

The Effects of an Academic Language Program on Student Reading Outcomes

NCEE 2022-007a
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation at IES



U.S. Department of Education

Miguel Cardona

Secretary

Institute of Education Sciences

Mark Schneider

Director

National Center for Education Evaluation and Regional Assistance

Matthew Soldner

Commissioner

Marsha Silverberg

Associate Commissioner

Tracy Rimdzius

Project Officer

The Institute of Education Sciences (IES) is the independent, nonpartisan statistics, research, and evaluation arm of the U.S. Department of Education. The IES mission is to provide scientific evidence on which to ground education practice and policy and to share this information in formats that are useful and accessible to educators, parents, policymakers, researchers, and the public.

We strive to make our products available in a variety of formats and in language that is appropriate for a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other IES product or report, we would like to hear from you. Please direct your comments to ncee.feedback@ed.gov.

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-15-C-0050 by MDRC. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

August 2022

This report is in the public domain. Although permission to reprint this publication is not necessary, it should be cited as:

Corrin, W., Zhu, P., Shih, M., Brown, K. T., Teres, J., Darrow, C., Nichols, A., & Lack, K. (2022). *The Effects of an Academic Language Program on Student Reading Outcomes* (NCEE 2022-007). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. <http://ies.ed.gov/ncee>.

This report is available on the Institute of Education Sciences website at <http://ies.ed.gov/ncee>.



The Effects of an Academic Language Program on Student Reading Outcomes

August 2022

William Corrin

Pei Zhu

Miki Shih

Kevin Thaddeus Brown, Jr.

Jed Teres

MDRC

Catherine Darrow

Austin Nichols

Kelly Lack

Abt Associates

CONTENTS

- LIST OF EXHIBITS..... iii
- INTRODUCTION..... 1
- APPENDIX A. THE ACADEMIC LANGUAGE PROGRAM 2
 - I. Program Selection and Brief Overview 2
 - II. Program Theory and Content 2
 - III. Training and Support 4
 - IV. Non-Study Access to WordGen Elementary Program Materials, Training, and Support 6
- APPENDIX B. STUDY DESIGN, DATA COLLECTION, AND ANALYTIC APPROACHES 7
 - I. Study Design..... 7
 - II. Data-Collection Activities..... 13
 - III. Analytic Approaches 18
- APPENDIX C. SUPPLEMENTAL INFORMATION ON FINDINGS IN THE REPORT 40
 - I. Additional Details on Program Impact Findings..... 40
 - II. Relationship Between Teacher Training, Instructional Practices, and Student Outcomes 49
 - III. Additional Details on Program Implementation 49
 - IV. Additional Information for Systematic Review..... 60
- APPENDIX D. SUPPLEMENTAL FINDINGS..... 66
 - I. Sensitivity Checks of Program Impacts on Student Outcomes..... 66
 - II. Supplemental Findings of Program Impacts on Student Outcomes 72
 - III. Supplemental Information About Contrasts in Program Implementation..... 89
- ENDNOTES 119
- REFERENCES..... 120

LIST OF EXHIBITS

A.1	Academic Language Program Theory of Change.....	3
A.2	Example of a Fifth-Grade Unit: Should Everyone Be Included?.....	4
B.1	School Eligibility Criteria and Recruitment of Study Districts.....	8
B.2	Comparison of Schools Remaining in the Study and Schools That Left the Study	9
B.3	Comparison of Study Schools and Public Elementary Schools Nationally.....	11
B.4	Comparison of Program and Non-Program Schools in the Study	12
B.5	Data-Collection Activities.....	15
B.6	Response Rates for Data Sources Used to Estimate Program Effects.....	16
B.7	Background Characteristic Comparison of Teacher Survey Respondents in Program and Non-Program Schools	17
B.8	Core Academic Language Domains and Skill Sets Measured by CALS-I.....	18
B.9	Checklist for Teacher Instructional Practices	21
B.10	Classroom Assessment Scoring System-Upper Elementary (CLASS-UE) Domains and Dimensions.....	23
B.11	Student Sample Formation for Impact Estimation on CALS-I and GMRT Tests	25
B.12	Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for CALS-I Analysis, Program Year (2017-2018).....	27
B.13	Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for CALS-I Analysis, Program Year (2017-2018)	28
B.14	Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for CALS-I Analysis, Program Year (2017-2018)	29
B.15	Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for GMRT Analysis, Program Year (2017-2018)	30
B.16	Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for GMRT Analysis, Program Year (2017-2018)	31
B.17	Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for GMRT Analysis, Program Year (2017-2018)	32

B.18	Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for State ELA Test Analysis, Program Year (2017-2018)	33
B.19	Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for State ELA Test Analysis, Program Year (2017-2018)	34
B.20	Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for State ELA Test Analysis, Program Year (2017-2018)	35
B.21	Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for State ELA Test Analysis, Follow-Up Year (2018-2019)	36
B.22	Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for State ELA Test Analysis, Follow-Up Year (2018-2019)	37
B.23	Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for State ELA Test Analysis, Follow-Up Year (2018-2019)	38
B.24	Realized Minimum Detectable Effects by Outcome and Sample	39
C.1	Estimated Impacts on Student Language and Reading Outcomes, Overall Sample, Program Year	41
C.2	Estimated Impacts on Student Language and Reading Outcomes, by District, Program Year	42
C.3	Estimated Impacts on Student Language and Reading Outcomes for English Learners and Students from Economically Disadvantaged Backgrounds, Program Year	44
C.4	Estimated Impacts on Students' Performances in State English Language Arts Tests in Follow-Up Year, Overall Sample, English Learners, and Students from Economically Disadvantaged Backgrounds	46
C.5	Estimated Impacts on the Use of Instructional Practices Important for Academic Language Development	47
C.6	Estimated Impacts on General Classroom Management Quality, as Measured by Classroom Assessment Scoring System-Upper Elementary (CLASS-UE)	48
C.7	Associations Between Training and Support and Teachers' Use of Program-Specific Instructional Practices, Program Schools.....	50
C.8	Associations Between Teachers' Use of Program-Specific Instructional Practices and Student Outcomes, Program Schools	51
C.9	Teacher Attendance at Initial Training, Overall and by District and Training Days.....	52
C.10	Teacher Attendance at Guidance Sessions, Overall and by District and Session	53
C.11	Teacher Attendance at Reflection Sessions, Overall and by District and Session	55
C.12	Amount of Training and Support Teachers Received During the Program Year	56

C.13	Estimated Differences in the Amount of Training and Professional Development Reported by Teachers, Program Year.....	57
C.14	Number of Available Instructional Days and Number of Curricular Units Covered, Overall and by District, Program Year.....	59
C.15	Top Five Implementation Challenges Reported by Coaches	60
C.16	Supplemental Information on Student Baseline Reading and Math Achievement, by Analysis Sample	61
C.17	Supplemental Information on Student Outcomes, by Outcome and Sample.....	63
C.18	Estimated Intraclass Correlation Coefficients and Explanatory Power of Covariates (R-square) in Impact Estimation Model.....	65
D.1	Model Specification Checks for Impacts on Student Outcomes for Program Year, Overall Sample and Subgroups.....	67
D.2	Sample Specification Checks for Impacts on Student Outcomes, for the Overall Sample and Subgroups	69
D.3	Background Characteristics Comparison of Students With and Without CALS-I or GMRT Scores	71
D.4	Estimated Impacts on Percentages of Students Meeting Proficiency Standards in State ELA Tests, by Respondent Status, Sample, and Year	73
D.5	Estimated Impacts on CALS-I Subscales.....	75
D.6	Estimated Impacts on State ELA Test Standardized Scores, by Sample and Year	76
D.7	Estimated Impacts on State Math Test Scores, by Sample and by Year.....	77
D.8	Estimated Impacts on Student Outcomes in the Program Year, by English Learner Status	79
D.9	Estimated Impacts on Student Outcomes in the Program Year, by Economic Background.....	81
D.10	Estimated Impacts on Student Outcomes in the Program Year, by Student Gender	83
D.11	Estimated Impacts on Student Outcomes in the Program Year, by Grade Level.....	85
D.12	Estimated Impacts on Student Outcomes in the Program Year, by Pre-Program Reading Level	87
D.13	Estimated Impacts on CALS-I Scores, by District-Level Subgroup	90
D.14	Estimated Impacts on CALS-I Scores, by Random Assignment Block Level Subgroup	93
D.15	Estimated Impacts on GMRT Scores, by District-Level Subgroups.....	95
D.16	Estimated Impacts on GMRT Scores, by Random Assignment Block Level Subgroup.....	98
D.17	Estimated Impacts on State ELA Test Performance, by District-Level Subgroup	100

D.18	Estimated Impacts on State ELA Test Performance, by Random Assignment Block Level Subgroup	103
D.19	Program Effects on Teachers' Use of Core Instructional Practices, by Item	105
D.20	Program Effects on Teachers' Use of Core Instructional Practices, by Site Starting Time, by Item	108
D.21	Relationship Between the Amount of Professional Development Teachers Reported and Student Outcomes.....	112
D.22	Relationship Between Teachers' Use of Core Instructions and Student Outcomes	113
D.23	Estimated Impacts on Teachers' Self-Reported Use of Instructional Practices that Support Academic Language Development, by Practice.....	115
D.24	Estimated Impacts on Teachers' Self-Reported Attitudes and Perceived Challenges	117

INTRODUCTION

Many schools have struggled to effectively help English learners and students from economically disadvantaged backgrounds perform as well as their more advantaged peers in reading achievement. Improving these students' reading performance is crucial in late elementary grades as they prepare for increasingly unfamiliar and subject-specific language in middle school. Existing research suggests that academic language, the formal language that students read, write, hear, and speak in their studies at school, is critical to their academic success. This study investigated a program designed to increase fourth- and fifth-grade students' ability to understand and use academic language and to improve their general reading skill through text-based and oral activities. The program provider offered training and ongoing professional development activities designed to support teachers' delivery of the associated curriculum in their classrooms. About sixty schools from six districts around the country were assigned at random to implement the program or continue with their typical language instruction programs and practices. The study gauged the program's effects on student outcomes by comparing the average student reading performance of these two groups of schools. This document provides supporting details on the academic language program being evaluated, the research activities carried out by the study team, and supplementary analyses for the findings presented in the report.

APPENDIX A. THE ACADEMIC LANGUAGE PROGRAM

This appendix provides additional information about WordGen Elementary, the academic language program evaluated in this study. The appendix begins with a brief discussion of the selection of the program and then turns to additional information about the program—its curriculum, classroom activities, and the program provider’s training and support plan. The appendix ends with information about access to the curricular materials as well as training and support under typical, non-study circumstances.

I. Program Selection and Brief Overview

For this evaluation, the Department of Education sought to evaluate the impact of an academic language program on the outcomes of students in the late elementary grades. The intent of the study was to assess the impact of a fully developed program with existing materials and prior implementation and not to conduct research on an emerging or new program still under development. The study team ran a competition to select such a program for evaluation. At the end of the summer of 2016, the team released a Request for Proposals from providers of existing academic language programs for fourth- and fifth-grade students. A panel of academic language experts reviewed and scored submissions, considering factors such as the quality of the proposed implementation plan, prior research evidence for the program proposed, and the provider’s prior experience implementing the program. The panel also participated in an in-person finalist presentation. The Strategic Education Research Partnership’s (SERP) proposal to provide WordGen Elementary materials and implementation support was selected for the evaluation.

II. Program Theory and Content

Theory of Change

The developers of WordGen Elementary theorized that with timely, targeted, and high-quality training and supports, teachers would develop instructional skills that would broaden students’ academic word knowledge, support academic skill development, and provide opportunities for students to practice through their delivery of the program’s 12-unit curriculum. Through exposure to the curriculum as delivered by these trained teachers, students’ academic language skills, reading comprehension, and English language arts (ELA) achievement would improve (see Exhibit A.1).

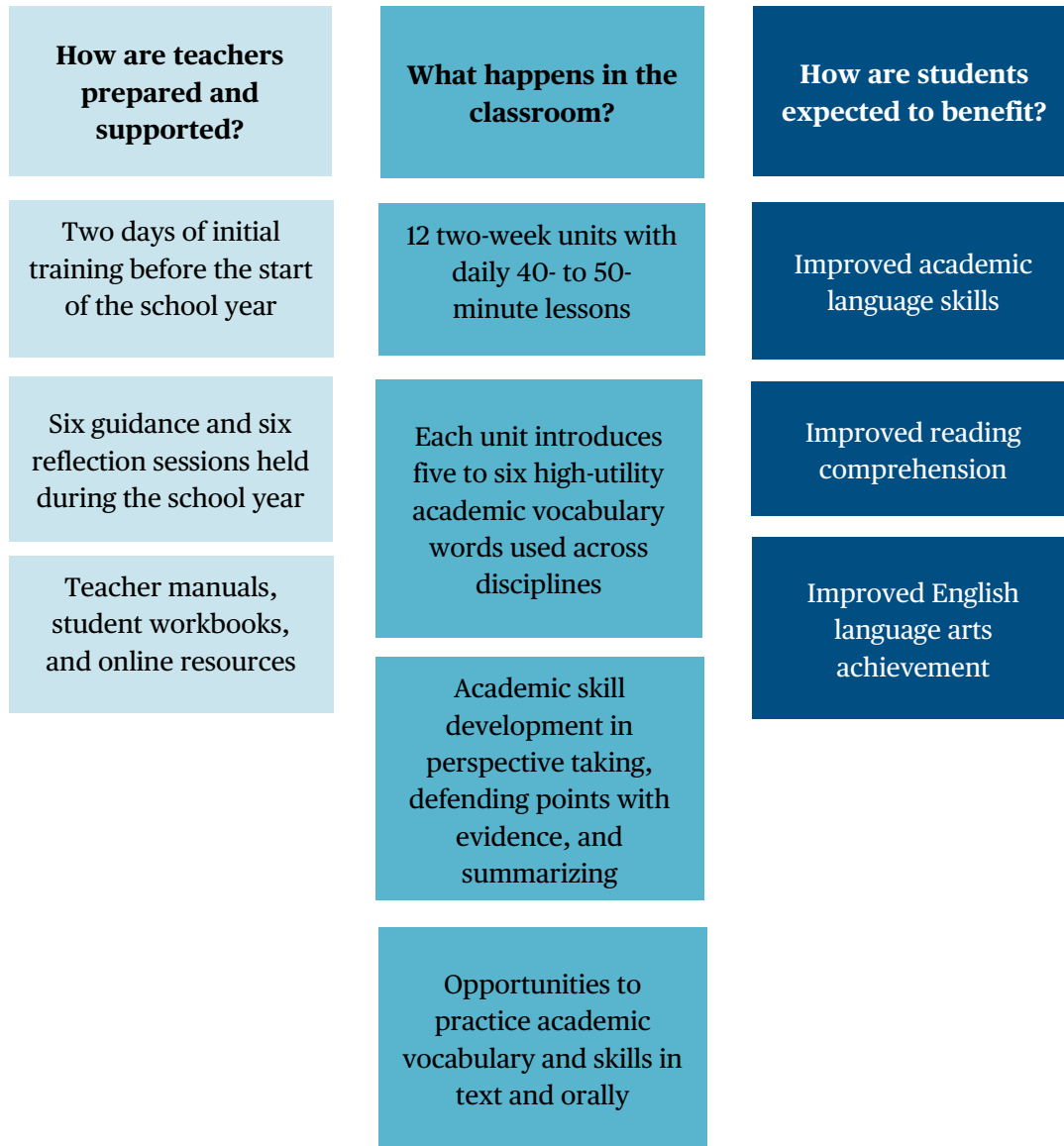
Curriculum and Instruction

The WordGen Elementary curriculum, developed for Grades 4 and 5, consists of 12 two-week units with 40-50-minute daily lessons. Each unit introduces approximately five to six high-utility academic “focus words”, focuses on a critical topic, and is designed to offer a variety of texts, word-learning activities, writing tasks, and opportunities for discussion and debate. Throughout a unit, students read, discuss, debate, and write about the focal topic using the focus words. WordGen Elementary is also designed to be applicable to multiple subject areas. The units’ content as well as their oral language and text-based activities are relevant to English language arts, humanities, social studies, math, and science. The curriculum is available at SERP’s website: <https://www.serp.institute.org/wordgen-elementary>.

