

Appendix

Appendix A1.1 Study characteristics: Jun-Aust, 1985 (randomized controlled trial)¹

Characteristic	Description
Study citation	Jun-Aust, H. (1985, March). <i>Individual differences in second language learning of Korean immigrant students</i> . Paper presented at the International Conference on Second/ Foreign Language Acquisition by Children, Oklahoma City, OK.
Participants	The study included 30 Korean English language learners in grades 1–6. ² All students participated in “pull-out” bilingual education conducted by English-speaking Korean teachers. Students who qualified for the study were identified as limited English proficient on the school district’s language proficiency test (Peabody Picture Vocabulary Test, PPVT) and on a reassessment of the PPVT just before the study began, scoring at or below the 20th percentile. All participating students were also recent immigrants to the United States (less than six months). Classes of students were randomly assigned into peer-pairing or non-peer-pairing conditions to avoid placing children from the same class in the intervention and comparison groups.
Setting	The study took place at two elementary schools located seven blocks apart in the Tacoma Public School District in Tacoma, Washington.
Intervention	The 14 Korean students in the intervention group participated in a 4.5-month peer-pairing program designed to increase social interaction, language development, and listening comprehension skills. When they started the program, the Korean students were asked to identify an English-speaking child from their classes with whom they would want to work. The chosen peers were then seated together by their classroom teachers, who asked the English-speaking peers to help the Korean students by explaining English to them, answering their questions, or being their friends.
Comparison	The 16 students in the comparison condition continued to participate in all regular classroom activities without the peer-pair program or teacher prompts to help peers learn English.
Primary outcomes and measurement	The primary outcomes were listening comprehension, oral language production, and actual classroom language behavior. Listening comprehension was measured by a researcher-developed assessment that required the student to listen to an audio tape of a monolingual English speaker and answer questions about daily tasks and Korean culture. Oral language production was assessed by asking students to tell stories in English about two pictures. Responses were audiotaped and scored according to a five-point rubric. Actual language behavior was evaluated with an event sampling classroom observation system that recorded when a target student was talking to or being addressed by a peer or the teacher.
Teacher training	Teachers attended a meeting that discussed second language learning and the purpose of using peer-pairs in the classroom and provided an operational definition of the concept. During the meeting teachers matched pairs according to the Korean student requests and created a new classroom seating chart for the pairs. Teachers were also instructed specifically to tell American peers to help their Korean peers to learn English by explaining to them, answering their questions, or just being friends (Jun-Aust, 1985, p. 14).

1. Jun-Aust (1985) examined the use of peer-pairing on student listening comprehension and English-language development. After students were assigned to a peer-pairing or a non-peer-pairing condition, students were rated by teachers and classroom peers as having low or high integrative motivation, or “the desire to be liked by others.” Jun-Aust presented posttest results by group (peer-pair condition vs. non-peer-pair condition) with high and low integrative motivation subpopulations in each group. The WWC pooled high and low integrative motivation subgroups within each condition to examine the effectiveness of overall peer-pairing versus non-peer pairing. Due to the report’s general focus on tutoring and peer-response groups, examining effects on high and low integrative subpopulations is beyond the scope of this report.
2. Minimal attrition occurred in this study. Thirty-three students qualified for participation. Two students from the peer-pairing group moved out of district, and one student from the comparison group moved out of district.

Appendix A1.2 Study characteristics: Prater & Bermudez, 1993 (randomized controlled trial)

Characteristic	Description
Study citation	Prater, D. L., & Bermudez, A. B. (1993). Using peer response groups with limited English proficient writers. <i>Bilingual Research Journal</i> , 17(1&2), 99–116.
Participants	The study included 46 English language learners in fourth grade who were randomly assigned to teachers and sections. Each teacher taught two sections, one randomly assigned to the peer-response intervention group and one to the comparison group. The intervention group included 27 students, of whom 25 were Hispanic, two were Asian-American, 16 were female, and 11 were male. The comparison group included 19 students, of whom 18 were Hispanic, one was Asian-American, 10 were female, and nine were male. Students ranged in age from 9 to 11 years old. All students had received English as a Second Language (ESL) or bilingual education services but were currently participating in general education fourth-grade classrooms. All students were considered by their teachers to have limited English proficiency that might put them at risk with respect to academic achievement.
Setting	The study took place at two elementary schools in the Houston, Texas, metropolitan area.
Intervention	Students participated in a four-week intervention that used small, mixed-ability peer response groups to provide feedback on group members' writing compositions. The 27 participating ELL students were randomly assigned to peer response groups consisting of four or five students. Peer response groups included both the ELL students participating in the study and students from the regular classroom. Generally, one or two ELL students were in each small group. During the first week, the teacher modeled how groups would work and demonstrated how students would respond to the writing of their peers. In the groups, the student author would read his or her composition, the group members would say what they liked about it, the student author would ask for help on a particular aspect, and the group members would suggest which parts of the composition to improve. During weeks two through four, students produced one composition a week. They met to select a topic, shared their first drafts, rewrote compositions based on group feedback, brought compositions to the group for final editing, incorporated changes, and wrote a final copy. For many of the peer group meetings, students assumed specific roles, with one student looking for errors in spelling, another for incomplete sentences, and another for capitalization and punctuation errors.
Comparison	Students in the comparison condition did individual composition writing (prewriting, drafting, revision, and editing) while students in the treatment condition participated in their peer response groups.
Primary outcomes and measurement	The primary outcome domain was written expression, which was assessed with a quality of composition score (holistic rubric score), total words written, total number of sentences written, and total number of idea units (single clauses) written. ¹
Teacher training	Information on teacher training was not provided.

