Appendix

Appendix A1 Study characteristics: Carlo, August, McLaughlin, Snow, Dressler, Lippman, Lively, & White, 2004 (randomized controlled trial with differential attrition)

Characteristic	Description
Study citation	Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., Lively, T. J., & White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39,* 188–215.
Participants	One-hundred forty-two English language learners[1] in the fifth grade participated in this study. Students were recruited from 16 classrooms in California, Virginia, and Massachusetts. Ninety-four English language learner students were in classrooms randomly assigned to the intervention group, and 48 students were in classrooms randomly assigned to the comparison group. Ninety percent (128 students) of the participants had pretest and posttest measures for at least one outcome. Follow-up contact with the first author revealed attrition in the comparison group; one classroom was not included in the analyses because a teacher left the study prior to intervention implementation, but after random assignment of classrooms to conditions (17 classrooms were originally assigned to conditions, but only 16 were in the analysis sample). In addition, some students in the overall sample received a pilot intervention in the fourth grade, and some did not. However, this intervention report focuses on fifth grade outcomes only.
Setting	The California site included classrooms from two schools that primarily served working class Mexican-American children in both bilingual and mainstream classes. The classrooms in Massachusetts were from a school that served working class, mostly Puerto Rican and Dominican students, within both bilingual and mainstream classes taught by bilingual teachers. The Virginia classrooms were recruited from an “English-medium” magnet school that served mainly working class Spanish speakers from the Caribbean and Central America.
Intervention	The intervention implemented in the study was adapted and published by the authors as the *Vocabulary Improvement Program for English Language Learners and Their Classmates (VIP).* Students read newspaper articles, diaries, documentaries, and historical and fictional accounts related to the topic of immigration. This 15-week intervention included 30–45 minutes of teaching four days a week and focused on 10–12 target words per week. On Mondays participants were given the weekly text to preview in Spanish. On Tuesdays the text was introduced in English, and target words in the text were discussed. On Wednesdays participants formed heterogeneous groups (based on English language proficiency) and completed two types of cloze activities. On Thursdays participants engaged in word association, synonym/antonym, and semantic feature analysis tasks. Then on Fridays either analysis of root words and derivation, or knowledge of multiple meanings of words was stressed. Three lessons were observed (during weeks 4, 9, and 13), revealing that six of the nine of the intervention group teachers implemented more than 70% of the key lesson elements, two 50%–60%, and one 35%.
Comparison	Students in the comparison group received their regular classroom instruction. The curriculum provided to the comparison group differed greatly across the schools in each region of the country. Teachers in the comparison group received some professional development in vocabulary teaching two years prior to the beginning of the intervention.
Primary outcomes and measurement	The study measured reading achievement using a researcher developed cloze measure. It measured English language development using measures titled Knowledge of Multiple Meanings of Words, Morphology, Word Mastery, Word Association, and the Peabody Picture Vocabulary Test-Revised (PPVT-R) (see Appendix A2 for more detailed descriptions of outcome measures). Assessments were given in the Fall and Spring of the academic year.
Teacher training	Researchers conducted biweekly Teacher Learning Community meetings with intervention group teachers, providing teachers with curriculum materials including detailed lesson plans, quasi-scripted lesson guides, overhead transparencies, worksheets, homework assignments, and all necessary reading materials. At these meetings, researchers facilitated discussions of practices that worked well in previous lessons and aspects of the curriculum that were problematic. The curriculum was not modified as a result of these meetings.

[1] Correspondence with the study authors revealed that they did not report treatment effects separately for English language learners and non-English language learners because their analyses did not show an interaction between treatment and language status. The WWC obtained English language learner subsample data from the authors for the purposes of this intervention report.

Appendix A2.1 Outcome measures in the reading achievement domain

Characteristic	Description
Cloze passages	For this researcher-developed measure, students read three stories with six cloze items per story. Each cloze item consists of a sentence with one word deleted; students are to supply the deleted word using contextual information to guide them. Ten of the 18 deleted words were taught during the intervention.