Students are introduced to the unit through a video newscast and a “Reader’s Theater” that introduces multiple perspectives on a high-interest topic. These topics are meant to stimulate student interest and promote engagement in related materials, written work, and discussion-based activities. Topics include questions about fairness, autonomy, self-identification, and freedom. By engaging students in interesting topics, the curriculum fosters academic language development, argumentation, perspective-taking, and writing.

The ten activities across a unit typically cluster into three categories. The first two to three activities are designed to provide the students with an introduction to the unit topic and focus words. The middle four to five activities focus on building background knowledge and strengthening academic skills. The last two to three activities are when students synthesize information and demonstrate critical reasoning, using the unit’s focus words and applying their academic


Exhibit A.1. Academic Language Program Theory of Change



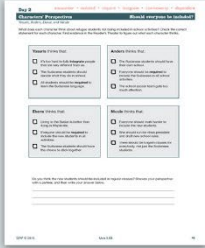
skills. Exhibit A.2 provides an example from a fifth-grade unit. As the unit progresses, students take on more prominent roles in the classroom, as their interactions, particularly their involvement in discussion and debate, are viewed as playing an essential role in the development of their academic language skills. This progression also means that the teacher has to be more and more deft with instruction, making sure that the classroom activities remain productive as the students become more active. For more information about the types of lesson activities or components represented within a unit, see <https://www.serpinstitute.org/wordgen-elementary/components>.

Exhibit A.2. Example of a Fifth-Grade Unit: Should Everyone Be Included?

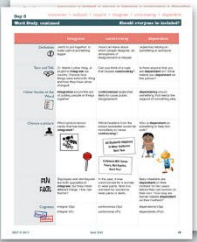
Introduction to the issue, different perspectives, and the focus words



Day 1
Action News/Reader's Theater




Day 2
Characters' Perspectives

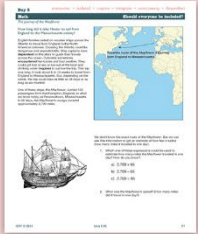


Day 3
Word Study


Building background knowledge across content areas



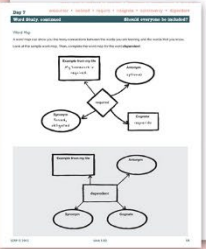
Day 4
Journals and Journeys



Day 5
Math Activity

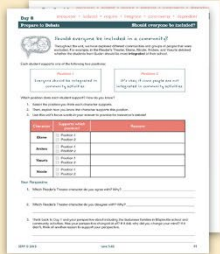


Day 6
Informational Text

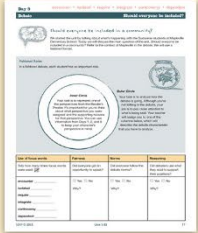


Day 7
Informational Text


Synthesizing information through discussion, debate, and argumentative writing



Day 8
Prepare for Debate



Day 9
Debate



Day 10
Writing Activity

NOTE: This graphic is provided by the Strategic Education Research Partnership.

III. Training and Support

The implementation of WordGen Elementary in this study relied on a combination of centralized training and local district support. SERP offered a centralized training for local district coaches and teachers integrated with its national summer institute, training them on the curricular content and the pedagogical practices integral to its delivery. Then

local coaches, with ongoing support from SERP, provided support to teachers as they implemented the program during the school year.

Coach Training

In the summer prior to the implementation year, academic language coaches attended a four-day training at Harvard University facilitated by the SERP team and academic language experts. The first day of this coaches' institute was focused on the role of coaches, an overview of the WordGen Elementary program, and a review of study-related expectations. Coaches attended the National Teachers' Institute for the second and third days of training. The final day was again dedicated to coaches and primarily focused on preparation for district-based teachers' institutes.

While seven of the coaches attended the four-day summer training, five coaches who were hired after the summer institute or were absent during the summer were unable to attend the full training. These coaches were expected to attend a district-based teachers' institute in another district (to cover the content of days 2 and 3 of the coaches' institute). SERP's lead coaches were to meet with these coaches the day before or after the district-based teachers' institute to brief them on the content of day 1 and day 4 of the coaches' institute that they had missed. Thus, these coaches still received some preparation and training prior to their own district's teachers' institute.

Teacher Training

Prior to the start of WordGen Elementary implementation, teachers were expected to attend a two-day teachers' institute located at a central location within each school district. This institute was facilitated by one of SERP's lead coaches and the district-specific local coach(es). The first day of training provided an overview of the WordGen Elementary program and an introduction to academic language. The second day focused on methods for supporting student discourse, supporting English learners, and acceptable adaptations to the program.

For teachers who were unable to attend the two-day training, some districts offered alternative date options and others offered make-up sessions or webinars recorded by SERP's lead coaches. District coaches facilitated teachers' access to this webinar, which reviewed key pieces of each training day and provided access to the materials used for each session for further independent exploration.

Coach Support

The central SERP coach, one of SERP's lead coaches, was expected to hold monthly calls for all local coaches to discuss WordGen Elementary implementation. As needed, the central coach would provide additional coaching via phone calls and email communication to individual coaches. The central coach was also expected to visit each district in the fall after implementation began and again in the spring. If a district required additional in-person help, the central coach would conduct a third site visit. During visits, the central coach would meet with the district coach, visit each school even if briefly, and observe classrooms of teachers needing additional support. The duration and intensity of visits would depend on the needs of coaches and associated schools.

Teacher Support

SERP planned for locally based coaches to deliver 20 hours of monthly support for teachers' implementation of WordGen Elementary at each school. The teacher support plan included both group-based and individualized professional learning. For the former, coaches provided "guidance" and "reflection" sessions. Coaches were to deliver 55-minute guidance sessions to teachers across the school year at intervals spanning approximately 20 instructional days. In each, WordGen Elementary experts introduced teachers to basic principles integral to WordGen Elementary instruction (for example, reasoning or argumentation) in preparation for delivery of the next two WordGen units. Each guidance session was to be recorded by a WordGen expert in advance for teachers at each school site to view as a team with the district coach. All teachers and coaches were expected to attend all six sessions. SERP also intended for coaches to host reflection sessions, held one to two weeks after each guidance session. Teachers at each school were expected to meet as a team with their local coach to discuss their experience using the practices in their classrooms.

Coaches were also supposed to provide individualized support to specific teachers who struggled with implementation. Coaches were encouraged to deliver support using a combination of in-person meetings, email communication, and phone calls. In addition, teachers received online support via webinars and an online WordGen Elementary community where they could access curricular materials and message boards.

IV. Non-Study Access to WordGen Elementary Program Materials, Training, and Support

At the time of this study, school district and school administrators could choose how to approach training teachers to support the adoption and implementation of the WordGen Elementary program. They could independently use the program materials (via free download or paying for printed student and teacher materials) without any support from SERP. They could send educators to an annual national summer training institute run by SERP. They could contract with SERP for training and support of teachers tailored to their district. If desired, districts could send educators to the national institute and have SERP provide local district support as well.

Since the implementation of WordGen Elementary for this study, SERP has further developed and refined virtual/online resources and training modules for educators. This includes an initial training module that replaced the summer institute starting in the summer of 2021, which districts and schools can access and use as they see best within their teacher professional development programming. Although still possible to arrange, tailored district support provided on-site by SERP is now rare. The curricular materials for students and teachers are still available for download at no cost and districts and schools can pay for printed materials.

APPENDIX B. STUDY DESIGN, DATA COLLECTION, AND ANALYTIC APPROACHES

This appendix describes the study design, the site recruitment process, and random assignment. It then introduces the data-collection activities and the resulting analytic samples. Lastly, the appendix presents the approaches used for the impact estimation and the exploratory analyses.

I. Study Design

This study used an experimental design that randomly selected schools to participate in the academic language program or continue with their usual strategies for language learning. The evaluation seeks to address the following research questions about the effects of the program:

1. What is the impact of WordGen Elementary on student achievement, including students' academic language skills, reading comprehension, and their general reading achievement?
2. What is the effect of WordGen Elementary on teacher instructional practices, particularly those important for academic language development?

This section describes how the team carried out this design through the recruitment and random assignment of schools.

Recruitment and Selection of Study Sites and Schools

The study team recruited 70 elementary schools from 6 school districts across the country to participate in the study. Because the study was particularly interested in the program's effects on English learners and students from economically disadvantaged backgrounds, the recruitment efforts targeted schools and districts with high concentrations of these students. It also screened districts and schools for their willingness and capacity to support program implementation in schools and to cooperate with the study's data collection efforts. Lastly, the team prioritized districts that did not already provide a similar language program to ensure that there would be meaningful contrast between the program and non-program schools.

Site recruitment occurred in three phases. In the first phase, the study team used data from the Common Core of Data (CCD, 2014-2015) and the Civil Rights Data Collection databases (CRDC, 2013-2014) to identify schools meeting eligibility criteria related to student populations (see Exhibit B.1 for list of eligibility criteria). This generated a list of 3,364 schools and 292 districts.

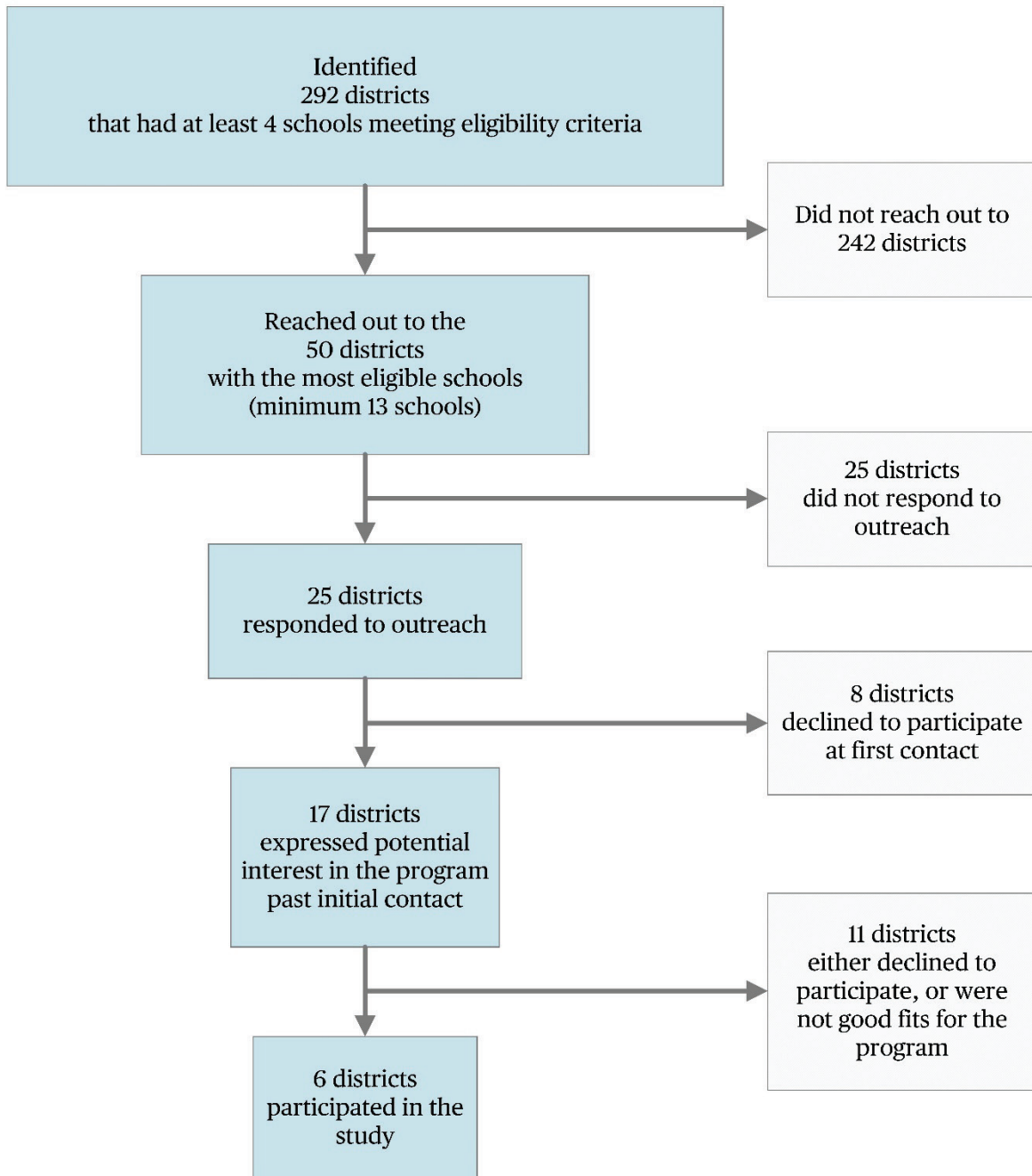
In the second phase, the study team reached out to the 50 districts with the largest number of eligible schools (minimum 13 schools). Of these 50 districts, 25 responded to the outreach. Eight of the 25 districts declined to participate in the study after a single interaction. After multiple interactions, 11 additional districts were removed from consideration, either because they declined to participate or because their policies or local context were not well-aligned with study goals. Districts that declined participation gave reasons that included reluctance to take on a new initiative, recent adoption of other ELA curricula or programs, or a lack of agreement between district ELA and English language learner coordinators on the value of participating.

