Appendix

Appendix A1 Study characteristics: Denton, Anthony, Parker, & Hasbrouck, 2004 (quasi-experimental design)

Characteristic	Description
Study citation	Denton, C. A., Anthony, J. A., Parker, R., and 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 report by Denton and colleagues covers two studies: one investigates the impacts of *Read Naturally* and the other investigates the impacts of *Read Well*. Ninety-three Hispanic ELL students (of which 45 were females) who were identified as having difficulty learning to read English participated in one of the two studies. All students who participated in both studies identified Spanish as their first language and were bilingual. The participants were in second through fifth grade (2nd = 22; 3rd = 37; 4th = 28; 5th = 6) and their ages ranged from 7 years to 12 years (average age = 9 years). The 63 students assigned to the *Read Naturally* study[1] were randomly assigned to either the treatment or control group. Three students in the control group were exposed to the treatment. These students were not reassigned to the treatment group. Instead, data from these students were removed from analysis.[2] Additionally, as requested by the participating schools, three students assigned to the control group were reassigned to the treatment group, and vice versa, one week after the study had begun. This renders the study a quasi-experimental design. The final sample consisted of 60 students (n = 32 in the treatment group and n = 28 in the control group).
Setting	The students attended one of five schools in a Central Texas school district. The district served 13,664 total students, 32% of whom were Hispanic and 56% of whom were identified as economically disadvantaged.
Intervention	The program occurred during pull-out tutoring sessions during the school day when the participants were not receiving their regular English instruction. Students involved with both programs (*Read Naturally* and *Read Well*) received an average of 22 tutoring sessions that were 40 minutes in length. The sessions consisted of repeated oral reading of connected text, vocabulary and comprehension instruction, and systematic monitoring of progress within 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 control group received the same regular English education curriculum as the treatment group but did not receive any additional tutoring.
Primary outcomes and measurement	The study measures in the reading achievement domain included a researcher-developed oral reading assessment[3] and three scales from the Woodcock Reading Mastery Tests-Revised: Word Identification, Word Attack, and Reading Comprehension (see Appendix A2 for more detailed descriptions of outcome measures).
Teacher training	Twenty-three undergraduate students studying special education who were enrolled in a class for teaching students with reading difficulties served as tutors. Tutors received training and were supervised by a graduate student experienced in *Read Naturally*.

1. Students were assigned to one of the two interventions, *Read Well* or *Read Naturally*, based on their pretest scores on the Word Attack subtest of the Woodcock Reading Master Tests-Revised (WRMT-R). Students who scored below the first-grade equivalency (< 1.0) were assigned to the *Read Well* study, and students whose grade equivalency score was higher than first grade (≥ 1.0) were assigned to the *Read Naturally* study.
2. Because data from three students in the comparison group were eliminated from the analysis and no data were eliminated from analysis in the treatment group, there was differential attrition (10% attrition in the comparison group and 0% attrition in the treatment group). The study did, however, demonstrate post-attrition equivalence.
3. Data from the researcher-developed oral reading assessment were not included in the study. Denton and colleagues (2004) stated that "logistical problems, some instances of potentially unreliable administration, and missing data points resulted in data that is invalid for analyses" (p. 296).

Appendix A2 Outcome measures in the reading achievement domain

Outcome measure	Description
Woodcock Reading Mastery Tests-Revised (WRMT-R): Word Identification subtest	The Word Identification subtest assesses basic reading skills by having participants read words presented in a list (as cited in Denton et al., 2004).
WRMT-R: Word Attack subtest	The Word Attack subtest assesses phonemic decoding by having participants read a list of nonsense words aloud (as cited in Denton et al., 2004).
WRMT-R: Passage Comprehension subtest	The Passage Comprehension subtest uses a cloze format that requires participants to read a passage that has an omitted word. The participants are asked to supply an appropriate word to complete the passage that they are reading (as cited in Denton et al., 2004).

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

Outcome measure	Study sample	Sample size (students)	Author's findings from the study: Mean outcome (standard deviation[2]) *Read Naturally* group	Author's findings from the study: Mean outcome (standard deviation[2]) Comparison group	WWC calculations: Mean difference[3] (*Read Naturally* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Denton et al., 2004 (quasi-experimental design)[7]								
WRMT-R: Word Identification	Grades 2–5	60	1.12 (11.64)	1.75 (9.65)	-0.63	-0.06	ns	0
WRMT-R: Word Attack	Grades 2–5	60	-0.22 (9.37)	0.97 (8.99)	-1.19	-0.13	ns	-5
WRMT-R: Passage Comprehension	Grades 2–5	60	2.13 (7.90)	0.71 (10.26)	1.42	0.16	ns	+6
Domain average[8] for reading achievement						-0.01	ns	0

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the improvement index.
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. Intervention and control group pre- to posttest change scores were used in the study authors' analyses and in the WWC calculations. 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, please 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 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.
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 *Read Naturally*, no corrections for clustering or multiple comparisons were needed.
8. This row provides the study average, which in this case 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 *Read Naturally* rating for the reading achievement domain

The WWC rates interventions 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.

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 WWC analysis found no statistically significant or substantively important effects in this domain.

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 study that met WWC evidence standards showed statistically significant positive effects. Further, there was only one study, and it did not meet WWC standards for a strong design.

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

  Met. The WWC analysis found no statistically significant or substantively important negative effects in this domain.

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 WWC analysis found no statistically significant or substantively important positive effects in this domain.

- 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. One study showed indeterminate effects and no study showed 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. The WWC analysis found no statistically significant or substantively important positive or negative effects in this domain.

- 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. The WWC analysis found no statistically significant or substantively important effects in this domain.

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

(continued)

Appendix A4 *Read Naturally* rating for the reading achievement 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. The WWC analysis found no statistically significant or substantively important negative effects in this domain.

- 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.

  Not met. The WWC analysis found no statistically significant or substantively important negative or positive effects in this domain.

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 is based on a strong design.

  Not met. No study that met WWC evidence standards showed statistically significant negative effects. Further, there was only one study, and it did not meet WWC standards for a strong design.

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

  Met. The WWC analysis found no statistically significant or substantively important positive effects in this domain.

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence[1]
		Schools	Students	
English language development	0	na	na	na
Mathematics achievement	0	na	na	na
Reading achievement	1	5	60	Small

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."