Appendix

Appendix A1.1 Study characteristics: Stevens & Slavin, 1995

Characteristic	Description
Study citation	Stevens, R. J., & Slavin, R. E. (1995). The Cooperative Elementary School: Effects on students' achievement, attitudes and social relations. *American Educational Research Journal, 32*(2), 321–351.
Participants	This study is a quasi-experiment conducted in five schools. Two treatment schools were selected by the investigators to implement the intervention, and three schools, matched on academic achievement, ethnicity, and socioeconomic background, were selected to serve as comparison schools. Classes in grades 2 through 6 in the treatment schools were matched with classes in the comparison schools based on pretest scores on the California Achievement Test (Reading, Language Arts, and Mathematics). The study's analytic sample included 411 students in 21 treatment classrooms and 462 students in 24 comparison classrooms. The study reported students' outcomes after two years of program implementation; these findings were used in the intervention ratings and can be found in Appendices A3.1 and A3.2. Additional findings reflecting students' outcomes after one year of program implementation can be found in Appendices A4.1 and A4.2
Setting	The study was conducted in five schools in one suburban school district in Maryland. The student populations of each school ranged from 4% to 15% minority students, and from 2% to 20% students received free or reduced-price lunch.
Intervention	Intervention schools implemented the Cooperative Elementary School model, a whole-school reform model that uses cooperative learning strategies across multiple content areas. Teachers used peer coaching and conducted their planning in a cooperative manner. Cooperative Elementary School emphasizes teacher involvement in site-based management and parent involvement in schools. The language arts/reading curriculum within Cooperative Elementary School is *Cooperative Integrated Reading and Composition*®. Daily lessons, which focus on story-related activities, direct instruction in reading comprehension, and integrated reading and language arts activities, incorporate team practice, peer assessment, and team/partner recognition. This program was phased in gradually during the first year of the two-year implementation.
Comparison	Reading activities consisted of students working in small reading groups using a basal series, workbooks, worksheets, and activities based on teacher-prepared materials. Language arts activities generally involved whole-class instruction using a published language arts series, as well as teacher-developed activities. Comparison schools did not use structured cooperative learning during classroom instruction, although occasional cooperative activities were used by some of the teachers. Comparison schools implemented some of the components of the Cooperative Elementary School model, but they did not implement *Cooperative Integrated Reading and Composition*®.
Primary outcomes and measurement	For both the pretest and the posttest, students took the Vocabulary, Reading Comprehension, Language Expression, and Language Mechanics subtests of the California Achievement Test. Scores were converted to z-scores in order to conduct analyses across the grades included in the study sample (grades 2–6). For a more detailed description of these outcome measures, see Appendices A2.1 and A2.2.
Staff/teacher training	Intervention teachers were trained in *Cooperative Integrated Reading and Composition*® prior to implementation. Subsequent trainings were conducted at two-month intervals during the school year. Trainers reviewed with treatment teachers a detailed manual that explains how to implement the program in the classroom. Trainers also provided simulated demonstrations of lessons. In addition, during the school year, members of the research staff observed treatment-group classes, participated in meetings with treatment-group teachers, and observed steering committee meetings in order to facilitate implementation of the program components.