In the third phase, six districts committed to study participation. These districts identified a total of 70 eligible schools for study participation. The study team included these schools in the study sample. The team recruited the study sample on a rolling basis. It recruited four of the six study districts by the summer of 2017, and enrolled the last two districts early in the fall of 2017, after the start of the program year. Exhibit B.1 presents the results of the recruitment process.

Exhibit B.1. School Eligibility Criteria and Recruitment of Study Districts

School Eligibility Criteria

- Title I schools that include grades 4 and 5
- At least 30 percent of the school’s students were English learners
- At least 55 percent of the school’s students were eligible for free or reduced price lunch



SOURCES: School eligibility based on Common Core of Data (CCD) from school year 2014-2015 and the Civil Rights Data Collection databases (CRDC) from school year 2013-2014.

Random Assignment

In the summer and fall of 2017, prior to the start of program implementation, the team randomly assigned about half of the 70 recruited schools to adopt the academic language program and the other half to continue with business as usual. The purpose of the random assignment was to create two groups of schools similar to each other before the start of the program. Thus, all subsequent differences in outcomes between these two groups could be attributed to the program.

The team conducted random assignment within each of the six study districts. If the proportion of the English learner population in the schools varied substantially in a district, the team stratified the schools in the district based on this proportion and did random assignment within these strata, also called random assignment blocks. In one study district, the team first stratified the schools into two groups based on recruitment timing: schools that confirmed their participation earlier were blocked together and randomized early to facilitate the rollout of the training. The remaining schools in that district were randomized later. Within each of these two groups, the team further stratified the schools based on the proportion of English learners in the school. This stratification helped ensure that the proportion of English learners was similar in the program and non-program schools. In the end, across 11 random assignment blocks, the study team randomly assigned 36 schools to the program condition and 34 schools to the non-program condition.

However, after random assignment, some schools decided that they could not accommodate the evaluation's requirements and decided to withdraw from the study. By late fall of 2017, 12 schools (4 program schools and 8 non-program schools) withdrew from the study. These 12 schools withdrew from the evaluation after the random assignment and should therefore be considered attrition to the study. The overall school-level attrition rate is therefore 17.1 percent, with a differential attrition rate of 11.9 percent (p-value = 0.169).

Overall, the schools that left the study and those that remained do not appear to be systematically different from one another based on their background characteristics (see Exhibit B.2). On the other hand, the 58 schools that remained in the study differed from the national sample of regular public elementary schools during the 2016-2017 school year.

Exhibit B.2. Comparison of Schools Remaining in the Study and Schools That Left the Study

Characteristic	Remaining Schools	Schools That Left	Estimated Difference	P-Value for Estimated Difference
Geographic region (% of schools)				
Northeast	46.6	91.7	-45.1*	0.004
South	0.0	0.0	0.0	1.000
Midwest	10.3	8.3	2.0	0.836
West	43.1	0.0	43.1*	0.004
Urban character (% of schools)				
Large or middle-sized city	79.3	100.0	-20.7	0.086
Urban fringe and town	20.7	0.0	20.7	0.086
Title I status (% of schools)	98.3	88.4	9.9	0.088

(continued)

Exhibit B.2 (continued)

Characteristic	Remaining Schools	Schools That Left	Estimated Difference	P-Value for Estimated Difference
Race/ethnicity (% of students)				
Hispanic	67.6	59.0	8.6	0.233
Black, non-Hispanic	15.4	21.1	-5.7	0.052
White, non-Hispanic	9.0	8.5	0.5	0.864
Asian	6.5	8.8	-2.3	0.719
Other	2.2	3.4	-1.1	0.058
Students with special education status (% of students)	14.8	16.6	-1.8	0.307
English learners (% of students)	36.4	30.2	6.2	0.163
Students eligible for free or reduced-price lunch (% of students)	81.5	80.9	0.6	0.900
Female (% of students)	49.0	47.8	1.2	0.257
Total school enrollment (number of students)	507	586	-78	0.253
Enrollment in Grade 4 or Grade 5 (number of students)	143	163	-20	0.274
Students at or above proficiency level				
State ELA test (% of fourth- and fifth-graders)	22.9	28.4	-5.4	0.266
State math test (% of fourth- and fifth-graders)	20.3	29.0	-8.7	0.114
Number of schools	58	12		

SOURCES: The Common Core of Data from school year 2016-2017, Office for Civil Rights data from school year 2015-2016, state reported school performance data from school year 2016-2017.

NOTES: Rounding may cause slight discrepancies in calculating means and differences. Sample size for each characteristic may vary due to missing values.

ELA = English Language Arts.

A two-tailed t-test with a null of zero difference is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was used to determine whether there is a systematic difference between schools remaining in the study and schools that left, with respect to the characteristics included in this table. The p-value for this test is 0.516.

These 58 schools differed from the national sample in terms of geographic location and urbanicity (see Exhibit B.3). The study schools also had a higher proportion of English learners and students eligible for free/reduced-price lunch, which was not surprising given the intent of recruitment to target school districts with high proportions of these students. The study schools also had a higher proportion of Hispanic students, which was often the case in districts with high proportions of English learners.

Exhibit B.3. Comparison of Study Schools and Public Elementary Schools Nationally^a

Characteristic	Study Schools	Public Elementary Schools Nationally	Estimated Difference	P-Value for Estimated Difference
Geographic region (% of schools)				
Northeast	46.6	13.6	33.0 *	0.000
South	0.0	43.0	-43.0 *	0.000
Midwest	10.3	17.7	-7.4	0.142
West	43.1	25.7	17.4 *	0.002
Urban character (% of schools)				
Large or middle-sized city	79.3	25.6	53.7 *	0.000
Urban fringe and town	20.7	74.4	-53.7 *	0.000
Race/ethnicity (% of students)				
Hispanic	67.6	31.5	36.2 *	0.000
Black, non-Hispanic	15.4	18.5	-3.2	0.350
White, non-Hispanic	9.0	41.9	-32.9 *	0.000
Asian	6.5	3.6	2.9 *	0.009
Other	2.2	6.9	-4.6 *	0.011
Students with special education status (% of students)	14.8	12.9	1.9 *	0.040
English learners (% of students)	36.4	16.5	19.8 *	0.000
Students eligible for free or reduced-price lunch (% of students)	81.5	71.2	10.3 *	0.000
Female (% of students)	49.0	48.4	0.6	0.147
Total school enrollment (number of students)	507	475	32	0.297
Enrollment in Grade 4 or Grade 5 (number of students)	143	149	-6	0.620
Number of schools	58	24,776		

SOURCES: The Common Core of Data from school year 2016-2017, Office for Civil Rights data from school year 2015-2016.

NOTES: Rounding may cause slight discrepancies in calculating means and differences. Sample size for each characteristic may vary due to missing values.

A two-tailed t-test with a null of zero difference is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was used to determine whether there is a systematic difference between the study schools and the national sample, with respect to the characteristics included in this table. The p-value for this test is 0.992.

^aThe national sample includes all public regular elementary schools eligible for school-wide Title I, serving students in Grades 4 and 5, that are not charter, magnet, or virtual schools.

The analytic sample of 58 participating schools included 32 schools in the program group and 26 schools in the non-program group. Despite the differential attrition rate, the remaining program and non-program schools exhibited no systematic differences in a range of observed school characteristics, as shown in Exhibit B.4.

Exhibit B.4. Comparison of Program and Non-Program Schools in the Study

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	P-Value for Estimated Difference
Geographic region (% of schools)				
Northeast	46.9	46.2	0.7	0.957
South	0.0	0.0	0.0	1.000
Midwest	12.5	7.7	4.8	0.558
West	40.6	46.2	-5.5	0.679
Urban character (% of schools)				
Large or middle-sized city	84.4	73.1	11.3	0.299
Urban fringe and town	15.6	26.9	-11.3	0.299
Title I status	100.0	96.9	3.1	0.364
Race/ethnicity (% of students)				
Hispanic	68.5	66.9	1.6	0.734
Black, non-Hispanic	15.5	14.7	0.8	0.699
White, non-Hispanic	8.5	8.2	0.2	0.925
Asian	5.3	8.3	-3.0	0.507
Other	2.3	2.0	0.3	0.429
Students with special education status (% of students)	15.8	14.4	1.4	0.264
English learners (% of students)	36.6	36.0	0.6	0.862
Students eligible for free or reduced-price lunch (% of students)	74.2	69.0	5.2	0.162
Female (% of students)	49.5	48.6	1.0	0.276
Total school enrollment (number of students)	509	518	-9	0.860
Enrollment in Grade 4 or Grade 5 (number of students)	141	150	-9	0.446
Students at or above proficiency level				
State ELA test (% of fourth- and fifth-graders)	20.8	24.8	-4.1	0.223
State math test (% of fourth- and fifth-graders)	17.7	23.6	-6.0	0.096
Number of schools	32	26		

(continued)

Exhibit B.4 (continued)

SOURCES: Common Core of Data from school year 2016-2017, Office for Civil Rights data from school year 2015-2016, state reported school performance data from school year 2016-2017.

NOTES: This table is based on the 32 program schools and 26 non-program schools that participated in the study. The estimated differences are regression-adjusted, controlling for the blocking of random assignment. Rounding may cause slight discrepancies in calculating means and differences. Sample size for each characteristic may vary due to missing values.

ELA = English Language Arts.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.887.

II. Data-Collection Activities

The study team carried out multiple data-collection activities during the year when the program was implemented (the 2017-2018 school year, “program year” hereafter) and the year after the program implementation had ended (the 2018-2019 school year, “the follow-up year” hereafter). This section describes the main data-collection activities and the instruments used for these activities. The next section includes information on the measures constructed from these data sources.

Study-Administered Tests

The study team administered two tests to fourth- and fifth-grade students in the study schools in the spring of the program year. They were the Core Academic Language Skills Instrument (CALSA-I) test that measures student academic language skills and the reading comprehension portion of the Gates-MacGinitie reading test (GMRT). The study administered these two tests on consecutive days in each school. Due to scheduling challenges, one of the program schools was only able to accommodate the CALSA-I test but not the GMRT test before the end of the school year. As a result, all 58 study schools participated in the CALSA-I testing, while 57 schools participated in the GMRT testing. Five of the six study districts required active parental consent for student-level data collection and one study district allowed parents to opt out of the study. The active parental consent requirement has become the norm in school random assignment studies and has made it more challenging to obtain high response rates. In this study, among the five districts that required active parental consent, only 57 percent of the students who received a consent form returned it. Among those who returned the form, 78 percent consented to study participation. The overall consent rate across all six study districts was 48 percent.

The team did not administer the CALSA-I and GMRT tests to the study sample students at baseline for several reasons. First, many districts were resistant to the idea of having multiple rounds of additional tests for their students, and it would have made recruitment of study sites very difficult if the team insisted on administering such tests at baseline. Second, logistically, it would not have been possible for the team to finish the active consent process in five of the six study districts in time for the pretest given when the districts committed to study participation. Instead, the team used students’ state standardized reading test scores from the baseline year as measures of their reading performance levels prior to the program and found no systematic difference between the program and non-program schools on this measure.

District Records Data Collection

The study team collected district records data for the cohort of students in Grades 4 and 5 in the program year. The team collected information on these students’ background characteristics and state standardized test scores for ELA and math for three school years: the year before the program year, the program year, and the follow-up year. The team used student demographic and achievement information collected from the year before the program year as covariates

in impact estimations and used students' state ELA test scores collected for the program year and the follow-up year as outcomes for student reading achievement. Five of the six study districts provided records data for all Grade 4 and Grade 5 students; one district only provided these data for students with parental consent.

Classroom Observation

The study randomly selected three classrooms across Grade 4 and Grade 5 in each study school for two rounds of classroom observations: the first round took place between November 27, 2017 and February 7, 2018; the second round occurred between March 12, 2018 and June 20, 2018. The majority of the selected classrooms in both program and non-program schools were observed two separate times. However, many classrooms in one of the late-start districts were observed only in the spring of 2018. Certified observers completed two instruments during each 40-minute observation: the academic language instructional practice checklist developed for this study and the Classroom Assessment Scoring System-Upper Elementary (CLASS-UE) instrument.¹ The study team used data collected from classroom observations to capture teachers' use of program-specific instructional practices, practices generally considered important for developing academic language, and general classroom management quality.

Teacher Survey

The study team administered online surveys to all fourth- and fifth-grade teachers in the study schools in the fall/early winter of 2017 and again in the spring of 2018. The survey asked teachers in the program schools to describe their program-related training and coaching experience and directed teachers from both program and non-program schools to describe more general professional development activities that occurred immediately before and during the program year. The survey also collected information on teacher background characteristics such as educational attainment, teaching experience, and certification.

Training and Coaching Attendance Records

The study team captured information on initial coach and teacher training through coach attendance records and teacher attendance records for all program teachers collected for the initial training events. The team used this information to calculate the extent of training delivery and participation. The team also collected attendance records for teachers' and coaches' participation in the scheduled professional development sessions during the program year—at each scheduled coaching event for the coaches and at each scheduled professional learning session for the program teachers.

Provider and Coach Reports

The study team gathered information about the delivery of ongoing implementation support for coaches and teachers from program provider and coach reports. The provider submitted periodic reports throughout the year, and coaches reported the supports they received from the developer and those they provided to program teachers four separate times from the fall of 2017 through the early summer of 2018. Each of the 11 coaches completed all requested reports for a response rate of 100 percent.

Exhibit B.5 summarizes these data sources, the data obtained from them, their collection times, and the unit of measure for each source.

