

Appendix

Appendix A1.1 Study characteristics: PCER Consortium, 2008 (randomized controlled trial)

Characteristic	Description
Study citation	Preschool Curriculum Evaluation Research (PCER) Consortium, (2008). <i>Doors to Discovery</i> and <i>Let's Begin with the Letter People</i> . In <i>Effects of preschool curriculum projects on school readiness</i> (pp. 85–98). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
Participants	The study, conducted during the 2003–2004 and 2004–2005 school years, included three groups: <i>Doors to Discovery</i> TM , <i>Let's Begin with the Letter People</i> [®] , and a control group. Nineteen full-day Head Start and public prekindergarten preschools were recruited for the study. From these 19 preschools, 95 teachers/classrooms were recruited, of which 76 were included in random assignment. The manuscript notes, and the authors confirmed, that the researchers randomly assigned the classrooms to three conditions (<i>Doors to Discovery</i> TM , <i>Let's Begin with the Letter People</i> [®] , and control); however, all classrooms within a preschool were assigned to the same condition. The resulting sample of teachers/classrooms included 25 <i>Doors to Discovery</i> TM classrooms, 24 <i>Let's Begin with the Letter People</i> [®] classrooms, and 27 control classrooms. Forty-five of the 76 classrooms were then randomly selected to participate in the PCER study. One of the 45 classrooms dropped out, leaving 14 <i>Doors to Discovery</i> TM classrooms, 15 <i>Let's Begin with the Letter People</i> [®] classrooms, and 15 control classrooms. Seven children whose parents had provided consent to participate in the study were randomly selected from each classroom, for a total of 308 children. ¹ The parental consent rate was 65% for the treatment group (combined <i>Doors to Discovery</i> TM and <i>Let's Begin with the Letter People</i> [®]) and 55% for the control group. The total number of participating children in the study at baseline was 297 (101 <i>Doors to Discovery</i> TM , 100 <i>Let's Begin with the Letter People</i> [®] , and 96 control). At baseline, children in the study averaged 4.6 years of age; 55% were male; and 43% were Hispanic, 30% were Caucasian, and 13% were African-American. The analysis sample for the <i>Doors to Discovery</i> TM study included 183 children (94 <i>Doors to Discovery</i> TM and 89 control). Depending on the outcome, child-level attrition ranged from 7% to 10%.
Setting	The <i>Doors to Discovery</i> TM study was conducted with children from 29 full-day preschool classrooms (14 <i>Doors to Discovery</i> TM and 15 control) selected from Head Start and public prekindergarten programs in Texas.
Intervention	<i>Doors to Discovery</i> TM is a prekindergarten curriculum that promotes learning in five areas associated with early literacy success: oral language, phonological awareness, concepts of print, alphabet knowledge and writing, and comprehension. Eight thematic units cover topics such as nature, friendship, communities, society, and health. Activities include teacher-directed, large- and small-group, and independent practice through activities tied to the curriculum. Family learning activities are also available. In the PCER study, each classroom's fidelity to the curriculum was rated on a four-point scale, ranging from “not at all” (0) to “high” (3). The average score for the <i>Doors to Discovery</i> TM classrooms was 2.13 on the measure.
Comparison	Control teachers used teacher-developed nonspecific curricula. Their classrooms were rated with the same fidelity measure used in the <i>Doors to Discovery</i> TM classrooms, which ranged from 0 to 3. The average score for the control classrooms was 1.0.
Primary outcomes and measurement	The outcome domains assessed were children's oral language, print knowledge, phonological processing, and math. Oral language was assessed with the Peabody Picture Vocabulary Test-III (PPVT-III) and the Test of Language Development-Primary III (TOLD-P:3) Grammatical Understanding subtest. Print knowledge was assessed with the Test of Early Reading Ability-III (TERA-3), the Woodcock-Johnson III (WJ-III) Letter-Word Identification subtest, and the WJ-III Spelling subtest. Phonological processing was assessed with the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP) Elision subtest. Math was assessed with the WJ-III Applied Problems subtest, the Child Math Assessment-Abbreviated (CMA-A), and the Shape Composition task. For a more detailed description of these outcome measures, see Appendices A2.1–2.4.
Staff/teacher training	Teachers received curriculum training prior to the start of the 2003–2004 school year. This was the second year of implementation of the treatment, and most of the teachers had been trained prior to the start of the 2002–2003 school year. New teachers each received 12 hours of training, and returning teachers each received six hours of training. The research team collected site-specific curriculum fidelity data three times during the preschool year. All classrooms were observed using the Teacher Behavior Rating Scale in fall and spring of the preschool year.