1. According to Prater & Bermudez (1993), the purpose of the study was to expand English language development through student discourse and writing. Written expression was considered under the English language development domain in this study due to the language and discourse facilitated during the peer response writing groups.

Appendix A1.3 Study Characteristics: Serrano, 1987 (randomized controlled trial)

Characteristic	Description
Study citation	Serrano, C. J. (1987). The effectiveness of cross-level peer involvement in the acquisition of English as a second language by Spanish-speaking migrant children. <i>Dissertation Abstracts International</i> , 48(07), 1682A. (UMI No. 8723140)
Participants	The study included 42 English language learners in grades 3–5. ¹ These students were native Spanish-speaking and were children of Mexican and Mexican-American migrant workers who seasonally reside in Florida to pick citrus fruits. English language learners were administered a pretest, the IDEA Oral Language Proficiency Test I (K-6) (Ballard, Tighe, & Dalton, 1982, as cited by Serrano, 1987) and were divided into two levels of English language proficiency. Students at each level were randomly assigned to one of three groups. Overall, 12 students were assigned to the bilingual tutoring group, 13 students were assigned to the English-only tutoring group, and 17 students were assigned to the comparison group. The analytic sample for the first and second interventions is 29 and 30 students respectively.
Setting	The study took place at one elementary school in the School District of Indian River County, Florida.
Intervention	Students participated in a three-month tutoring program. Two versions of the program were examined: a tutoring group where the ELL tutee worked with a bilingual (somewhat proficient in both English and Spanish) student tutor and a tutoring group where the ELL tutee worked with an English-speaking tutor who did not speak Spanish. Students were assigned to their tutors based on age, sex, and grade level criteria. Tutoring included daily 20-minute sessions. A total of 37 sessions were implemented in the study for a total of 12.3 hours of tutoring. Tutoring focused on English language instruction and included lessons on life skills and every day tasks. For example, tutors introduced vocabulary, played a cassette tape that asked tutees to respond to directions and commands, and used a set of pictures to help ask comprehension questions. Each tutoring lesson focused on a life skill task (such as caring for a cut).
Comparison	Students in the comparison condition did not receive tutoring. The control group consisted of whole-group second language instruction led by the teacher.
Primary outcomes and measurement	The primary outcome was oral language proficiency as measured by the IDEA Oral Language Proficiency Test I (K-6) (Ballard, Tighe, & Dalton, 1982, as cited by Serrano, 1987). The test assesses syntax, comprehension, vocabulary, and verbal expression.
Teacher training	Student tutors participated in a series of 20-minute training sessions before tutoring began. Training content included explanations and demonstrations of effective second language teaching, modeling instructions, prompting, asking questions, and managing time and behavior. Role-playing was also included in training where the trainer played the role of the learner to help tutors practice tutoring skills.

1. The study began with 50 students. Minor attrition occurred, with eight students moving out of the district during the implementation of the study. Of the eight students, three left the bilingual tutor group, four left the English-only tutor group, and one left the comparison group.

Appendix A2 Outcome measures in the English language development domain

Outcome measure	Description
Listening comprehension	Listening comprehension was measured with an individually-administered, researcher-developed assessment that required a student to listen to an audio tape of a monolingual English speaker and answer questions about daily tasks and Korean culture (as cited in Jun-Aust, 1985).
Oral language production	Oral language production was assessed by asking students to tell stories in English about two pictures. Responses were audiotaped and scored according to a five-point rubric (as cited in Jun-Aust, 1985).
Language behavior	Actual language behavior was evaluated based on an event time sampling classroom observation system that recorded when a target student was talking to or being addressed by a peer or the teacher. The language behaviors were charted at 10-second intervals during four 3-minute observations: two observations during classroom instruction and two observations during recess (as cited by Jun-Aust, 1985).
Composition quality	A six-point holistic scoring guide was used to determine overall English writing quality. Each composition was scored by two independent readers. Scores that diverged more than one point were read by a third reader who assigned a final score. Cohen's Kappa was calculated on the unarbitrated scores and yielded a reliability coefficient of 0.94 on the pretest and 0.92 on the posttest (as cited in Prater & Bermudez, 1993).
Total words written	The number of total words in a composition (as cited in Prater & Bermudez, 1993).
Total sentences written	The number of total sentences in a composition (as cited in Prater & Bermudez, 1993).
Total idea units written	The number of total independent or dependent single clauses in a composition (as cited in Prater & Bermudez, 1993).
IDEA Oral Language Proficiency Test (IPT I)	A standardized measure of oral language proficiency in syntax, comprehension, vocabulary, and verbal expression. Verbal and visual stimuli are presented to the student to elicit speech which is then assessed for correctness, appropriateness, and completeness (as cited in Serrano, 1987a,b).