Exhibit B.5. Data-Collection Activities

Data Source	Data Obtained	Time of Data Collection	Unit of Measure (Respondent)
Data to measure effects on students			
Core Academic Language Skills Instrument (CALSI)	Student test scores	Spring 2018	Students
Gates-MacGinitie reading test (GMRT): reading comprehension part	Student test scores	Spring 2018	Students
District records	Student state test scores; student background characteristics	Fall of 2018 Fall of 2019	Students
Data to measure implementation in classes and schools			
Classroom observations	Teachers' use of program-specific practices (program schools only); teachers' use of practices important for developing academic language; Classroom Assessment Scoring System-Upper Elementary (CLASS-UE)	Round 1: November 2017-February 2018; Round 2: March-June 2018	Classrooms (study observer)
Teacher survey	Teachers' professional development training and support experience during the program year; teacher background characteristics	Round 1: November 2017-January 2018; Round 2: April-June 2018	Teachers
Teacher attendance data	Teachers' attendance at the Teachers' Institute and guidance and reflection sessions, for program-school teachers only	Teachers' Institute: Summer/Fall 2017 Guidance and reflection sessions: ongoing throughout the 2017-2018 school year	Teachers
Program provider and coach reports	Delivery and participation of training and ongoing support	Ongoing throughout 2017-2018 school year	Report (provider/coaches)

Exhibit B.6 presents the response rates for data sources used in the primary impact and implementation analyses. This exhibit shows that for the CALSI and GMRT tests, largely due to low rates of parental consent, the overall response rates were low for both the program and non-program group. The response rates were above 80 percent among students with parental consent, and the overall response rate did not differ by program status. The teacher survey response rate, however, differed between the program and non-program groups. Despite the program teachers' higher response rate, the program and non-program teachers who responded to the survey were similar to each other in their background characteristics (see Exhibit B.7).

Exhibit B.6. Response Rates for Data Sources Used to Estimate Program Effects

Measure (%)	All Study Schools	Program Schools	Non-Program Schools	Estimated Difference	P-Value for Estimated Difference
<u>Student outcomes</u>					
Study-administered tests					
Consent form returned ^a	56.7	56.5	59.6	-3.1	0.558
Consent rate among returned forms ^a	78.2	79.2	75.3	3.9	0.154
Overall consent rate	48.1	49.5	47.8	1.8	0.646
Response rate among students for whom consent was obtained					
CALS-I	85.2	83.1	88.2	-5.1	0.127
GMRT	83.3	81.0	88.1	-7.1	0.103
Overall response rate					
CALS-I	41.0	41.5	42.1	-0.5	0.887
GMRT	40.0	40.2	41.4	-1.2	0.755
State ELA test					
Program year	88.1	87.1	88.3	-1.3	0.403
Follow-up year	75.7	74.6	77.5	-2.9	0.167
<u>Classroom observations</u>					
First round (winter of the program year)	79.3	84.4	73.1	11.3	0.690
Second round (spring of the program year)	98.3	100.0	96.2	3.8	0.360
<u>Teacher surveys</u>					
First round (winter of the program year)	78.9	82.6	73.8	8.7 *	0.010
Second round (spring of the program year)	67.7	70.3	64.6	5.7 *	0.030

SOURCE: Authors' calculations based on student- and teacher/classroom-level data collected and compiled by the study team.

NOTES: This table is based on the 32 program schools and 26 non-program schools that participated in the study. The estimated differences are regression-adjusted, controlling for the blocking of random assignment. Rounding may cause slight discrepancies in calculating means and differences.

CALS-I = Core Academic Language Skills Instrument, ELA = English Language Arts, GMRT = Gates-MacGinitie Reading Test.

A two-tailed t-test with a null of zero difference is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

^aNumbers reported in this row only include information for the five study districts that required active parental consent.

Exhibit B.7. Background Characteristic Comparison of Teacher Survey Respondents in Program and Non-Program Schools

Characteristic	Program Schools	Non-Program Schools	Chi-Square or T-Value	P-Value for Estimated Difference
Highest degree attained (% of teachers)			0.44	0.933
Associate/bachelor's degree	24.1	27.3		
Master's degree	60.9	59.6		
Educational specialist or professional diploma	7.5	6.1		
Certificate of advanced studies/doctorate	7.5	7.1		
License, certificate, or endorsement to teach English learners (% of teachers)	39.9	51.5	3.12	0.077
Major field of study (% of teachers)			2.53	0.865
Elementary education	49.6	51.5		
Special education	8.3	9.1		
English as a Second Language or bilingual education	9.0	7.1		
Other education (for example, administration)	11.3	9.1		
English or language arts	11.3	8.1		
Social sciences	6.8	8.1		
Other	3.8	7.1		
Primary language (% of teachers)			0.10	0.949
English	48.2	48.2		
Spanish	41.0	39.8		
Other	10.8	12.0		
Teaching experience (% of teachers)				
Years worked as elementary or secondary-level teacher	12.2	12.4	-0.11	0.916
Years worked in current school	7.6	7.3	0.29	0.772
Years teaching fourth or fifth grade	6.3	6.5	-0.29	0.775
Number of teachers	156	130		

SOURCE: Teacher survey data collected in the spring of 2018.

NOTES: This table is based on teachers' response to the teacher survey in the spring of 2018. Sample size for each characteristic may vary due to missing values. A logistic regression model was conducted to see if there are systematic differences between the two groups, with the treatment condition as the outcome and the baseline characteristics variables as predictors, controlling for RA blocks. The p-value for the Chi-square test is 0.255.

* Indicates the estimated difference is statistically significant at the 0.05 level with a two-tailed test.

III. Analytic Approaches

This section presents the analytic approaches the team used to estimate the effects of the program. It starts by describing the key measures used in the evaluation. It then describes the analytical models and samples used for the impact estimation and approaches used for exploring the relationships among program implementation features and their effects on student outcomes.

Student Reading Performance and Classroom Instructional Practice Measures

This subsection provides information on the key measures used in this evaluation, including those for students’ language and reading performance, for teachers’ use of practices that are generally important for academic language development in classes, for teachers’ use of program-specific practices, and for overall classroom quality. It also provides information on the construction and reliabilities of these measures.

Academic Language Skills

The study used the Core Academic Language Skills Instrument (CALSI) to measure students’ academic language skills. The CALSI test addresses the high-utility language skills that correspond to linguistic features prevalent in oral and written academic discourse across school content areas and infrequent in everyday conversations. Examples of such features include: knowledge of logical connectives such as *nevertheless*, *consequently*; knowledge of structures that pack dense information such as nominalizations or embedded clauses; and knowledge of structures for organizing argumentative texts.² One benefit of using the CALSI is that it does not overly focus on academic vocabulary. It therefore contributes to the call from the field to expand the conceptualization of academic language as encompassing more than vocabulary knowledge alone.³ Exhibit B.8 below describes the domains and skills measured by the CALSI.

Exhibit B.8. Core Academic Language Domains and Skill Sets Measured by CALSI

Core Academic Language Domain	Task	Skill Measured
Word-Level Skills		
Tracking participants and ideas	Task 2. Tracking themes	Skill in identifying or producing the terms or phrases used to refer to the same participants or themes throughout an academic text (for example, <i>Water evaporates at 100 degrees Celsius. <u>This process...</u></i>)
Organizing analytic texts	Task 3. Organizing texts	Skill in organizing analytic texts, especially argumentative texts, according to the conventional academic (for example, <i>thesis, argument, counterargument, conclusion</i>) and paragraph-level structures (for example, <i>compare/contrast; problem/solution</i>)
Recognizing academic language	Task 6. Identifying definitions	Skill in recognizing more academic language when contrasted with more colloquial language in communicative contexts where academic language use is expected (for example, <i>more academic vs. more colloquial definitions of nouns</i>)

(continued)

Exhibit B.8 (continued)

Core Academic Language Domain	Task	Skill Measured
Sentence-Level Skills		
Connecting ideas logically	Task 1. Connecting ideas	Skill in comprehending and using “connectives” prevalent in academic texts to signal relationships between ideas (for example, <i>consequently, on the one hand...on the other hand</i>)
Unpacking/packing dense information	Task 5. Comprehending sentences (lower level)	Skill in comprehending and using complex words and complex sentences that facilitate concise communication (for example, nominalizations, embedded clauses, expanded noun phrases)
Understanding/expressing a writer's viewpoint	Task 7. Sure/unsure	Skill in understanding or using markers that signal a writer’s viewpoint, especially “epistemic stance markers,” those that signal a writer’s degree of certainty in relationship to a claim (for example, <i>Certainly, it is unlikely that</i>)
Discourse		
Unpacking/packing dense information	Task 4. Breaking words (higher level)	Skill in comprehending and using complex words and complex sentences that facilitate concise communication (for example, nominalizations, embedded clauses, expanded noun phrases)
Understanding/expressing metalinguistic vocabulary	Task 8. Understanding responses	Skill in understanding or expressing precise meanings, in particular, in using language to make thinking and reasoning visible, known as metalinguistic vocabulary (for example, <i>hypothesis, generalization, argument</i>)

The CALS-I test used in this study is a 45-minute paper and pencil test tailored for students in Grades 4 to 6. This form of the test has been shown to have a reliability measure of 0.93 and has scores on a vertically equated scale across these three grades.⁴ Two research papers documented that the CALS-I score is highly correlated with at least one state ELA assessment (Massachusetts), as well as with the Gates-MacGinitie reading comprehension test.⁵

Reading Comprehension Skill

The study used the reading comprehension part of the Gates-MacGinitie Reading Test (fourth edition) to measure students’ reading comprehension skills and serve as a common general reading measure across all study schools. This comprehension test generally takes about 45 minutes to administer with paper and pencil. It has 46 items measuring the ability to read and comprehend passages of a prose and simple verse nature. The test yields a single overall comprehension performance score. The Gates-MacGinitie is a nationally normed assessment. Normative scores were developed in 2005-2006 with a sampling plan based on geographic region, family income, enrollment size, parents’ years of schooling, and other factors. Studies have shown the reliability of the comprehension test to be above 0.85.⁶

English Language Arts Achievement

The study used students’ performance level on state standardized English language arts tests to measure their reading and language skills. Specifically, the team used the change in the percentage of students scoring at or above the state standard for proficiency in an average study school to measure the program’s effect on students’ broad skills. This

measure can provide useful information for understanding the program's effects on policy-relevant thresholds related to state proficiency expectations. The team also constructed an alternative measure using standardized state test scores to measure students' reading and language skills. The team standardized the scores from different state tests, converting them to z-scores. Specifically, within each state and grade, the team subtracted the non-program group mean from each student's test score and divided the result by the standard deviation of the non-program group. This standardization allowed the impact analysis to pool data across states with different state tests. Exhibit D.6 presents impact findings for this alternative measure.

Teachers' Use of Program-Specific Practices and Practices Considered Important for Developing Academic Language

Trained observers from the study team completed a multi-purpose checklist during classroom observation to capture instructional practices in the classroom. One section of the checklist contained 16 items that captured the degree to which teachers in program schools delivered the core instructional practices specific to the program as intended by the developer. The score from this section is called the fidelity score. The second section, with 15 items, measured the degree to which teachers in both the program and non-program schools delivered instructional practices emphasized by the program but not unique to its curriculum. The score from this section is called the alignment score.

For each item, teachers received a score of one if they used the practice specified in the item during the observation period. The team added up the item-level scores to get a total score and three sub-scores for each of the core instructional components for each list. For each observation round, each program-school teacher would receive a fidelity score (range = 0-16) for their use of program-specific practices, and each teacher from both program and non-program schools would receive an alignment score (range = 0-15) for their use of practices considered important for academic language development. The study team then averaged scores across observations for each teacher.⁷ In some cases, an observer could have credited a program teacher with an item on the alignment list but not an analogous item on the fidelity list because the teacher might have facilitated academic language development but might not have used the approach that was unique to the program. As a result, it was possible a program teacher could receive different scores for analogous items on the two lists.

The checklists were developed using a draft implementation instrument piloted by the program provider in prior implementations of the program. To adapt this tool for this evaluation, the study team frequently consulted the provider to ensure that the checklist included the core program components. The team deliberately constructed these items to be low inference so that observers could reliably complete this checklist and the CLASS-UE instrument simultaneously.

The study team further grouped the items on both checklists into three categories that correspond to the three core instructional components emphasized by the program: word knowledge instruction, academic skill instruction, and provision of practice opportunities. The groupings were constructed by reference to elements in the logic model, and reliability is estimated as the ratio of the variance of a teacher random effect to total variance (teacher effect variance plus residual variance). The team then calculated sub-scores for each component. Exhibit B.9 lists the items from both checklists by these categories and provides their corresponding reliability information.