1. PCER Consortium (2008, p. 88) reported that eight children were selected from each classroom. In response to a query, the study authors noted that eight children were randomly selected for the site-specific study; however, only seven children were randomly selected for the PCER Consortium study.

Appendix A1.2 Study characteristics: Christie et al., 2003 (randomized controlled trial)

Characteristic	Description
Study citation	Christie, J., Roskos, K., Vukelich, C., & Han, M. (2003, June). The effects of a well-designed literacy program on young children's language and literacy development. In F. Lamb-Parker, J. Hagen, R. Robinson, & H. Rhee. (Eds.), <i>The first eight years—pathways to the future: Implications for research, policy, and practice</i> (pp. 447–448). Proceedings of the Head Start National Research Conference. New York: Mailman School of Public Health, Columbia University.
Participants	In this study, four Head Start classrooms—two serving English-speaking children and two serving Spanish-speaking children—were blocked on primary language of the children and randomly assigned to implement either <i>Doors to Discovery</i> TM or the <i>Creative Curriculum</i> [®] . One additional classroom served a mixed-language group and was assigned to implement <i>Doors to Discovery</i> TM . Since this classroom was not assigned at random, it was omitted from WWC analyses. At baseline, the four-classroom study included 35 children in the <i>Doors to Discovery</i> TM group and 28 children in the control group. The four-classroom analysis sample was substantially smaller, containing 21 children in the <i>Doors to Discovery</i> TM group and 16 children in the control group. This translates to a child-level attrition rate of 41%. Baseline differences between the treatment and control groups were large, but not statistically significant. For the analytic sample, the baseline difference was (in standard deviation units) 0.40 for the Peabody Picture Vocabulary Test (PPVT), 0.45 for Get Ready to Read!, and –0.29 for Concepts of Print.
Setting	The study was conducted with Head Start classrooms in a large metropolitan area in the southwest United States.
Intervention	Teachers in the intervention classrooms used three units from the <i>Doors to Discovery</i> TM curriculum: Vroom! Vroom!; Build It Big!; and Tabby Tiger's Diner. Each unit was taught for 4 weeks.
Comparison	The control classrooms used the existing curriculum, which the authors described as loosely based on the <i>Creative Curriculum</i> [®] .
Primary outcomes and measurement	The outcomes assessed were children's oral language and print knowledge. Oral language was assessed with the PPVT. Print knowledge was assessed with Get Ready to Read! and Concepts of Print. All assessments were conducted in English (J. Christie, personal communication, January 23, 2009). For a more detailed description of these outcome measures, see Appendices A2.1–2.2.
Staff/teacher training	No information on training was provided.

Appendix A2.1 Outcome measures for the oral language domain

Outcome measure	Description
Peabody Picture Vocabulary Test-3rd Edition (PPVT-III)	A standardized measure of children's receptive vocabulary where children show understanding of a spoken word by pointing to a picture that best represents the meaning (as cited in PCER Consortium, 2008).
Test of Language Development-Primary III (TOLD-P:3) Grammatical Understanding subtest	A standardized measure of children's ability to comprehend the meaning of sentences by selecting pictures that most accurately represent the sentence (as cited in PCER Consortium, 2008).

Appendix A2.2 Outcome measures for the print knowledge domain

Outcome measure	Description
Test of Early Reading Ability-III (TERA-3)	A standardized measure of children's developing reading skills with three subtests: alphabet, conventions, and meaning (as cited in PCER Consortium, 2008). ¹
Woodcock-Johnson III Letter-Word Identification subtest	A standardized measure of identification of letters and reading of words (as cited in PCER Consortium, 2008).
Woodcock-Johnson III Spelling subtest	A standardized measure that assesses children's prewriting skills, such as drawing lines, tracing, and writing letters (as cited in PCER Consortium, 2008).
Concepts of Print	An eight-item measure of concepts of print, adapted from the Developing Skills Checklist, which assesses children's knowledge of book handling; the difference between print and pictures; the concepts of "letter", "word", and "number"; and several conventions of print, e.g., left-right sequence and capitalization (J. Christie, personal communication, January 23, 2009).
Get Ready to Read!	An early literacy screening tool that measures print recognition, concepts of print, book concepts, and phonemic awareness (J. Christie, personal communication, January 23, 2009).