Appendix A1.2 Study characteristics: Jewell, 1994

Characteristic	Description
Study citation	Jewell, M. E. (1994). The effect of classroom-based follow-up assistance on mainstream reading and language arts instruction (Doctoral dissertation, University of Washington, 1994). *Dissertation Abstracts International, 55*(11A), 107–3473.
Participants	This study is a quasi-experiment that initially included a sample of 51 second- to sixth-grade classrooms assigned to one of three conditions: (1) comparison; (2) treatment, receiving *Cooperative Integrated Reading and Composition*® training in the summer preceding implementation; and (3) treatment, receiving the program training as well as follow-up support during the school year.[1] The treatment classrooms were matched with comparison classrooms on the Gates–MacGinitie pretest scores. This review focuses on comparisons of the 15 classrooms taught by teachers who received either program training or program training with follow-up support, and the 15 classrooms in the comparison group.
Setting	The study took place in four schools in one district in the United States. The participating elementary schools served 9% to 27% minority students, and less than 15% of the student population received special education services.
Intervention	There were two forms of the *Cooperative Integrated Reading and Composition*® intervention: (1) training only and (2) training plus follow-up support. The intervention group participated in teacher-led basal-related activities, partner reading, story-related writing, reading of words aloud, word meaning activities, story retelling, spelling, direct instruction in reading comprehension, home reading, integrated language arts and writing, weekly tests, and cooperative learning groups of four students. The program was implemented in intervention classrooms for seven to eight months.
Comparison	Comparison group teachers continued to teach in accordance with their own style and used the regular district-adopted reading and language arts program (basal materials). All comparison schools used the same reading and language arts curricula: *Houghton Mifflin Reading* (Durr et al., 1989)[2] and the *Silver Burdett & Ginn English* series (Ragno, Toth, & Gray, 1988).[3]
Primary outcomes and measurement	For both the pretest and the posttest, students took the Gates–MacGinitie Reading Test Comprehension and Vocabulary subtests and the Basic Academic Skills Sample Reading Proficiency subtest. For a more detailed description of these outcome measures, see Appendices A2.1 and A2.2.
Staff/teacher training	The study's investigator provided *Cooperative Integrated Reading and Composition*® training and follow-up assistance. The training took place during five 4-hour sessions (spanning one week) during the summer preceding the study. Teachers were provided with lesson plans that were aligned to the district's basal series curriculum. Teachers practiced the *Cooperative Integrated Reading and Composition*® components and received feedback from the investigator and peers. All *Cooperative Integrated Reading and Composition*®–trained teachers participated in two follow-up meetings during the school year. Teachers assigned to receive classroom-based follow-up assistance also had the investigator observe lessons (on average, about 10 observations), provide feedback, demonstrate teaching procedures, and make recommendations for future lessons.

1. Results from these analyses are not included in this report because the treatment groups and the comparison group were not equivalent at baseline.
2. Durr, W. K., Pikulski, J. J., Bean, R. M., Cooper, J. D., Glaser, N. A., Greenlaw, M. J., & Schoephoerster, H. (1989). *Houghton Mifflin Reading*. Boston: Houghton Mifflin.
3. Ragno, N. N., Toth, M. D., & Gray, B. G. (1988). *Silver Burdett & Ginn English*. Morristown, NJ: Silver Burdett & Ginn.

Appendix A2.1 Outcome measures for the comprehension domain

Outcome measure	Description
Reading comprehension construct	
California Achievement Test (CAT)—Reading Comprehension subtest	The CAT is a norm- and criterion-referenced annual test. The Reading Comprehension subtest is administered to grades 1 through 12 and focuses on students' use of reading comprehension strategies. Passages reflect a wide range of narrative, expository, contemporary, and traditional texts. The test measures information recall, meaning construction, form analysis, and meaning evaluation of seven different selections (as cited in Stevens & Slavin, 1995).
Gates–MacGinitie Reading Tests—Comprehension subtest	The Comprehension subtest of the Gates–MacGinitie Reading Test measures each student's ability to read and understand different types of prose. The test contains 11 passages of various lengths and subjects and 48 questions (as cited in Jewell, 1994).
Vocabulary development construct	
California Achievement Test (CAT)—Vocabulary subtest	The CAT is a norm- and criterion-referenced annual test. The Vocabulary subtest contains 20 items measuring same-meaning, opposite-meaning words; multi-meaning words; words in context; and the meaning of affixes (as cited in Stevens & Slavin, 1995).
Gates–MacGinitie Reading Tests—Vocabulary subtest	The Vocabulary subtest of the Gates–MacGinitie Reading Test measures a student's reading vocabulary. The test contains 45 questions that measure word knowledge by asking students to choose one word or phrase that means most nearly the same as a presented word. The vocabulary test takes 20 minutes to administer (as cited in Jewell, 1994).