Exhibit B.9. Checklist for Teacher Instructional Practices

	Use of WordGen (WG) program-specific practices (fidelity score, program schools only)		Use of practices that support academic language development (alignment score, program and non-program schools)	
Word Knowledge Instruction				
1	"Program Teacher introduced, reviewed, or called attention to the use of WG target words"	0,1	Teacher introduced, reviewed, or called attention to the use of vocabulary word(s)	0,1
2	"Program: WG target words were visually displayed or posted in classroom"	0,1	Vocabulary word(s) were visually displayed or posted in the classroom	0,1
3	"Program: Teacher referred to/prompted students to use a WG Word Study chart of target words"	0,1	Teacher referred to/prompted students to use visual display or graphic organizer of vocabulary word(s)/definitions (with skills listed)	0,1
4	At least 4 program-specific vocabulary words introduced	0,1		
Range of possible scores for Word Knowledge Instruction		0-4	Range of possible scores for Word Knowledge Instruction	0-3
Scale reliability for Word Knowledge Instruction		0.90	Scale reliability for Word Knowledge Instruction	0.67
Academic Skill Instruction				
5	"Program: Teacher introduced central WG question"	0,1		
6	"Program: Teacher introduced learning objective of the WG lesson or WG guiding question"	0,1	Teacher introduced learning objective of the lesson or guiding question	0,1
7	"Students followed norms or teacher reminded student of norms"	0,1	Students followed norms or teacher reminded students of norms	0,1
8	"Teacher reminded students that they must provide reasons and evidence"	0,1	Teacher reminded students that they must provide reasons and evidence to support their positions	0,1
9	"Closure was established by summarizing key positions or interesting exchanges"	0,1	Closure was established by summarizing key points or interesting observations	0,1

(continued)

Exhibit B.9 (continued)

	Use of WordGen (WG) program-specific practices (fidelity score, program schools only)		Use of practices that support academic language development (alignment score, program and non-program schools)	
	Range of possible scores for Academic Skill Instruction	0-5	Range of possible scores for Academic Skill Instruction	0-4
	Scale reliability for Academic Skill Instruction	0.80	Scale reliability for Academic Skill Instruction	0.66
Provision of Practice Opportunities				
10	"Program: Lesson delivered smoothly"	0,1		
11	"Program: All students had access to WG workbook"	0,1	All students had access to workbook, textbook, or other curricular material used for learning	0,1
12	"Students read text passage from program-specific workbook"	0,1	Teacher prompted students to read text passage (silently or aloud), or teacher read aloud text passage	0,1
13			Teacher instructed students to read independently (silently, without read-aloud support)	0,1
14			IF TEXT READ ALOUD: All students could see text while listening to read-aloud	0,1
15			Teacher modeled/prompted students to use comprehension strategies during reading	0,1
16	"Students participated in a debate or in a pre-debate or post-debate discussion"	0,1	Students participated in a classroom discussion	0,1
17	Debate, pre-debate, or post-debate focused on program-specific topic/question	0,1		
18	"Teacher asked two or more open-ended questions"	0,1	Teacher asked two or more open-ended questions	0,1
19	"At least three distinct positions were represented, reviewed, or discussed"	0,1	At least two opinions or viewpoints were represented, reviewed, or discussed	0,1
	Range of possible scores for Provision of Practice Opportunities	0-7	Range of possible scores for Provision of Practice Opportunities	0-8
	Scale reliability for Provision of Practice Opportunities	0.87	Scale reliability for Provision of Practice Opportunities	0.71
	Range of possible total scores for each observation	0-16	Range of possible total scores for each observation	0-15
	Scale reliability of total scores	0.97	Scale reliability of total scores	0.80

CLASS-UE Score

The study used the CLASS-UE instrument to capture the program-independent instructional quality in both program and non-program classrooms.⁸ The CLASS-UE is grounded in research on the relationships between instruction and student achievement and assesses the classroom environment for learning in three domains:

1. Emotional support: the extent to which teachers show responsiveness to students’ individual academic, social, and developmental needs.
2. Classroom organization: the extent to which the teacher manages both classroom routines and student behavior to maximize instructional time.
3. Instructional support: the extent to which teachers scaffold student learning by providing concrete, process-oriented feedback, engage students in higher-order thinking skills (such as problem-solving and analysis), and engage in instructional dialogue that extends students’ understanding of content and ability to apply concepts to novel contexts.

There is a wealth of psychometric data available for the CLASS-UE, from the Measures of Effective Teaching (MET) studies and other research.⁹ The CLASS-UE has exhibited strong psychometric properties across many grade levels.¹⁰ Exhibit B.10 lists the domains and dimensions included in this instrument. Note that “Student Engagement” is a stand-alone dimension that is not incorporated into any of the three domains.

Exhibit B.10. Classroom Assessment Scoring System-Upper Elementary (CLASS-UE) Domains and Dimensions

Domain	Dimension
Emotional support	Positive climate
	Teacher sensitivity
	Regard for adolescent perspectives
Classroom organization	Behavior management
	Productivity
	Negative climate
Instructional support	Instructional learning formats
	Content understanding
	Analysis and inquiry
	Quality of feedback
	Instructional dialogue
	Student engagement

Impact Estimation Approach

This section describes the statistical models used to estimate the effects of the academic language program on student and classroom outcomes.

Estimating Program Effects

The study used the following two-level hierarchical model to estimate the effects of the program on student outcomes.

$$Y_{ik} = \sum_m \alpha_{0m} B_{mik} + \sum_n \beta_n T_k D_{nik} + \sum_l \gamma_l X_{lik} + \mu_k + \omega_{ik} \quad (1)$$

Where

- Y_{ik} = Outcome measure for student i in school k ;
- B_{mik} = 1 if student i in school k is in random assignment block m , 0 otherwise;
- D_{nik} = 1 if student i in school k is in district n , 0 otherwise;
- T_k = 1 if school k is a participating school, 0 if it is a nonparticipating school;
- X_{lik} = l th covariate for student i with teacher j in school k ; and
- μ_k, ω_{ik} = School- and student-level random errors respectively assumed to be independently and identically distributed.

This model estimates separate program impacts for each district and then averages them across the districts, weighting each district’s estimate in proportion to the number of participating schools in the district. Therefore, these findings represent the effects of the program on the average program school in the study sample. Given that this study used a school-level random assignment design, it is preferable that the findings represent the program’s effect on the average program school in the sample. This approach provides an explicit way to achieve this goal.

Covariates: The model includes the random assignment block indicators as fixed effects to account for the blocked random assignment design. Wherever appropriate, grade indicators are also included as fixed effects to account for possible variation in outcome levels across grades. For the analysis on student outcomes, the model also includes standardized English language arts and math test scores from the baseline year, students’ age, race and ethnicity, gender, English learner status, special education status, and district-provided poverty indicators.

Missing Values: The program effect analyses did not include observations with missing outcome values. For missing covariate values, the team replaced the missing data with zeros and added an indicator for a given covariate’s missing status to the model. Research has demonstrated that this approach, known as the “dummy variable imputation method,” is unlikely to create estimation bias that is larger than 0.05 standard deviations in an experimental setting.¹¹

The team estimated the model using the PROC MIXED procedure in SAS. The estimated standard errors account for the clustering of students within schools. The reported program group average outcome values are the weighted average of unadjusted mean outcomes across districts, with the number of program schools in each district as weight. The non-program group average outcome values are calculated as the program group average minus the estimated average program effect.

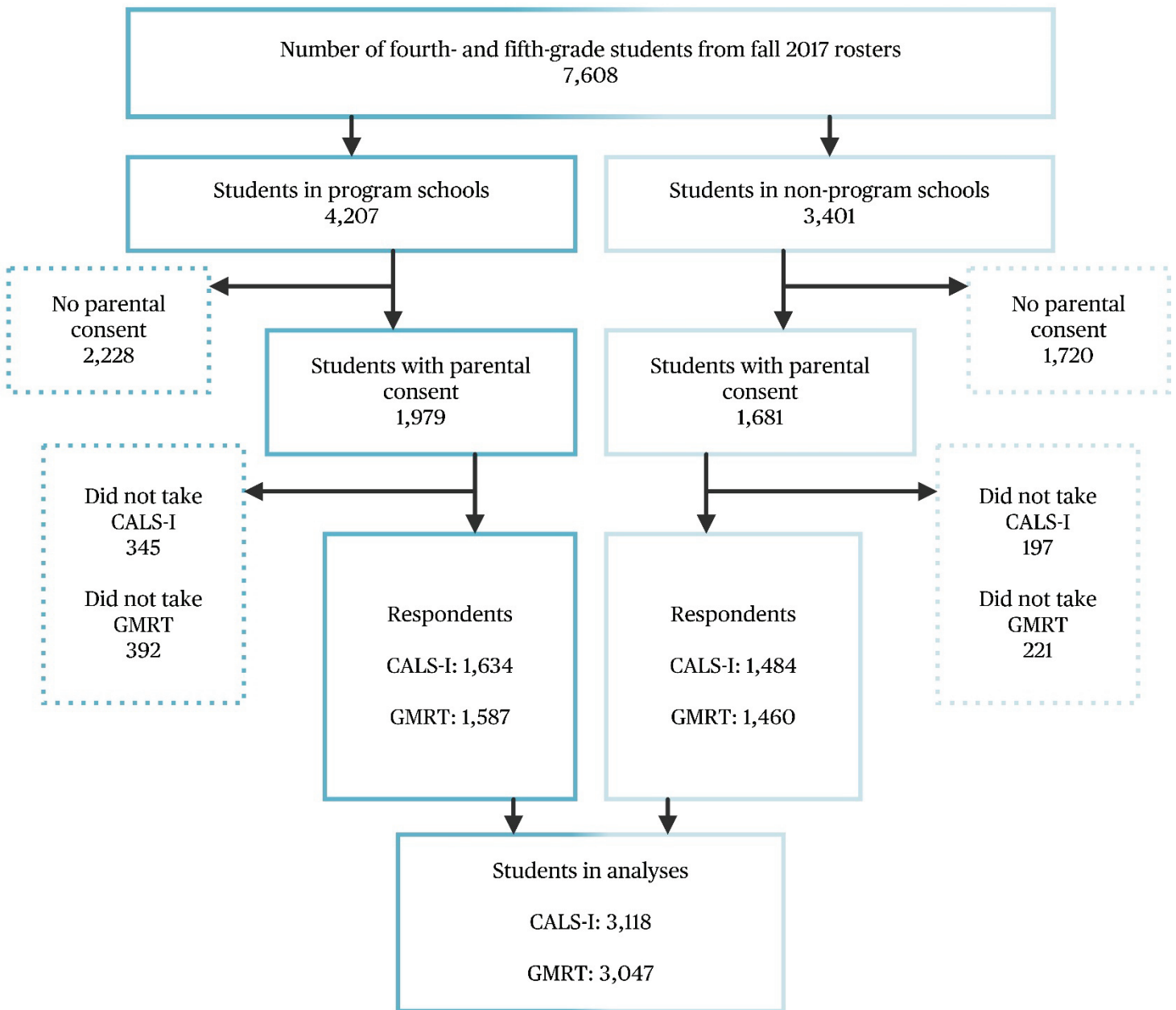
The team estimated this model separately for each student outcome from the program year and the follow-up year for the overall sample and for the key student subgroups. The team also used a fully interacted version of this model where each element in the model interacted with a subgroup indicator to test if the program impact varied by subgroup (see Appendix D for results). To estimate how the program affected teachers’ instructional practice alignment score, the team used a linear regression model similar to Equation (1) but accounts for the nesting of teachers within schools by using a cluster-robust standard error estimator.

Samples Used for Program Effect Estimation

In general, the impact estimation used all students with non-missing outcomes who were in Grades 4 and 5 in the study schools during the program year. This sample definition allowed the team to maximize the sample size and the statistical power of each analysis and improved the generalizability of the findings. Specifically, for the impact analyses on the

CALS-I and GMRT test scores, the team used all Grade 4 and Grade 5 students with parental consent and CALS-I or GMRT test scores from the spring of the program year. Exhibit B.11 shows the formation of these samples. For the impact analyses on the state ELA test scores in the program year, the team used all Grade 4 and Grade 5 students with non-missing state ELA test scores in that year. The follow-up year analysis used students with non-missing state ELA test scores from the spring of the follow-up year in Grades 4 or 5 in the study schools in the program year. Students in

Exhibit B.11. Student Sample Formation for Impact Estimation on CALS-I and GMRT Tests



SOURCE: Authors' calculations based on Core Academic Language Skills Instrument (CALS-I) and Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018 and compiled by the study team.

the program and non-program groups shared similar baseline characteristics across these samples. Exhibits B.12-B.23 present findings from such comparisons for each of the samples and relevant key subgroups listed below. Appendix D provides impact findings based on alternative sample definitions, which generally confirm the findings presented in the report.

Realized Minimum Detectable Effect

A common way to convey a study’s statistical power is through the minimum detectable effect (MDE) or the minimum detectable effect size (MDES). Formally, the MDE is the smallest true program impact that can be detected with a reasonable degree of power (in this study, 80 percent) for a given level of statistical significance (in this study, 5 percent for a two-tailed test). The MDES is the MDE scaled as an effect size. In other words, it is the MDE divided by the standard deviation of the unaffected outcome of interest (in this case, the non-program group standard deviation). Exhibit B.24 reports the realized values of the minimum detectable effects and the corresponding minimum detectable effect sizes for estimating program impacts on student outcomes based on the actual data and analytical approaches used in this study.

Exploring the Relationship Between Teacher Training, Program Implementation Features, and Student Outcomes

This section describes the analytic approach used to explore the relationship between teachers’ training and support, teachers’ use of instructional practices and student outcomes. It illustrates the analytic approach using the analysis of the relationship between teachers’ instructional practices and student outcome as an example.

The study used a hierarchical linear model (HLM) to analyze these relationships. Simply put, measures of teacher practices were added to the impact model in place of the treatment status indicator. The following modified model was used for this analysis:

$$Y_{ik} = \sum_m \alpha_{0m} B_{mik} + \beta TP_k + \sum_l \gamma_l X_{lik} + \mu_k + \omega_{ik} \quad (2)$$

Where TP_k is the teacher practice measure for school k , and all other variables are defined as in Equation (1). Here β is the conditional relationship between student outcomes and a given teacher/class-level intermediate measure, holding other covariates constant.

The teacher practice measures examined in this analysis included the scores constructed from the classroom observation checklist, including the total score and sub-scores for program teachers’ use of program-specific practices (the fidelity score) and the total and sub-scores for all teachers’ use of practices that were considered important to academic language development (the alignment score). Because the study did not observe all fourth- and fifth-grade classrooms in the study schools but rather randomly selected a set of classrooms for observation, these practice measures were not available for all classrooms. In addition, the observation data cannot be linked to individual students at classroom level. To use all available student outcome data in this analysis, the team aggregated these practice measures to the school level and merged them to the student data. Therefore, these measures were included as a school-level variable in Equation (2). This approach assumed that the observed teachers’ average practice was representative of the practice level at a given school, a reasonable assumption given that the team randomly sampled classrooms for observation.

The study conducted separate analyses for the program schools and for all study schools. To assess the relationship between teachers’ use of program-specific practices and student outcomes, the study estimated Equation (2) using data from the program schools only (see Exhibit C.8). To assess the relationship between teachers’ use of general practices important for academic language and student outcomes, the study estimated the model using data from all study schools (see Exhibit D.22).

Similar models were used to estimate the relationship between teachers’ reported amount of initial training and ongoing support and their fidelity scores (see Exhibit C.7).