1. By name, this measure sounds like it should be captured under the early reading and writing domain; however, the description of the measure identifies constructs that are pertinent to print knowledge, such as knowing the alphabet, understanding print conventions, and environmental print.

Appendix A2.3 Outcome measures for the phonological processing domain

Outcome measure	Description
Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP) Elision subtest	A measure of children's ability to identify and manipulate sounds in spoken words, using word prompts and picture plates for the first nine items and word prompts only for later items (as cited in PCER Consortium, 2008).

Appendix A2.4 Outcome measures for the math domain

Outcome measure	Description
Woodcock-Johnson III Applied Problems subtest	A standardized measure of children's ability to solve numerical and spatial problems, presented verbally with accompanying pictures of objects (as cited in PCER Consortium, 2008).
Child Math Assessment-Abbreviated (CMA-A) Composite Score	The average of four subscales: (1) solving addition and subtraction problems using visible objects, (2) constructing a set of objects equal in number to a given set, (3) recognizing shapes, and (4) copying a pattern using objects that vary in color and identity from the model pattern (as cited in PCER Consortium, 2008).
Building Blocks, Shape Composition task	Modified for PCER from the Building Blocks assessment tools. Children use blocks to fill in a puzzle and are assessed on whether they fill the puzzle without gaps or hangovers (as cited in PCER Consortium, 2008).

Appendix A3.1 Summary of study findings included in the rating for the oral language domain¹

Outcome measure	Study sample	Sample size (classrooms/ children)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ⁴ (Doors to Discovery™ – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
			Doors to Discovery™ group ³	Comparison group				
PCER Consortium, 2008 (randomized controlled trial)⁸								
PPVT-III	Preschoolers	29/183	94.63 (18.20)	91.33 (18.12)	3.30	0.15	ns	+6
TOLD-P:3 Grammatic Understanding subtest	Preschoolers	29/183	10.19 (3.06)	9.33 (2.71)	0.86	0.17	ns	+7
Average for oral language (PCER Consortium, 2008)⁹						0.16	na	+6
Christie et al., 2003 (randomized controlled trial)⁸								
PPVT-III	Preschoolers	4/37	35.98 (19.32)	30.25 (17.09)	5.73	0.30	ns	+12
Average for oral language (Christie et al., 2003)⁹						0.30	na	+12
Domain average for oral language across all studies⁹						0.23	na	+9

ns = not statistically significant

na = not applicable

PPVT-III = Peabody Picture Vocabulary Test-III

TOLD-P:3 = Test of Language Development Primary, Third Edition

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the oral language domain. Follow-up findings from PCER Consortium (2008) are not included in these ratings, but are reported in Appendix A4.1.
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. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted. For Christie et al. (2003), the treatment group means are the sum of the control group means and the mean difference, which is adjusted for pretest. The standard deviations were pooled across classrooms.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted. For the study by Christie et al. (2003), the WWC excluded one non-randomly assigned classroom, so the means, standard deviations, effect sizes, and significance levels in this report may differ from those reported in the original study.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's *d* based on a repeated measures analysis).
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 results favorable to the intervention group.
8. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections for clustering or multiple comparisons were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant. In the case of Christie et al. (2003), the WWC corrected for clustering.
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 print knowledge domain¹

Outcome measure	Study sample	Sample size (classrooms/ children)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ⁴ (Doors to Discovery™ – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
		Doors to Discovery™ group ³	Comparison group					
PCER Consortium, 2008 (randomized controlled trial)⁸								
TERA-3	Preschoolers	29/182	93.4 (17.22)	92.76 (17.86)	0.64	0.06	ns	+2
WJ-III Letter-Word Identification subtest	Preschoolers	29/183	108.82 (14.56)	106.04 (13.82)	2.78	0.10	ns	+4
WJ-III Spelling subtest	Preschoolers	29/183	98.91 (12.56)	97.37 (12.63)	1.54	0.06	ns	+2
Average for print knowledge (PCER Consortium, 2008)⁹						0.07	na	+3
Christie et al., 2003 (randomized controlled trial)⁸								
Concepts of Print	Preschoolers	4/37	4.48 (1.51)	2.82 (1.49)	1.66	1.08	ns	+37
Get Ready to Read!	Preschoolers	4/37	8.62 (3.96)	7.06 (3.81)	1.56	0.39	ns	+16
Average for print knowledge (Christie et al., 2003)⁹						0.74	na	+27
Domain average for print knowledge across all studies⁹						0.41	na	+16

ns = not statistically significant

na = not applicable

TERA-3 = Test of Early Reading Ability-III

WJ-III = Woodcock-Johnson III

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the print knowledge domain. Follow-up findings from PCER Consortium (2008) are not included in these ratings, but are reported in Appendix A4.2.
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. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted. For Christie et al. (2003), the treatment group means are the sum of the control group means and the mean difference, which is adjusted for pretest. The standard deviations were pooled across classrooms.