Appendix A2.2 Outcome measures for the general literacy achievement domain

Outcome measure	Description
California Achievement Test (CAT)—Language Mechanics subtest	The Language Mechanics and Language Expression subtests of the CAT work together to measure a broad range of language and writing skills, including the ability to apply standard usage and writing conventions and to develop effective sentences and paragraphs. The Language Mechanics subtest contains 20 items that measure skills in the mechanics of capitalization and punctuation. Editing skills are measured in the context of passages presented in various formats (for grades 4–12) (as cited in Stevens & Slavin, 1995).
California Achievement Test (CAT)—Language Expression subtest	The Language Mechanics and Language Expression subtests of the CAT work together to measure a broad range of language and writing skills, including the ability to apply standard usage and writing conventions and to develop effective sentences and paragraphs. The Language Expression subtest contains 20 items that measure skills in language usage and sentence structure. The items measure skills in the use of various parts of speech, formation and organization of sentences and paragraphs, and writing for clarity (for grades 4–12) (as cited in Stevens & Slavin, 1995).
Basic Academic Skills Sample—Reading Proficiency subtest	The Basic Academic Skills Sample is a group-administered, curriculum-based assessment of students' reading proficiency, measured via maze procedure using three passages averaging 263 words. In this maze procedure, every seventh word in the passage is replaced by a multiple choice question item containing the correct word and two distractors. Students' scores reflect the number of correct multiple choice selections, given one minute per passage (as cited in Jewell, 1994).

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

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Cooperative Integrated Reading and Composition®* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Cooperative Integrated Reading and Composition®* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Stevens & Slavin, 1995[7,8]								
CAT—Reading Vocabulary	Grades 2–6	45/873	0.10 (0.98)	–0.11 (1.01)	0.21	0.21	Statistically significant	+8
CAT—Reading Comprehension	Grades 2–6	45/873	0.15 (0.98)	–0.13 (1.01)	0.28	0.28	Statistically significant	+11
Average for comprehension (Stevens & Slavin, 1995)[9]						**0.25**	**Statistically significant**	**+10**
Jewell, 1994[7,10]								
GMRT—Vocabulary subtest	Grades 2–6	30 classrooms	50.48 (21.06)	49.07 (21.06)	1.41	0.07	ns	+3
GMRT—Comprehension subtest	Grades 2–6	30 classrooms	52.20 (21.06)	49.16 (21.06)	3.04	0.14	ns	+6
Average for comprehension (Jewell, 1994)[9]						**0.11**	**ns**	**+4**
Domain average for comprehension across all studies[9]						**0.18**	**na**	**+7**

ns = not statistically significant
na = not applicable
CAT = California Achievement Test
GMRT = Gates–MacGinitie Reading Test

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the comprehension 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. 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 WWC Procedures and Standards Handbook, Appendix B.
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 for the intervention group.

(continued)

Appendix A3.1 Summary of study findings included in the rating for the comprehension domain *(continued)*

7. 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. In the case of Stevens and Slavin (1995), correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study. A correction for clustering was not needed, as the authors used HLM analyses, which accounted for multi-level data (of students nested within classrooms and schools). In the case of Jewell (1994), no corrections for clustering or multiple comparisons were needed, as analyses were performed at the classroom level and findings were not statistically significant.
8. The intervention group and comparison group mean outcome values for Stevens and Slavin (1995) are the HLM-fitted two-year posttest means.
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.
10. The group mean values reported for Jewell (1994) are the pretest group means plus the gain scores. Standard deviations of 21.06 are from the normative student sample.

Appendix A3.2 Summary of study findings included in the rating for the general literacy achievement domain[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Cooperative Integrated Reading and Composition®* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Cooperative Integrated Reading and Composition®* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Stevens & Slavin, 1995[7,8]								
CAT—Language Mechanics subtest	Grades 2–6	45/873	0.05 (0.97)	–0.05 (1.02)	0.10	0.10	ns	+4
CAT—Language Expression subtest	Grades 2–6	45/873	0.11 (0.96)	–0.10 (1.03)	0.21	0.21	Statistically significant	+8
Average for general literacy achievement (Stevens & Slavin, 1995)[9]						**0.16**	**ns**	**+6**
Jewell, 1994[7,10]								
BASS—Reading Proficiency subtest	Grades 2–6	30 classrooms	100.72 (15.00)	101.86 (15.00)	–1.14	–0.08	ns	–3
Average for general literacy achievement (Jewell, 1994)[9]						**–0.08**	**ns**	**–3**
Domain average for general literacy achievement across all studies[9]						**0.04**	**na**	**+2**

ns = not statistically significant
na = not applicable
CAT = California Achievement Test
BASS = Basic Academic Skills Sample

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the general literacy 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.
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 WWC Procedures and Standards Handbook, Appendix B.
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 for the intervention group.
7. 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. In the case of Stevens and Slavin (1995), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study. A correction for clustering was not needed, as the authors used HLM analyses, which accounted for multi-level data (of students nested within classrooms and schools). In the case of Jewell (1994), no corrections for clustering or multiple comparisons were needed.
8. The intervention group and comparison group mean outcome values for Stevens and Slavin (1995) are the HLM-fitted two-year posttest means.
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.
10. The group mean values reported for Jewell (1994) are the pretest group means plus the gain scores. Standard deviations of 15.00 are from the normative student sample.