Exhibit B.12. Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for CALS-I Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.90	8.91	-0.01	0.02	-0.02	0.481
Female (%)	53.1	49.6	3.4	2.06	0.07	0.102
Race/ethnicity (%)						
Hispanic	71.1	69.4	1.7	5.09	0.04	0.747
Black, non-Hispanic	12.0	11.6	0.4	2.14	0.01	0.863
White, non-Hispanic	8.8	7.9	1.0	2.46	0.03	0.691
Asian	6.6	9.2	-2.7	4.23	-0.10	0.532
Other	1.5	1.6	-0.1	0.57	-0.01	0.882
Students in Grade 4 in 2017-2018 (%)	55.2	49.1	6.1	3.31	0.12	0.074
Students with low-income status (%)	85.8	81.9	3.9	3.56	0.10	0.282
English learners (%)	27.4	31.3	-3.9	3.74	-0.09	0.298
Students with special education status (%)	9.8	10.2	-0.4	2.45	-0.01	0.877
Students meeting proficiency standards on state ELA test (%)	27.9	29.5	-1.5	4.40	-0.03	0.733
Average state ELA test standardized score	0.14	0.11	0.04	0.10	0.04	0.718
Students meeting proficiency standards on state math test (%)	29.2	29.3	0.0	5.24	0.00	0.997
Average state math test standardized score	0.14	0.13	0.01	0.12	0.01	0.950
Number of students ^a	1,634	1,484				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on all fourth- and fifth-grade students who were enrolled in the 58 study schools in the spring of 2018, had parental consent, and took the study-administered Core Academic Language Skills Instrument (CALS-I) test in the spring of 2018.

ELA = English Language Arts.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.426.

^aThe sample size reported here is for the full sample of students included in the estimation of program effect on CALS-I. Sample size for each characteristic may vary due to missing values.

Exhibit B.13. Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for CALS-I Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Impact
Age (years)	8.87	8.87	0.00	0.04	0.00	0.964
Female (%)	50.1	45.7	4.3	3.93	0.09	0.279
Race/ethnicity (%)						
Hispanic	81.5	79.1	2.5	6.60	0.06	0.710
Black, non-Hispanic	3.1	1.8	1.3	2.19	0.06	0.559
White, non-Hispanic	7.5	5.9	1.5	2.90	0.06	0.596
Asian	7.1	11.7	-4.6	5.43	-0.15	0.403
Other	0.8	0.6	0.2	0.96	0.02	0.841
Students in Grade 4 in 2017-2018 (%)	59.1	57.8	1.3	4.85	0.03	0.783
Students with low-income status (%)	92.7	87.4	5.3	3.52	0.16	0.138
Students with special education status (%)	14.4	13.5	0.9	3.52	0.02	0.798
Students meeting proficiency standards on state ELA test (%)	6.7	8.7	-2.0	2.97	-0.07	0.506
Average state ELA test standardized score	-0.46	-0.52	0.06	0.11	0.07	0.605
Students meeting proficiency standards on state math test (%)	11.8	12.2	-0.4	5.15	-0.01	0.937
Average state math test standardized score	-0.32	-0.34	0.03	0.13	0.03	0.833
Number of students	400	386				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade English learners who were enrolled in study schools in the spring of 2018, had parental consent, and took the study administered Core Academic Language Skills Instrument (CALS-I) test in the spring of 2018.

ELA = English Language Arts.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.801.

Exhibit B.14. Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for CALS-I Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.89	8.91	-0.02	0.02	-0.03	0.285
Female (%)	54.1	49.8	4.3	2.28	0.09	0.065
Race/ethnicity (%)						
Hispanic	73.9	72.7	1.2	5.11	0.03	0.812
Black, non-Hispanic	12.1	9.9	2.1	2.18	0.07	0.335
White, non-Hispanic	6.5	6.2	0.3	2.06	0.01	0.901
Asian	6.1	9.4	-3.3	4.40	-0.14	0.460
Other	1.4	1.5	-0.2	0.64	-0.01	0.808
Students in Grade 4 in 2017-2018 (%)	55.7	49.9	5.8	3.42	0.12	0.100
English learners (%)	30.2	33.8	-3.6	3.89	-0.08	0.365
Students with special education status (%)	10.4	11.0	-0.6	2.61	-0.02	0.820
Students meeting proficiency standards on state ELA test (%)	25.6	27.0	-1.4	4.22	-0.03	0.740
Average state ELA test standardized score	0.08	0.03	0.05	0.10	0.05	0.613
Students meeting proficiency standards on state math test (%)	26.3	26.9	-0.6	4.86	-0.01	0.904
Average state math test standardized score	0.06	0.07	-0.01	0.11	-0.01	0.946
Number of students	1,238	1,045				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade students from disadvantaged backgrounds who were enrolled in the 58 study schools in the spring of 2018, had parental consent, and took the study administered Core Academic Language Skills Instrument (CALS-I) test in the spring of 2018.

ELA = English Language Arts.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.471.

Exhibit B.15. Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for GMRT Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.92	8.93	-0.01	0.02	-0.02	0.536
Female (%)	52.6	49.8	2.8	2.09	0.06	0.188
Race/ethnicity (%)						
Hispanic	71.2	69.4	1.7	5.09	0.04	0.736
Black, non-Hispanic	12.0	11.9	0.0	2.27	0.00	0.994
White, non-Hispanic	9.2	8.1	1.1	2.51	0.04	0.663
Asian	6.6	8.8	-2.2	4.10	-0.08	0.599
Other	1.1	1.4	-0.3	0.54	-0.03	0.539
Students in Grade 4 in 2017-2018 (%)	54.0	49.3	4.7	3.48	0.09	0.182
Students with low-income status (%)	85.9	82.4	3.5	3.51	0.09	0.323
English learners (%)	26.7	31.8	-5.1	3.85	-0.11	0.194
Students with special education status (%)	9.5	9.1	0.4	2.04	0.01	0.858
Students meeting proficiency standards on state ELA test (%)	28.5	29.3	-0.8	4.32	-0.02	0.851
Average state ELA test standardized score	0.16	0.11	0.05	0.10	0.05	0.651
Students meeting proficiency standards on state math test (%)	29.4	28.8	0.6	5.31	0.01	0.912
Average state math test standardized score	0.15	0.14	0.01	0.12	0.01	0.904
Number of students ^a	1,587	1,460				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on all fourth- and fifth-grade students who were enrolled in the 58 study schools in the spring of 2018, had parental consent, and took the study administered Gates-MacGinitie Reading Test (GMRT) in the spring of 2018.

ELA = English Language Arts.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.459.

^aThe sample size reported here is for the full sample of students included in the estimation of program effect on GMRT. Sample size for each characteristic may vary due to missing values.

Exhibit B.16. Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for GMRT Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.86	8.88	-0.02	0.04	-0.03	0.591
Female (%)	49.4	46.3	3.2	3.93	0.06	0.427
Race/ethnicity (%)						
Hispanic	81.0	78.6	2.4	6.71	0.06	0.725
Black, non-Hispanic	3.2	2.1	1.1	2.34	0.05	0.641
White, non-Hispanic	7.8	6.2	1.6	2.91	0.06	0.577
Asian	7.5	11.3	-3.8	5.47	-0.12	0.494
Other	0.5	0.6	-0.1	1.11	-0.02	0.914
Students in Grade 4 in 2017-2018 (%)	59.4	58.3	1.1	4.99	0.02	0.831
Students with low-income status (%)	91.6	87.6	4.0	3.28	0.12	0.228
Students with special education status (%)	12.6	12.1	0.6	3.32	0.02	0.862
Students meeting proficiency standards on state ELA test (%)	7.4	7.8	-0.4	3.08	-0.01	0.901
Average state ELA test standardized score	-0.42	-0.52	0.09	0.10	0.11	0.368
Students meeting proficiency standards on state math test (%)	13.8	11.1	2.7	5.24	0.08	0.610
Average state math test standardized score	-0.27	-0.33	0.06	0.12	0.07	0.611
Number of students	382	379				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade English learners who were enrolled in study schools in the spring of 2018, had parental consent, and took the study administered Gates-MacGinitie Reading Test (GMRT) in the spring of 2018.

ELA = English Language Arts.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.809.

Exhibit B.17. Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for GMRT Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.91	8.93	-0.02	0.02	-0.02	0.425
Female (%)	53.5	49.7	3.8	2.30	0.08	0.109
Race/ethnicity (%)						
Hispanic	74.2	72.7	1.6	5.17	0.04	0.765
Black, non-Hispanic	11.7	10.2	1.4	2.22	0.05	0.520
White, non-Hispanic	6.9	6.5	0.4	2.09	0.02	0.834
Asian	6.1	8.8	-2.6	4.30	-0.11	0.542
Other	1.1	1.4	-0.3	0.58	-0.03	0.553
Students in Grade 4 in 2017-2018 (%)	54.3	50.2	4.1	3.54	0.08	0.257
English learners (%)	29.3	34.0	-4.7	3.95	-0.10	0.246
Students with special education status (%)	10.2	9.8	0.4	2.20	0.01	0.864
Students meeting proficiency standards on state ELA test (%)	26.2	27.1	-0.9	4.16	-0.02	0.834
Average state ELA test standardized score	0.10	0.05	0.05	0.10	0.05	0.594
Students meeting proficiency standards on state math test (%)	26.6	26.9	-0.2	4.93	-0.01	0.960
Average state math test standardized score	0.07	0.08	-0.01	0.11	-0.01	0.947
Number of students	1,205	1,028				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade students from disadvantaged backgrounds who were enrolled in the 58 study schools in the spring of 2018, had parental consent, and took the study administered Gates-MacGinitie Reading Test (GMRT) in the spring of 2018.

ELA = English Language Arts.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.489.

Exhibit B.18. Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for State ELA Test Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.99	8.99	0.00	0.02	0.00	0.950
Female (%)	50.2	47.3	2.9	1.56	0.06	0.071
Race/ethnicity (%)						
Hispanic	69.1	68.3	0.8	4.78	0.02	0.866
Black, non-Hispanic	15.1	13.6	1.5	2.21	0.04	0.493
White, non-Hispanic	8.2	7.7	0.5	2.25	0.02	0.823
Asian	6.0	8.9	-2.9	3.92	-0.11	0.466
Other	1.5	1.4	0.1	0.40	0.01	0.726
Students in Grade 4 in 2017-2018 (%)	51.2	50.0	1.2	1.73	0.02	0.495
Students with low-income status (%)	86.3	82.7	3.6	3.22	0.10	0.264
English learners (%)	32.5	34.9	-2.4	2.94	-0.05	0.417
Students with special education status (%)	13.2	13.0	0.2	1.78	0.01	0.916
Students meeting proficiency standards on state ELA test (%)	23.1	26.5	-3.4	3.49	-0.08	0.341
Average state ELA test standardized score	-0.03	0.01	-0.04	0.09	-0.04	0.670
Students meeting proficiency standards on state math test (%)	23.7	28.0	-4.3	4.17	-0.10	0.309
Average state math test standardized score	-0.04	0.06	-0.09	0.11	-0.10	0.383
Number of students ^a	3,984	3,468				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on all fourth- and fifth-grade students who were enrolled in the 58 study schools in the spring of 2018 and had a valid score for the state English Language Arts (ELA) test in the spring of 2018.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.694.

^aThe sample size reported here is for the full sample of students included in the estimation of program effect on the state ELA test in the program year. Sample size for each characteristic may vary due to missing values.

Exhibit B.19. Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for State ELA Test Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.96	8.96	0.00	0.03	0.01	0.908
Female (%)	45.5	44.3	1.2	2.63	0.02	0.643
Race/ethnicity (%)						
Hispanic	82.1	81.6	0.5	5.98	0.01	0.928
Black, non-Hispanic	2.8	1.9	0.9	1.40	0.05	0.513
White, non-Hispanic	6.7	5.4	1.3	2.72	0.05	0.641
Asian	7.7	10.3	-2.5	4.95	-0.08	0.610
Other	0.7	0.8	-0.2	1.35	-0.02	0.906
Students in Grade 4 in 2017-2018 (%)	55.5	56.4	-0.9	3.59	-0.02	0.808
Students with low-income status (%)	90.0	89.7	0.4	2.99	0.01	0.907
Students with special education status (%)	16.2	18.0	-1.8	2.51	-0.05	0.476
Students meeting proficiency standards on state ELA test (%)	6.5	6.5	0.0	2.39	0.00	0.993
Average state ELA test standardized score	-0.56	-0.56	0.00	0.10	0.00	0.983
Students meeting proficiency standards on state math test (%)	10.3	13.0	-2.7	3.68	-0.08	0.469
Average state math test standardized score	-0.43	-0.38	-0.05	0.11	-0.06	0.650
Number of students	1,041	937				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade English learners who were enrolled in study schools in the spring of 2018 and had a valid score for the state English Language Arts (ELA) test in the spring of 2018.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.948.

Exhibit B.20. Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for State ELA Test Analysis, Program Year (2017-2018)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.99	9.00	0.00	0.02	0.00	0.896
Female (%)	51.1	47.9	3.2	1.90	0.06	0.103
Race/ethnicity (%)						
Hispanic	72.2	71.3	0.8	4.74	0.02	0.860
Black, non-Hispanic	14.6	12.3	2.3	2.24	0.06	0.313
White, non-Hispanic	6.0	6.0	0.0	1.97	0.00	0.992
Asian	5.7	8.8	-3.1	4.02	-0.13	0.445
Other	1.5	1.4	0.1	0.47	0.01	0.867
Students in Grade 4 in 2017-2018 (%)	51.3	50.0	1.3	1.80	0.03	0.462
English learners (%)	34.4	37.4	-3.1	2.86	-0.06	0.290
Students with special education status (%)	13.8	13.8	-0.1	1.81	0.00	0.977
Students meeting proficiency standards on state ELA test (%)	20.3	23.0	-2.7	3.05	-0.06	0.384
Average state ELA test standardized score	-0.09	-0.07	-0.02	0.08	-0.02	0.809
Students meeting proficiency standards on state math test (%)	21.3	24.5	-3.2	3.87	-0.08	0.409
Average state math test standardized score	-0.11	-0.02	-0.09	0.10	-0.09	0.382
Number of students	3,253	2,715				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade students from disadvantaged backgrounds who were enrolled in the 58 study schools in the spring of 2018 and had a valid score for the state English Language Arts (ELA) test in the spring of 2018.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.979.

