcome measure	Description
Orthographic Choice test	The Orthographic Choice test measured orthographic awareness by presenting 17 pairs of pronounceable pseudowords. One pseudoword of each pair contained a letter pair that never occurs in English in the initial or final position, and the other word contained an orthographically appropriate letter pair in the same position (e.g., filv, filk). The students were asked, "You are going to see pairs of letter strings that are not words. One of them looks more like a word than the other. I want you to circle the word that looks more like a word than the other. Which one has spelling that is more like a word?" The maximum score of this task was 17 (as cited by Kemp, 2006).
The Morphological Relatedness Test (MRT)–Written version	The MRT assessment consisted of 40 items divided equally between the Written and the Oral/Written versions. Students determined whether or not the second word in each pair was derived from the first word and circled either "yes" or "no" after each pair. In the Written version, students silently read the items before marking their answers. The items included in this assessment were pairs of words adopted from Mahony (1993) and some additional pairs that Mann (2000) created. Each version of the test contained 15 related pairs and five unrelated pairs or foils. The maximum score of both versions of the MRT was 20 (as cited by Kemp, 2006).
The Morphological Relatedness Test (MRT)–Oral/Written version	The MRT assessment consisted of 40 items divided equally between the Written and the Oral/Written versions. Students determined whether or not the second word in each pair was derived from the first word and circled either "yes" or "no" after each pair. In the Oral/Written version, the experimenter read each item aloud. The items included in this assessment were pairs of words adopted from Mahony (1993) and some additional pairs that Mann (2000) created. Each version of the test contained 15 related pairs and five unrelated pairs or foils. The maximum score of both versions of the MRT was 20 (as cited by Kemp, 2006).

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

Outcome measure	Study sample	Sample size (students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Read Naturally*® group[3]	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[4] (*Read Naturally*® – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Kemp, 2006[8]								
TOWRE Sight Word subtest	3rd grade	39	57.90 (12.15)	57.74 (10.27)	0.16	0.01	ns	+1
TOWRE Phonemic Decoding Efficiency subtest	3rd grade	39	29.15 (12.97)	27.58 (9.91)	1.57	0.13	ns	+5
DIBELS Oral Reading Fluency subtest	3rd grade	39	94.08 (32.00)	89.37 (27.92)	4.71	0.15	ns	+6
Stanford Diagnostic Reading Test Comprehension subtest	3rd grade	39	30.98 (6.29)	30.37 (6.44)	0.61	0.09	ns	+4
Stanford Diagnostic Reading Test Vocabulary subtest	3rd grade	39	29.10 (6.22)	28.63 (7.58)	0.47	0.07	ns	+3
Average for reading achievement (Kemp, 2006)[9]						**0.09**	**ns**	**+4**
Denton, Anthony, Parker, & Hasbrouck, 2004[8]								
WRMT-R Word Identification subtest	2nd–5th grade	60	95.37 (11.65)	95.94 (9.65)	–0.57	–0.05	ns	–2
WRMT-R Word Attack subtest	2nd–5th grade	60	96.74 (9.37)	98.04 (9.00)	–1.30	–0.14	ns	–6
WRMT-R Passage Comprehension subtest	2nd–5th grade	60	90.45 (7.90)	89.28 (10.26)	1.17	0.13	ns	+5
Average for reading achievement (Denton, Anthony, Parker, & Hasbrouck, 2004)[9]						**–0.02**	**ns**	**–1**
Domain average for reading achievement across all studies[9]						**0.04**	**na**	**+1**

ns = not statistically significant
na = not applicable
TOWRE = Test of Oral Word Reading Efficiency
DIBELS = Dynamic Indicators of Basic Early Literacy Skills
WRMT-R = Woodcock Reading Mastery Test–Revised

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

(continued)

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

3. The *Read Naturally*® group means for each of the five reading achievement outcomes reported by Kemp (2006) represent the posttest mean of the control group plus the mean difference calculated by the WWC using the difference-in-differences approach. The *Read Naturally*® group means for each of the three WRMT-R outcomes reported by Denton, Anthony, Parker, & Hasbrouck (2004) represent the adjusted posttest mean calculated using analysis of covariance results obtained by personal communication with the authors. These results were not included or referenced in Denton et al. (2004); those findings are based on the assumption that there were no problems with random assignment. However, students were moved between the treatment and comparison groups after the study began, and according to WWC standards, that resulted in this study being categorized as a quasi-experimental design rather than a randomized controlled trial. Baseline differences between the treatment and comparison groups were large enough to require statistical adjustment. Thus, the analysis of covariance results obtained from the author are presented.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The mean difference for each of the five reading achievement outcomes reported by Kemp (2006) reflects the mean difference between treatment and control groups calculated by the WWC using the difference-in-differences approach. The mean difference for each of the three WRMT-R outcomes reported by Denton, Anthony, Parker, & Hasbrouck (2004) reflects the adjusted mean difference between treatment and control groups obtained from analysis of covariate results provided to the WWC by personal communication with the authors.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. Effect sizes for the Kemp (2006) study are based on the difference-in-differences approach. Effect sizes for Denton et al. (2004) are based on results from a student-level ANCOVA.
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 favorable results for the intervention group.
8. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. For the *Read Naturally*® studies summarized here, no corrections for clustering were needed; however, multiple comparisons corrections were needed.
9. The WWC-computed average effect sizes for each study and for the domain across studies are simple averages rounded to two decimal places. The average improvement indices are calculated from the average effect sizes.