Appendix A4.1 Summary of one-year implementation findings for the comprehension domain[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Cooperative Integrated Reading and Composition®* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Cooperative Integrated Reading and Composition®* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Stevens & Slavin, 1995[7,8]								
CAT—Reading Vocabulary	Grades 2–6	45/873	0.08 (0.99)	–0.09 (1.01)	0.17	0.17	Statistically significant	+7
CAT—Reading Comprehension	Grades 2–6	45/873	0.08 (1.01)	–0.05 (0.99)	0.13	0.13	ns	+5

ns = not statistically significant

CAT = California Achievement Test

1. This appendix presents one-year findings for measures that fall in the comprehension domain. Two-year findings were used for rating purposes and are presented in Appendix A3.1
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. 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 WWC Procedures and Standards Handbook, Appendix B.
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 for the intervention group.
7. 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. In the case of Stevens and Slavin (1995), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study. A correction for clustering was not needed, as the authors used HLM analyses, which accounted for multi-level data (of students nested within classrooms and schools).
8. The intervention group and comparison group mean outcome values for Stevens and Slavin (1995) are the HLM-fitted one-year posttest means.

Appendix A4.2 Summary of one-year implementation findings for the general literacy achievement domain[1]

Outcome measure	Study sample	Sample size (classrooms/ students)	Authors' findings from the study: Mean outcome (standard deviation)[2] — *Cooperative Integrated Reading and Composition®* group	Authors' findings from the study: Mean outcome (standard deviation)[2] — Comparison group	WWC calculations: Mean difference[3] (*Cooperative Integrated Reading and Composition®* – comparison)	WWC calculations: Effect size[4]	WWC calculations: Statistical significance[5] (at $\alpha = 0.05$)	WWC calculations: Improvement index[6]
Stevens & Slavin, 1995[7,8]								
CAT—Language Mechanics subtest	Grades 2–6	45/873	0.00 (0.99)	0.01 (1.00)	–0.01	–0.01	ns	–0.4
CAT—Language Expression subtest	Grades 2–6	45/873	0.04 (1.01)	–0.04 (0.99)	0.08	0.08	ns	+3

ns = not statistically significant

CAT = California Achievement Test

1. This appendix presents one-year findings for measures that fall in the general literacy achievement domain. Two-year findings were used for rating purposes and are presented in Appendix A3.2.
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. 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 WWC Procedures and Standards Handbook, Appendix B.
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 for the intervention group.
7. 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. In the case of Stevens and Slavin (1995), a correction for multiple comparisons was needed, so the significance levels may differ from those reported in the original study. A correction for clustering was not needed, as the authors used HLM analyses, which accounted for multi-level data (of students nested within classrooms and schools).
8. The intervention group and comparison group mean outcome values for Stevens and Slavin (1995) are the HLM-fitted one-year posttest means.

Appendix A5.1 *Cooperative Integrated Reading and Composition®* 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 *Cooperative Integrated Reading and Composition®* as having potentially positive effects for adolescent learners. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, or negative effects) were not considered, as *Cooperative Integrated Reading and Composition®* 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. One study showed a statistically significant 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 study 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. One study showed a statistically significant positive effect.

AND

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

  Met. No study showed a statistically significant or substantively important negative effect.

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 *Cooperative Integrated Reading and Composition*® 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 *Cooperative Integrated Reading and Composition*® as having potentially positive effects for adolescent learners. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, or negative effects) were not considered, as *Cooperative Integrated Reading and Composition*® 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. One study showed a statistically significant 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 study 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. One study showed a statistically significant positive effect.

AND

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

  Met. No study showed a statistically significant or substantively important negative effect.

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	na	na	na	na
Reading fluency	na	na	na	na
Comprehension	2	9	1,460[2]	Medium to large
General literacy achievement	2	9	1,460[2]	Medium to large

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 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.
2. This number is an estimate, as the exact number of students is not available for Jewell (1994).