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

Characteristic	Description
Knowledge of multiple meanings of words (polysemy production)	For this researcher-developed cloze measure, students generate as many sentences as possible conveying the different meanings of words with multiple meanings (such as ring and place). Correct responses are scored based on the frequency of the response in the response pool, and each correct response receives one or more points. Common responses receive one point, intermediate responses two points, and infrequent responses three points.
Morphology	For this researcher-developed measure, students are to provide the base form of 27 derived words after being given the derived word and then hearing a sentence with the base form of the derived word omitted. The students are to write the correct form of the target word. Fewer than a third of the words were included in the intervention. Example item: The derived word is read to the participant ("discussion"). Then, the student is provided with a lean sentence context ("What did he want to ______?") and is asked to provide the word that fits into the sentence ("discuss"—the base word). Other examples of base word-derived pairs include remark-remarkable, nation-national, and migration-migrate.
Word mastery	For this researcher-developed measure, students are presented with 36 target words. Each of the target words is included in the curriculum and followed by four short definitions. Students must select the correct definition from the four definitions. All 36 words were taught during the intervention about two to three weeks before they were tested.
Word association measures (depth of word knowledge)	This task measures the depth of word knowledge by assessing students' knowledge of the relationship between words. Twenty target words each appear in the center of separate pages with six other words around the periphery of the pages. Students must draw a line from the target word to the three most closely related words printed on the periphery. Half of the target words were taught during the intervention.
Peabody Picture Vocabulary Test-Revised (PPVT-R)	For the PPVT-R, students are read a word and then must select the picture related to that word from the four pictures displayed. This is a widely used standardized test.

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

			Author's findings from the study					
			Mean outcome (standard deviation[2])		WWC calculations			
Outcome measure	Study sample	Sample size (classrooms)	*VIP* group	Comparison group	Mean difference[3] (*VIP* – comparison)	Effect size[4]	Statistical significance[5] (at $\alpha = 0.05$)	Improvement index[6]
Carlo et al., 2004 (randomized controlled trial with differential attrition)[7]								
Cloze passages	Grade 5	16	2.20 (3.74)	0.28 (4.01)	1.92	0.50	ns	+19
Domain average[8] for reading achievement						0.50	ns	+19

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 students' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
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, please see the 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 Carlo et al. (2004), a correction for clustering was needed for the finding in the reading achievement domain.
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 English language development domain[1]

			Author's findings from the study					
			Mean outcome (standard deviation[2])		WWC calculations			
Outcome measure	Study sample	Sample size (classromos)	*VIP* group	Comparison group	Mean difference[3] (*VIP* – comparison)	Effect size[4]	Statistical significance[5] (at $\alpha = 0.05$)	Improvement index[6]
Carlo et al., 2004 (randomized controlled trial with differential attrition)[7]								
Knowledge of multiple meanings of words (polysemy production)	Grade 5	16	2.38 (3.20)	0.60 (2.51)	1.78	0.59	ns	+22
Morphology	Grade 5	16	16.36 (29.05)	10.93 (30.20)	5.43	0.18	ns	+7
Word mastery	Grade 5	16	8.76 (6.78)	2.24 (5.15)	0.40	1.03	Statistically significant	+35
Word association measures	Grade 5	16	4.70 (6.75)	1.55 (7.74)	3.15	0.44	ns	+17
Peabody Picture Vocabulary Test-Revised (PPVT-R)	Grade 5	16	15.13 (21.54)	17.48 (20.86)	–2.35	–0.11	ns	–4
Domain average[8] for English language development						0.43	ns	+17

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 students' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
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, please see the 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 the Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Carlo et al. (2004), a correction for clustering was needed for findings in the English language development domain.
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 A4.1 *VIP* rating for the reading achievement domain

The WWC rates the effects of an intervention 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 *VIP* as potentially positive. It did not meet the criteria for *positive effects* because it had only one study. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *VIP* 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. *VIP* had one study that showed substantively important positive effects.

- 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 reviewed did not show any statistically significant or substantively important negative effects.

Other ratings considered

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

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

  Not met. The one study reviewed met WWC evidence standards with reservations and showed a substantively important, but not statistically significant positive effect.

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

  Met. The one study reviewed did not demonstrate any 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 *VIP* rating for the English language development domain

The WWC rates the effects of an intervention 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 *VIP* as potentially positive. It did not meet the criteria for *positive effects* because it had only one study. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *VIP* 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. *VIP* had one study that showed substantively important positive effects.

- 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 reviewed did not show any statistically significant or substantively important negative effects.

Other ratings considered

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

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

  Not met. The one study reviewed met WWC evidence standards with reservations and showed a substantively important, but not statistically significant positive effect.

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

  Met. The one study reviewed did not demonstrate any 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.