(continued)

Appendix A3.2 Summary of study findings included in the rating for the print knowledge domain¹ *(continued)*

4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted. For the study by Christie et al. (2003), the WWC excluded one non-randomly assigned classroom, so the means, standard deviations, effect sizes, and significance levels in this report may differ from those reported in the original study.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's *d* based on a repeated measures analysis).
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 results favorable to the intervention group.
8. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections for clustering or multiple comparisons were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant. In the case of Christie et al. (2003), the WWC corrected for clustering.
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 A3.3 Summary of study findings included in the rating for the phonological processing domain¹

Outcome measure	Study sample	Sample size (classrooms/ children)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ⁴ (<i>Doors to Discovery</i> TM – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
		<i>Doors to Discovery</i> TM group ³	Comparison group					
PCER Consortium, 2008 (randomized controlled trial)⁸								
Pre-CTOPPP Elision subtest	Preschoolers	29/182	10.78 (4.18)	10.11 (4.64)	0.67	0.18	ns	+7
Domain average for phonological processing (PCER Consortium, 2008)⁹						0.18	na	+7

ns = not statistically significant

na = not applicable

Pre-CTOPPP = Preschool Comprehensive Test of Phonological and Print Processing

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the phonological processing domain. Follow-up findings from PCER Consortium (2008) are not included in these ratings, but are reported in Appendix A4.3.
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. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's *d* based on a repeated measures analysis).
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 results favorable to the intervention group.
8. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections for clustering or multiple comparisons were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant.
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 A3.4 Summary of study findings included in the rating for the math domain¹

Outcome measure	Study sample	Sample size (classrooms/ children)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ⁴ (Doors to Discovery™ – comparison)	Effect size ⁵	Statistical significance ⁶ (at $\alpha = 0.05$)	Improvement index ⁷
		Doors to Discovery™ group ³	Comparison group					
PCER Consortium, 2008 (randomized controlled trial)⁸								
WJ-III Applied Problems subtest	Preschoolers	29/183	99.53 (13.24)	99.28 (16.60)	0.25	0.01	ns	+0
CMA-A Composite	Preschoolers	29/183	0.68 (0.20)	0.65 (0.24)	0.03	0.13	ns	+5
Shape Composition	Preschoolers	29/183	1.61 (0.84)	1.72 (0.69)	-0.11	-0.13	ns	-5
Domain average for math (PCER Consortium, 2008)⁹						0.00	na	+0

ns = not statistically significant

na = not applicable

WJ-III = Woodcock-Johnson III

CMA-A = Child Math Assessment-Abbreviated

1. This appendix reports findings considered for the effectiveness rating and the average improvement indices for the math domain. Follow-up findings from PCER Consortium (2008) are not included in these ratings, but are reported in Appendix A4.4.
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. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted.
5. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's d based on a repeated measures analysis).
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 results favorable to the intervention group.
8. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections for clustering or multiple comparisons were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant.
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 A4.1 Summary of follow-up findings for the oral language domain¹

Outcome measure	Study sample	Sample size ³ (classrooms/ children)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ⁵ (Doors to Discovery™ – comparison)	Effect size ⁶	Statistical significance ⁷ (at $\alpha = 0.05$)	Improvement index ⁸
		Doors to Discovery™ group ⁴	Comparison group					
PCER Consortium, 2008 (randomized controlled trial)⁹								
PPVT-III	Kindergarteners	nr/152	98.13 (17.46)	94.00 (16.01)	4.13	0.18	ns	+7
TOLD-P:3 Grammatic Understanding subtest	Kindergarteners	nr/155	10.41 (3.19)	10.08 (2.80)	0.33	0.06	ns	+2

ns = not statistically significant

nr = not reported

PPVT-III = Peabody Picture Vocabulary Test-III

TOLD-P:3 = Test of Language Development Primary, Third Edition

1. This appendix presents follow-up findings considered for measures that fall in the oral language domain. End-of-preschool scores 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. The PCER Consortium (2008) study included 149 kindergarten classrooms across all three conditions in this study (*Doors to Discovery™*, control, and *Let's Begin with the Letter People®*). The number of classrooms for *Doors to Discovery™* and the control group is likely about two-thirds of the total.
4. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted.
6. For an explanation of the effect size calculation, see WWC Computations and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's *d* based on a repeated measures analysis).
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
8. 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 results favorable to the intervention group.
9. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant.