Appendix A3 Summary of study findings included in the rating for the English language development domain¹

Outcome measure	Study sample	Sample size (schools/students)	Authors' findings from the study			WWC calculations			
			Mean outcome (standard deviation ²)		Mean difference ³ (Peer Tutoring – comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶	
			Peer Tutoring group	Comparison group					
Jun-Aust, 1985 (randomized controlled trial)⁷									
Listening comprehension	Grades 1–6	2/30	9.00 (2.90)	7.70 (1.90)	1.30	0.51	ns	+19	
Oral language production	Grades 1–6	2/30	20.8 (5.80)	17.8 (6.30)	3.00	0.48	ns	+19	
Language behavior – talking to peer	Grades 1–6	2/30	14.00 (6.99)	5.35 (5.29)	8.65	1.34	Statistically significant	+41	
Language behavior – addressed from subject to peer	Grades 1–6	2/30	11.60 (4.87)	2.90 (2.41)	8.70	2.16	Statistically significant	+48	
Language behavior – talking to teacher	Grades 1–6	2/30	1.05 (1.33)	0.90 (2.07)	0.15	0.09	ns	+3	
Language behavior – addressed from teacher to subject	Grades 1–6	2/30	0.45 (0.78)	0.50 (1.84)	0.05	0.04	ns	+1	
Average⁸ for English language development (Jun-Aust, 1985)						0.77	Statistically significant	+22	
Prater & Bermudez, 1993 (randomized controlled trial)⁷									
Composition quality	Grade 4	2/46	2.33 (1.01)	2.16 (1.26)	0.17	0.15	ns	+6	
Total words written	Grade 4	2/46	100.22 (50.52)	70.37 (42.63)	29.85	0.62	ns	+23	
Total sentences written	Grade 4	2/46	8.52 (6.07)	6.68 (4.51)	1.84	0.33	ns	+13	
Total idea units written	Grade 4	2/46	15.93 (8.32)	9.89 (7.81)	6.04	0.73	Statistically significant	+27	
Average⁸ for English language development (Prater & Bermudez, 1993)						0.46	ns	+17	
Serrano, 1987 (randomized controlled trial)^{7,9}									
IDEA Oral Language Proficiency Test (IPT I)	Grades 3–5 with bilingual tutors ¹⁰	1/29	14.20 (22.40)	11.30 (15.20)	–2.90	–0.16	ns	+7	

(continued)

Appendix A3 Summary of study findings included in the rating for the English language development domain¹ (continued)

Outcome measure	Study sample	Sample size (schools/students)	Authors' findings from the study					
			Mean outcome (standard deviation ²)		WWC calculations			
			Peer Tutoring group	Comparison group	Mean difference ³ (Peer Tutoring – comparison)	Effect size ⁴	Statistical significance ⁵ (at $\alpha = 0.05$)	Improvement index ⁶
IDEA Oral Language Proficiency Test (IPT I)	Grades 3–5 with English-only tutors ¹⁰	1/30	12.20 (19.30)	11.30 (15.20)	0.90	0.05	ns	+2
Average⁸ for English language development (Serrano, 1987)						0.11	ns	+5
Domain average⁸ for language development across all studies						0.56	na	+17

ns = not statistically significant

na = not applicable

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices.
2. The intervention group mean equals the comparison group mean plus the mean difference. 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. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, see [Technical Details of WWC-Conducted Computations](#).
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition versus the percentile rank of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
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 Jun-Aust (1985) and Prater & Bermudez (1993), corrections for clustering and multiple comparisons were needed. Corrections for clustering and multiple comparisons did not change reported statistical significance for Jun-Aust (1985). Corrections for multiple comparisons did change Prater & Bermudez (1993) outcomes for total words written from statistically significant to non-significant.
8. 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 size. Domain averages are calculated as the average of study averages.
9. Intervention and control group pretest to posttest change scores were used in the WWC calculations.
10. WWC viewed the bilingual tutoring and English-only tutoring as two separate outcomes rather than two different interventions because the tutoring intervention by both bilingual and English-only tutors was not substantially different.

Appendix A4 *Peer Tutoring and Response Groups* 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 *Peer Tutoring and Response Groups* as having positive effects. The remaining ratings (potentially positive effects, mixed effects, no discernible effects, potentially negative effects, negative effects) were not considered because *Peer Tutoring and Response Groups* was assigned the highest applicable rating.

Rating received

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.

Met. Two of the three studies reviewed in this domain showed statistically significant positive effects. Both studies met WWC evidence standards for a strong design.

AND

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

Met. None of the studies reviewed 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. See the [WWC Intervention Rating Scheme](#) for a complete description.

Appendix A5 Extent of evidence by domain

Outcome domain	Number of studies	Sample size		Extent of evidence ¹
		Schools	Students	
Reading achievement	0	0	0	na
Mathematics achievement	0	0	0	na
English language development	3	5	118	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.”