Exhibit B.21. Comparison of Background Characteristics of All Students in Program and Non-Program Study Schools for State ELA Test Analysis, Follow-Up Year (2018-2019)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.96	8.95	0.01	0.02	0.01	0.710
Female (%)	50.7	47.6	3.1	1.62	0.06	0.063
Race/ethnicity (%)						
Hispanic	68.9	68.7	0.2	4.81	0.00	0.961
Black, non-Hispanic	15.4	13.2	2.2	2.18	0.06	0.314
White, non-Hispanic	8.1	7.6	0.5	2.25	0.02	0.832
Asian	6.3	9.0	-2.8	3.91	-0.10	0.481
Other	1.3	1.3	0.0	0.40	0.00	0.943
Students in Grade 4 in 2017-2018 (%)	53.0	51.3	1.8	1.91	0.04	0.358
Students with low-income status (%)	86.3	81.7	4.6	3.12	0.12	0.147
English learners (%)	33.1	35.4	-2.3	3.27	-0.05	0.483
Students with special education status (%)	13.9	12.7	1.2	1.88	0.03	0.536
Students meeting proficiency standards on state ELA test (%)	23.2	26.8	-3.6	3.42	-0.08	0.298
Average state ELA test standardized score	-0.02	0.02	-0.05	0.09	-0.05	0.623
Students meeting proficiency standards on state math test (%)	24.0	28.0	-3.9	4.38	-0.09	0.375
Average state math test standardized score	-0.04	0.06	-0.10	0.11	-0.10	0.390
Number of students ^a	3,420	2,982				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on all fourth- and fifth-grade students who were enrolled in the 58 study schools in the spring of 2018 and had a valid score for the state English Language Arts (ELA) test in the spring of 2019.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.911.

^aThe sample size reported here is for the full sample of students included in the estimation of program effect on the state ELA test in the follow-up year. Sample size for each characteristic may vary due to missing values.

Exhibit B.22. Comparison of Background Characteristics of English Learners in Program and Non-Program Study Schools for State ELA Test Analysis, Follow-Up Year (2018-2019)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.97	8.96	0.01	0.03	0.02	0.726
Female (%)	46.5	43.3	3.2	2.74	0.06	0.247
Race/ethnicity (%)						
Hispanic	81.8	80.7	1.1	5.91	0.03	0.857
Black, non-Hispanic	2.7	2.1	0.6	1.37	0.03	0.680
White, non-Hispanic	6.6	6.4	0.1	2.74	0.01	0.959
Asian	8.3	10.5	-2.2	5.00	-0.07	0.666
Other	0.6	-0.2	0.9	0.86	0.13	0.317
Students in Grade 4 in 2017-2018 (%)	55.2	59.6	-4.4	3.81	-0.09	0.254
Students with low-income status (%)	89.9	87.8	2.1	3.12	0.06	0.509
Students with special education status (%)	17.4	17.7	-0.2	2.76	-0.01	0.937
Students meeting proficiency standards on state ELA test (%)	6.5	7.0	-0.4	2.53	-0.02	0.863
Average state ELA test standardized score	-0.56	-0.54	-0.03	0.10	-0.03	0.799
Students meeting proficiency standards on state math test (%)	11.2	13.7	-2.5	3.93	-0.07	0.525
Average state math test standardized score	-0.43	-0.35	-0.08	0.11	-0.09	0.481
Number of students	963	822				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade English learners who were enrolled in study schools in the spring of 2018 and had a valid score for the state English Language Arts (ELA) test in the spring of 2019.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.884.

Exhibit B.23. Comparison of Background Characteristics of Students from Disadvantaged Backgrounds in Program and Non-Program Study Schools for State ELA Test Analysis, Follow-Up Year (2018-2019)

Characteristic	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.96	8.96	0.01	0.02	0.01	0.780
Female (%)	51.5	48.5	3.1	1.91	0.06	0.117
Race/ethnicity (%)						
Hispanic	71.6	71.8	-0.2	4.77	0.00	0.970
Black, non-Hispanic	15.2	12.2	3.0	2.22	0.08	0.186
White, non-Hispanic	6.2	5.7	0.5	1.95	0.02	0.804
Asian	5.8	8.9	-3.1	4.02	-0.12	0.444
Other	1.2	1.3	-0.1	0.49	-0.01	0.869
Students in Grade 4 in 2017-2018 (%)	53.2	51.4	1.7	2.07	0.03	0.407
English learners (%)	34.9	37.7	-2.7	3.19	-0.06	0.396
Students with special education status (%)	14.6	13.6	0.9	1.93	0.03	0.631
Students meeting proficiency standards on state ELA test (%)	20.3	23.6	-3.3	3.05	-0.08	0.286
Average state ELA test standardized score	-0.09	-0.05	-0.03	0.08	-0.03	0.704
Students meeting proficiency standards on state math test (%)	21.6	24.6	-3.0	4.08	-0.07	0.464
Average state math test standardized score	-0.11	-0.02	-0.09	0.10	-0.10	0.384
Number of students	2,745	2,303				

SOURCE: 2016-2017 school records data obtained for this study.

NOTES: This table is based on fourth- and fifth-grade students from disadvantaged backgrounds who were enrolled in the 58 study schools in the spring of 2018 and had a valid score for the state English Language Arts (ELA) test in the spring of 2019.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was used to determine whether there is a systematic difference between the students in the program schools and the non-program schools, with respect to the characteristics included in this table. The p-value for this test is 0.607.

Exhibit B.24. Realized Minimum Detectable Effects by Outcome and Sample

Outcome	Minimum Detectable Effect Size	Minimum Detectable Effect (MDE)	Unit of Measure for MDE
Full sample			
CALS-I	0.13	4.31	Scaled score
GMRT	0.12	4.91	Scaled score
State ELA test proficiency			
Program year	0.15	6.9	Percentage
Follow-up year	0.15	6.9	Percentage
English learners			
CALS-I	0.16	5.30	Scaled score
GMRT	0.19	7.50	Scaled score
State ELA test proficiency			
Program year	0.16	7.3	Percentage
Follow-up year	0.16	7.8	Percentage
Students from economically disadvantaged backgrounds			
CALS-I	0.13	4.41	Scaled score
GMRT	0.13	5.36	Scaled score
State ELA test proficiency			
Program year	0.15	6.8	Percentage
Follow-up year	0.14	6.7	Percentage

SOURCES: Authors' calculations based on Core Academic Language Skills Instrument (CALS-I) and Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018, and district records data collected for the 2016-2017, 2017-2018, and 2018-2019 school years.

NOTES: The minimum detectable effect sizes in this table are calculated by multiplying the standard error of the estimated effects by 2.8 and dividing by the standard deviations of students' spring 2018 test scores in the non-program schools, assuming a statistical significance level of 0.05.

ELA = English Language Arts.

Full sample non-program group student-level CALS-I standard deviation = 32.76; GMRT standard deviation = 39.99; state ELA test percentage proficient standard deviation = 46.16.

It is important to note that this analysis is correlational rather than experimental, so any observed relationships between teacher training, teacher practices, and student outcomes might be due to the effects of unobserved factors that happen to be correlated with teacher practices rather than true causal effects. It is also important to note that conducting multiple hypothesis tests on the associations between different outcomes and explanatory variables increases the likelihood of concluding that a given estimated relationship is statistically significant, when in fact such association does not exist (this is known as a type I error or a false positive). In particular, one would expect to see one false positive for every 20 hypothesis tests conducted when $p < 0.05$ is selected as the criterion for statistical significance. Therefore, findings from these correlational analyses are exploratory and are for hypothesis-generating purposes only. They should be interpreted with caution.

APPENDIX C. SUPPLEMENTAL INFORMATION ON FINDINGS IN THE REPORT

This appendix provides additional details on the study findings presented in the report. It starts with information on the estimated program impacts on student outcomes and classroom instructional practices. It then shows supplemental information on the implementation of training and ongoing supports for teachers. Lastly, it reports additional information that a systematic review might need to assess the impact findings for student outcomes.

I. Additional Details on Program Impact Findings

This section provides supplementary information on the program impact findings presented in the report. These findings include details of the estimated program impacts on students' reading outcomes, teachers' use of core instructional practices in classrooms, and general classroom management quality. It also provides information on the exploratory analysis of the relationship between instructional practices and student outcomes.

Program Impacts on Student Language and Reading Outcomes

Exhibit 3 in the report presents the estimated program effects on students' test performances on the Core Academic Language Skills Instrument (CALIS-I), the Gates-MacGinitie Reading Test (GMRT), and the state standardized English language arts (ELA) test for fourth- and fifth-graders in the study schools. Exhibit C.1 presents more details of these findings, including the corresponding standard errors, confidence intervals, and p-values for each impact estimate.

The report also mentions that the estimated program impacts do not vary significantly across the six study districts. Exhibit C.2 presents the magnitudes and confidence intervals of each study district's impact estimates for the three key student outcomes. These figures show that even though the impact estimates' magnitudes appear to vary from district to district, their confidence intervals largely overlap. This indicates that the district-level impact estimates cannot be statistically distinguished from each other.

The report presents program impacts on students' language and reading outcomes for two groups of students who might stand to benefit from the program: the English learners and students from economically disadvantaged backgrounds (Exhibits 4 and 5). The findings generally showed that the program did not produce any effects on the reading performance of these two groups of students. Exhibit C.3 provides details of the findings for these two groups of students.

The study tracked the cohort of fourth- and fifth-graders enrolled in the study schools in the program year (2017-2018) for an additional year (2018-2019), after the program-provided training and support had ended. Most of these students were in fifth and sixth grades in the follow-up year. The team obtained about 76 percent of these students' scores on the state standardized reading tests in the follow-up year (see Exhibit B.6). The estimated program impacts on state reading test performance were not different from zero for the overall sample and the two key subgroups (Exhibit C.4).

Program Impacts on Classroom Outcomes

Exhibit 8 in the report illustrates that the program expanded teachers' use of core instructional practices that are important for academic language development. However, the increase mostly came from the increased use of word knowledge instruction. Exhibit C.5 presents details of the estimated program effects on teachers' use of instructional practices in the classrooms. In particular, it shows the number of specific practices teachers used during the classroom observation period overall and in each core practice category: word knowledge instruction, academic skill instruction, and provision of practice opportunities. It also converted these counts into the coverage rate (percentage of practices/items covered) for all practices and for each category separately.

To provide context for the general classroom environment, the team also collected information on general classroom management and teaching practices during classroom observations using the CLASS-UE instrument. Appendix B described the measures from this instrument. The team found no difference between the program and non-program school classrooms across all CLASS-UE measures (Exhibit C.6).

Exhibit C.1. Estimated Impacts on Student Language and Reading Outcomes, Overall Sample, Program Year

Outcome	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	95 Percent Confidence Interval	
							Lower Bound	Upper Bound
CALS-I scaled score	496.05	498.04	-1.99	1.54	-0.06	0.204	-5.10	1.12
GMRT scaled score	478.27	481.29	-3.02	1.75	-0.08	0.092	-6.56	0.52
GMRT grade equivalence	3.90	4.10						
Percentage of students meeting proficiency standards on state ELA test	26.5	28.0	-1.5	2.46	-0.03	0.546	-6.5	3.5
Number of schools	32	26						

SOURCES: Core Academic Language Skills Instrument (CALS-I) data (sample size = 3,118 students) and the Gates-MacGinitie Reading Test (GMRT) data (sample size = 3,047 students) collected in the spring of 2018. State standardized English Language Arts (ELA) test data (sample size = 7,452 students) and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: The student sample includes all students with valid outcome measures.

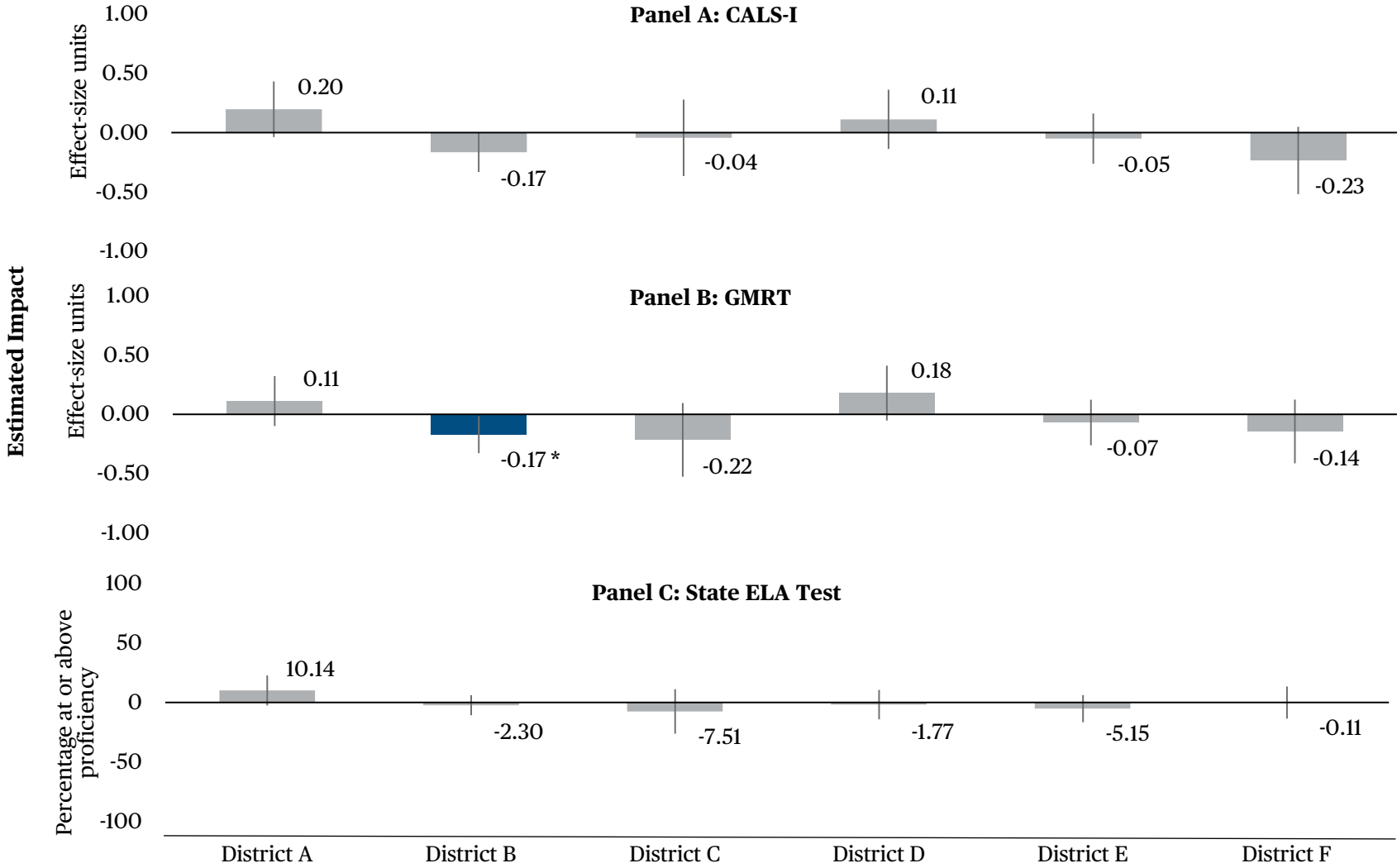
The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, special education status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted averages of the observed district means for students from the program schools (using the number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit C.2. Estimated Impacts on Student Language and Reading Outcomes, by District, Program Year