Appendix A4.2 Summary of follow-up findings for the print knowledge domain¹

Outcome measure	Study sample	Sample size ³ (classrooms/ children)	Authors' findings from the study		WWC calculations				
			Mean outcome (standard deviation) ²		Mean difference ⁵ (Doors to Discovery™ – comparison)	Effect size ⁶	Statistical significance ⁷ (at $\alpha = 0.05$)	Improvement index ⁸	
			Doors to Discovery™ group ⁴	Comparison group					
PCER Consortium, 2008 (randomized controlled trial)⁹									
TERA-3	Kindergarteners	nr/155	93.38 (18.88)	93.96 (16.47)	–0.58	–0.05	ns	–2	
WJ-III Letter-Word Identification subtest	Kindergarteners	nr/155	106.99 (14.82)	109.53 (13.57)	–2.54	–0.09	ns	–4	
WJ-III Spelling subtest	Kindergarteners	nr/155	100.51 (14.84)	103.46 (13.14)	–2.95	–0.12	ns	–5	

ns = not statistically significant

nr = not reported

TERA-3 = Test of Early Reading Ability-III

WJ-III = Woodcock-Johnson III

1. This appendix presents follow-up findings for measures that fall in the print knowledge domain. End-of-preschool scores 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. The PCER Consortium (2008) study included 149 kindergarten classrooms across all three conditions in this study (*Doors to Discovery™*, control, and *Let's Begin with the Letter People®*). The number of classrooms for *Doors to Discovery™* and the control group is likely about two-thirds of the total.
4. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted.
6. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's *d* based on a repeated measures analysis).
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
8. 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 results favorable to the intervention group.
9. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant.

Appendix A4.3 Summary of follow-up findings for the phonological processing domain¹

Outcome measure	Study sample	Sample size ³ (classrooms/ children)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ⁵ (Doors to Discovery™ - comparison)	Effect size ⁶	Statistical significance ⁷ (at $\alpha = 0.05$)	Improvement index ⁸
			Doors to Discovery™ group ⁴	Comparison group				
PCER Consortium, 2008 (randomized controlled trial)⁹								
CTOPP Elision subtest	Kindergarteners	nr/155	4.68 (3.84)	5.04 (4.24)	-0.36	-0.09	ns	-4

ns = not statistically significant

nr = not reported

CTOPP = Comprehensive Test of Phonological Processing

1. This appendix presents follow-up findings for measures that fall in the phonological processing domain. End-of-preschool scores were used for rating purposes and are presented in Appendix A3.3.
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 PCER Consortium (2008) study included 149 kindergarten classrooms across all three conditions in this study (*Doors to Discovery™*, control, and *Let's Begin with the Letter People®*). The number of classrooms for *Doors to Discovery™* and the control group is likely about two-thirds of the total.
4. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted.
6. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's *d* based on ANCOVA).
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
8. 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 results favorable to the intervention group.
9. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant.

Appendix A4.4 Summary of follow-up findings for the math domain¹

Outcome measure	Study sample	Sample size ³ (classrooms/ children)	Authors' findings from the study		WWC calculations			
			Mean outcome (standard deviation) ²		Mean difference ⁵ (Doors to Discovery™ – comparison)	Effect size ⁶	Statistical significance ⁷ (at $\alpha = 0.05$)	Improvement index ⁸
		Doors to Discovery™ group ⁴	Comparison group					
PCER Consortium, 2008 (randomized controlled trial)⁹								
WJ-III Applied Problems subtest	Kindergarteners	nr/155	101.84 (10.95)	102.40 (11.38)	–0.56	–0.02	ns	–1
CMA-A Composite	Kindergarteners	nr/155	0.68 (0.16)	0.72 (0.14)	–0.04	–0.16	ns	–6
Shape Composition	Kindergarteners	nr/155	2.40 (0.79)	2.51 (0.69)	–0.11	–0.12	ns	–5