Appendix A3.2 Summary of study findings included in the rating for the English language development domain[1]

Outcome measure	Study sample	Sample size (students)	Author's findings from the study: Mean outcome (standard deviation)[2] — *Read Naturally*® group[3]	Author's findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[4] (*Read Naturally*® – comparison)	WWC calculations: Effect size[5]	WWC calculations: Statistical significance[6] (at $\alpha = 0.05$)	WWC calculations: Improvement index[7]
Kemp, 2006[8]								
Orthographic Choice test	3rd grade	39	13.17 (2.14)	12.84 (2.39)	0.33	0.14	ns	+6
MRT–Written	3rd grade	39	13.26 (3.03)	12.21 (2.51)	1.05	0.37	ns	+14
MRT–Oral/Written	3rd grade	39	13.13 (2.03)	12.68 (2.31)	0.45	0.20	ns	+8
Domain average for English language development[9]						**0.24**	**na**	**+9**

ns = not statistically significant
na = not applicable
MRT = Morphological Relatedness Test

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the English language development domain.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. The *Read Naturally*® group means for each of the three English language development outcomes reported by Kemp (2006) were calculated as the posttest mean of the control group plus the mean difference calculated by the WWC using the difference-in-differences approach.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. The mean difference for each of the three English language development outcomes reported by Kemp (2006) reflects the mean difference between treatment and control groups calculated by the WWC using the difference-in-differences approach.
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 favorable results for the intervention group.
8. The level of statistical significance was reported by the study authors or, when necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For the formulas the WWC used to calculate the statistical significance, see WWC Procedures and Standards Handbook, Appendix C for clustering and WWC Procedures and Standards Handbook, Appendix D for multiple comparisons. For the *Read Naturally*® study summarized here, no corrections for clustering were needed; however, multiple comparisons corrections were needed.
9. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A4.1 *Read Naturally*® rating for the reading achievement 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 reading achievement, the WWC rated *Read Naturally*® as having no discernible effects for English language learners.

Rating received

No discernible effects: No affirmative evidence of effects.

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

  Met. *Read Naturally*® has no studies showing statistically significant or substantively important effects for reading achievement.

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. *Read Naturally*® has no studies showing statistically significant positive effects in reading achievement.

AND

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

  Met. *Read Naturally*® has no studies showing statistically significant or substantively important negative effects in reading achievement.

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. *Read Naturally*® does not have any studies showing 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. *Read Naturally*® does not have any studies showing a statistically significant or substantively important negative effect, but there are two studies showing indeterminate effects, and none showing statistically significant or substantively important positive 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 study showing a statistically significant or substantively important effect, either positive or negative.

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.

(continued)

Appendix A4.1 *Read Naturally*® rating for the reading achievement domain *(continued)*

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 showing a statistically significant or substantively important effect.

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

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

  Not met. No studies showing a statistically significant or substantively important negative effect in reading achievement.

OR

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

  Not met. No studies showing a statistically significant or substantively important negative effect in reading achievement.

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 showing a statistically significant negative effect in reading achievement.

AND

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

  Met. No studies showing a statistically significant or substantively important positive effect in reading achievement.

Appendix A4.2 *Read Naturally*® rating for the English language development 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 English language development, the WWC rated *Read Naturally*® as having no discernible effects for English language learners.

Rating received

No discernible effects: No affirmative evidence of effects.

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

  Met. *Read Naturally*® has no studies showing statistically significant or substantively important effects for English language development.

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. *Read Naturally*® has no studies showing statistically significant positive effects in English language development.

AND

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

  Met. *Read Naturally*® has no studies showing statistically significant or substantively important negative effects in English language development.

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. *Read Naturally*® does not have any studies showing 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. *Read Naturally*® does not have any studies showing a statistically significant or substantively important negative effect, there is only one study showing indeterminate effects, and no studies show statistically significant or substantively important positive 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 study showing a statistically significant or substantively important effect, either positive or negative.

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.

(continued)

Appendix A4.2 *Read Naturally*® rating for the English language development domain *(continued)*

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 showing a statistically significant or substantively important effect.

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

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

Not met. No studies showing a statistically significant or substantively important negative effect in English language development.

OR

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

Not met. No studies showing a statistically significant or substantively important negative effect in English language development.

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 showing a statistically significant negative effect in English language development.

AND

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

Met. No studies showing a statistically significant or substantively important positive effect in English language development.

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
Reading achievement	2	8	99	Small
English language development	1	3	39	Small
Mathematics achievement	0	na	na	na

na = not applicable/not studied

1. A rating of "medium to large" requires at least two studies and two schools across studies in one domain and a total sample size across studies of at least 350 students or, assuming 25 students in a class, at least 14 classrooms across studies. Otherwise, the rating is "small." For more details on the extent of evidence categorization, see the WWC Procedures and Standards Handbook, Appendix G.