ns = not statistically significant

nr = not reported

WJ-III = Woodcock-Johnson III

CMA-A = Child Math Assessment-Abbreviated

1. This appendix presents follow-up findings for measures that fall in the math domain. End-of-preschool scores were used for rating purposes and are presented in Appendix A3.4.
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 PCER Consortium (2008) study included 149 kindergarten classrooms across all three conditions in this study (*Doors to Discovery™*, control, and *Let's Begin with the Letter People®*). The number of classrooms for *Doors to Discovery™* and the control group is likely about two-thirds of the total.
4. In PCER Consortium (2008), the treatment group mean equals the unadjusted control group mean and the covariate-adjusted mean difference. Standard deviations are unadjusted.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group. In the case of PCER Consortium (2008), the mean differences are covariate-adjusted.
6. For an explanation of the effect size calculation, see WWC Procedures and Standards Handbook, Appendix B. In the case of PCER Consortium (2008), the WWC used the effect sizes reported by the study authors (Cohen's *d* based on a repeated measures analysis).
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
8. 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 results favorable to the intervention group.
9. 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. For the formulas the WWC used to calculate 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 PCER Consortium (2008), no corrections were needed because the analysis corrected for clustering by using HLM, and no impacts were statistically significant.

Appendix A5.1 *Doors to Discovery*[™] rating for the oral language 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 oral language, the WWC rated *Doors to Discovery*[™] as having potentially positive effects. The remaining ratings (mixed effects, no discernible effects, potentially negative, negative) were not considered, as *Doors to Discovery*[™] 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 of two studies that measured oral language showed a 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.

Met. Neither of the two studies that measured oral language showed a statistically significant or substantively important negative effect. One study showed a substantively important positive effect, and one study showed no effect.

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. Neither of the two studies that measured oral language showed a statistically significant positive effect.

AND

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

Met. Neither of the two studies that measured oral language 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. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A5.2 *Doors to Discovery*TM rating for the print knowledge 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 print knowledge, the WWC rated *Doors to Discovery*TM as having potentially positive effects. The remaining ratings (mixed effects, no discernible effects, potentially negative, negative) were not considered, as *Doors to Discovery*TM 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 of the two studies that measured print knowledge showed a 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.

Met. Neither of the two studies that measured print knowledge showed a statistically significant or substantively important negative effect. One study showed a substantively important positive effect, and one study showed an effect that was not statistically significant or substantively important.

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. Neither of the two studies that measured print knowledge showed a statistically significant positive effect.

AND

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

Met. Neither of the two studies that measured print knowledge 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. For a complete description, see the WWC Procedures and Standards Handbook, Appendix E.

Appendix A5.3 *Doors to Discovery*TM rating for the phonological processing 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 phonological processing, the WWC rated *Doors to Discovery*TM as having no discernible effects. The remaining ratings (potentially negative, negative) were not considered, as *Doors to Discovery*TM was assigned the highest applicable rating.

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 one study that measured phonological processing showed no statistically significant or substantively important effect.

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 that measured phonological processing showed no statistically significant effect.

AND

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

Met. The one study that measured phonological processing did not show a statistically significant or substantively important negative effect.

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 one study that measured phonological processing showed no statistically significant or substantively important 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. The one study that measured phonological processing showed no statistically significant or substantively important effect. No other studies measured phonological processing.

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 one study that measured phonological processing showed no statistically significant or substantively important effect. No other studies measured phonological processing.

(continued)

Appendix A5.3 *Doors to Discovery™* rating for the phonological processing 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. The one study that measured phonological processing showed no statistically significant or substantively important effect. No other studies measured phonological processing.
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 A5.4 *Doors to Discovery*[™] rating for the math 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 math, the WWC rated *Doors to Discovery*[™] as having no discernible effects. The remaining ratings (potentially negative, negative) were not considered, as *Doors to Discovery*[™] was assigned the highest applicable rating.

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 one study that measured math showed no statistically significant or substantively important effect.

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 that measured math showed no statistically significant or substantively important positive effect.

AND

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

Met. The one study that measured math did not show a statistically significant or substantively important negative effect.

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 one study that measured math showed no statistically significant or substantively important 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. The one study that measured math showed no statistically significant or substantively important effect. No other studies measured math.

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 one study that measured math showed no statistically significant or substantively important effect. No other studies measured math.

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. The one study that measured math showed no statistically significant or substantively important effect. No other studies measured math.

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	
Oral language	2	33	220	Medium to large
Print knowledge	2	33	220	Medium to large
Phonological processing	1	29	182	Small
Early reading and writing	0	na	na	na
Cognition	0	na	na	na
Math	1	29	183	Small

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.