(continued)

Exhibit C.2 (continued)

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data (sample size = 3,118) and the Gates-MacGinitie Reading Test (GMRT) data (sample size = 3,047) collected in the spring of 2018. State standardized English Language Arts (ELA) test data (sample size = 7,452) and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: The student sample includes all students with valid outcome measures.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school members in the analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

Exhibit C.3. Estimated Impacts on Student Language and Reading Outcomes for English Learners and Students from Economically Disadvantaged Backgrounds, Program Year

Outcome	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	95 Percent Confidence Interval	
							Lower Bound	Upper Bound
English learners								
CALS-I scaled score	480.88	481.58	-0.71	1.89	-0.02	0.711	-4.54	3.13
GMRT scaled score	462.72	465.25	-2.53	2.68	-0.06	0.352	-7.96	2.91
GMRT grade equivalence	3.37	3.43						
Percentage of students meeting proficiency standards on state ELA test	11.3	10.3	1.0	2.62	0.02	0.702	-4.3	6.3
Number of schools	28	21						
Students from economically disadvantaged backgrounds								
CALS-I scaled score	495.68	496.85	-1.17	1.57	-0.04	0.460	-4.35	2.00
GMRT scaled score	478.14	480.06	-1.92	1.91	-0.05	0.321	-5.78	1.94
GMRT grade equivalence	3.90	4.05						
Percentage of students meeting proficiency standards on state ELA test	25.4	25.8	-0.5	2.44	-0.01	0.845	-5.4	4.4
Number of schools	32	26						

(continued)

Exhibit C.3 (continued)

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: English learners are identified by their status in the 2016-2017 school year. The sample includes 786 students with CALIS-I scores, 761 students with GMRT scores, and 1,978 students with state ELA scores. Schools from one district are not included in this subgroup analysis because no English learners participated in the program. Students from economically disadvantaged backgrounds are identified by their family income status in the 2016-2017 school year. The sample includes 2,283 students with CALIS-I scores, 2,233 students with GMRT scores, and 5,968 students with state ELA scores.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample. None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit C.4. Estimated Impacts on Students' Performances in State English Language Arts Tests in the Follow-Up Year, Overall Sample, English Learners, and Students from Economically Disadvantaged Backgrounds

Sample	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	95% Confidence Interval	
							Lower Bound	Upper Bound
Full sample	26.9	28.2	-1.3	2.48	-0.03	0.614	-6.3	3.7
English learners	12.8	10.8	1.9	2.78	0.04	0.490	-3.7	7.6
Students from economically disadvantaged backgrounds	26.5	26.1	0.4	2.40	0.01	0.883	-4.5	5.2

SOURCES: State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2018-2019 school years.

NOTES: The overall sample includes 6,405 students with state ELA test scores for the 2018-2019 school year. English learners (1,785 students) are identified by their status in the 2016-2017 school year. Schools from one district are not included in this subgroup analysis because no English learners participated in the program. Students from economically disadvantaged backgrounds (5,051 students) are identified by their family income status in the 2016-2017 school year.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit C.5. Estimated Impacts on the Use of Instructional Practices Important for Academic Language Development

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact
Word knowledge instruction (item coverage, range = 0-3)	2.00	1.41	0.59*	0.15	0.64	0.000
Percentage of items covered	66.8	47.0				
Academic skill instruction (item coverage, range = 0-4)	2.52	2.43	0.09	0.13	0.08	0.489
Percentage of items covered	63.0	60.8				
Provision of practice opportunities (item coverage, range = 0-8)	5.60	5.47	0.13	0.25	0.07	0.595
Percentage of items covered	70.0	68.4				
Total (item coverage, range = 0-15)	10.13	9.31	0.82*	0.39	0.27	0.040
Percentage of items covered	67.5	62.1				
Number of schools	32	25				
Number of classrooms	93	72				

SOURCE: Classroom observation data collected in the 2017-2018 school year.

NOTES: The sample includes 165 fourth- and fifth-grade regular classrooms from 57 study schools.

The impacts are estimated using linear models that account for the nested structure of the data, with classrooms nested within schools. The models control for the blocking of random assignment.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

Exhibit C.6. Estimated Impacts on General Classroom Management Quality, as Measured by Classroom Assessment Scoring System-Upper Elementary (CLASS UE)

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact
Emotional support	4.51	4.53	-0.03	0.11	-0.04	0.828
Classroom organization	5.76	5.82	-0.07	0.12	-0.10	0.567
Instructional support	3.63	3.56	0.07	0.12	0.08	0.548
Student engagement	4.86	4.77	0.09	0.15	0.10	0.550
Overall	4.69	4.67	0.02	0.11	0.02	0.881
Number of schools	32	25				
Number of classrooms	91	71				

SOURCE: Classroom observation data collected in the 2017-2018 school year using the CLASS-UE instrument.

NOTES: The sample includes 162 fourth- and fifth-grade regular classrooms from 57 study schools.

The impacts are estimated using linear models that account for the nested structure of the data, with classrooms nested within schools. The models control for the blocking of random assignment.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

II. Relationship Between Teacher Training, Instructional Practices, and Student Outcomes

To learn more about the mechanisms through which the program might affect student outcomes, the study examined the relationship between the amount of training teachers received and their use of the core instructional practices, and between teachers' use of instructional practices that are aligned with the program and student outcome. As described in Appendix B, this kind of analysis assessed whether more usage of such practices is associated with better student outcomes, but the findings do not imply causality and should be interpreted with caution.

The program's theory of change suggests that the program intended to provide teachers with training and support that encourage and facilitate their use of certain instructional practices aligned with academic language learning. In turn, teachers' use of such practices could lead to better language and reading performances for students (see Exhibit A.1). The correlational analysis shows that more training and ongoing support for academic language instruction received by the program-school teachers was only positively associated with teachers' use of word knowledge instruction. It was not consistently associated with the other two core components of instruction promoted by the program: instructions on academic skills or provision of opportunities for student practice (see Exhibit C.7). The analysis further shows that more use of these practices overall was positively associated with students' CALS-I score (see Exhibit C.8). Such a positive correlation appeared to be driven by the positive associations between academic skill instruction and provision of practice opportunities and CALS-I scores. The association between word knowledge instruction and CALS-I scores is on the cusp of being statistically significant with a p-value of 0.053, just over the 0.05 threshold. The links between word knowledge instruction and other student outcomes were not statistically significant (Exhibit C.8).

III. Additional Details on Program Implementation

As described in the program's theory of change (see Exhibit A.1), the implementation of the program included teachers' participation in the program-provided teacher training and ongoing support, as well as teachers' delivery of the curricular units in their classrooms. This section provides more detailed information about these two aspects of program implementation and presents the facilitators and challenges to implementation as reported by program providers and coaches.

Delivery and Attendance of Training and Ongoing Support for Teachers

As discussed in the report, the training and support provided to and received by the program-school teachers were less than what had been planned (Exhibit 7). Data from attendance records and teacher surveys support this finding.

Initial Training: The report shows that 35 percent of teachers received two days, or 16 hours, of initial training as planned initially. Specifically, in four study districts that started the program on time at the beginning of the program year, the provider delivered two days of training, and about 74 percent of teachers participated in both days of training. In two districts that started the program late in the fall of the program year, the program provided an adapted single day of condensed training that covered the same topics, and about 87 percent of teachers in these two districts participated in the training (Exhibit C.9).

Guidance Sessions: The program planned to provide six "guidance sessions" to teachers. These sessions were intended to introduce teachers to basic principles (for example, reasoning, argumentation) that were integral to the program's core instructional components and were designed to prepare teachers to deliver the next two curricular units. The program provider was able to offer guidance sessions to all study districts. Teachers' attendance rates varied by district and by session. The average teacher attendance rate was 61 percent across the six guidance sessions, and the attendance rate for a given session ranged from 41 percent to 87 percent (Exhibit C.10).

Exhibit C.7. Associations Between Training and Support and Teachers' Use of Program-Specific Instructional Practices, Program Schools

Measure	Total		Word Knowledge Instruction		Academic Skill Instruction		Provision of Practice Opportunities	
	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value
Teacher-reported training								
Initial training total hours (number of hours)	0.00	0.901	0.02	0.028*	-0.01	0.286	-0.01	0.044*
Ongoing support total score (range = 0-9)	0.20	0.294	0.19	0.035*	0.06	0.565	-0.04	0.618

SOURCES: Classroom observation data collected in the 2017-2018 school year. Teacher survey data collected in the 2017-2018 school year.

NOTES: The sample includes teachers in 32 program schools who responded to the teacher survey and answered questions about the amount of initial training and ongoing support they received during the program year. The correlations between the amount of training and support and the fidelity scores are estimated with a multilevel regression with teachers nested within schools. The model also controls for the blocking of random assignment.

The estimated correlation reflects the amount of fidelity score change that is associated with one hour of change in initial training or one unit of change in ongoing support. A two-tailed t-test with a null of zero correlation is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

Exhibit C.8. Associations Between Teachers' Use of Program-Specific Instructional Practices and Student Outcomes, Program Schools

	Academic Language Skills (CALS-I score)		Reading Comprehension Skills (GMRT score)		Reading Achievement (% at or Above State Proficiency Level)	
	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value
Total	1.68	0.004*	0.57	0.454	1.22	0.122
Word knowledge instruction	2.67	0.053	0.95	0.578	3.30	0.129
Academic skill instruction	3.56	0.008*	1.68	0.239	1.09	0.609
Provision of practice opportunities	3.08	0.022*	0.46	0.788	2.19	0.383

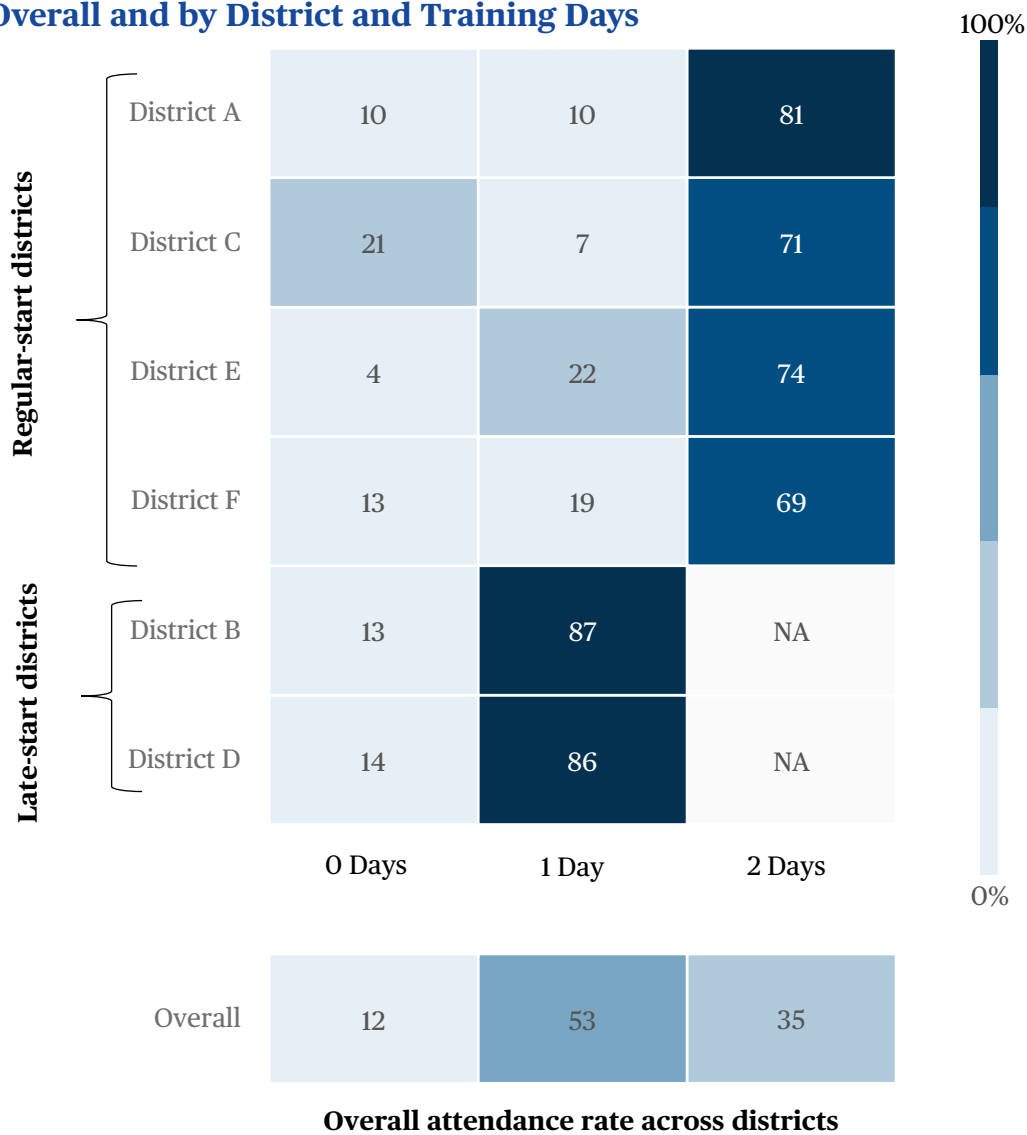
SOURCES: Classroom observation data collected in the 2017-2018 school year. Core Academic Language Skills Instrument (CALS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: The sample used in this analysis includes all fourth- and fifth-graders with nonmissing values for respective outcomes and all explanatory variables in a given model. The correlations between the explanatory variables and each outcome are estimated with a multilevel, multivariate regression with students nested within schools. The regression models also control for the blocking of random assignment and for student background characteristics such as grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The estimated correlation reflects the amount of outcome change (in the unit of the outcome measure) that is associated with one unit of change in a given explanatory variable, controlling for all covariates in the model.

A two-tailed t-test with a null of zero correlation is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

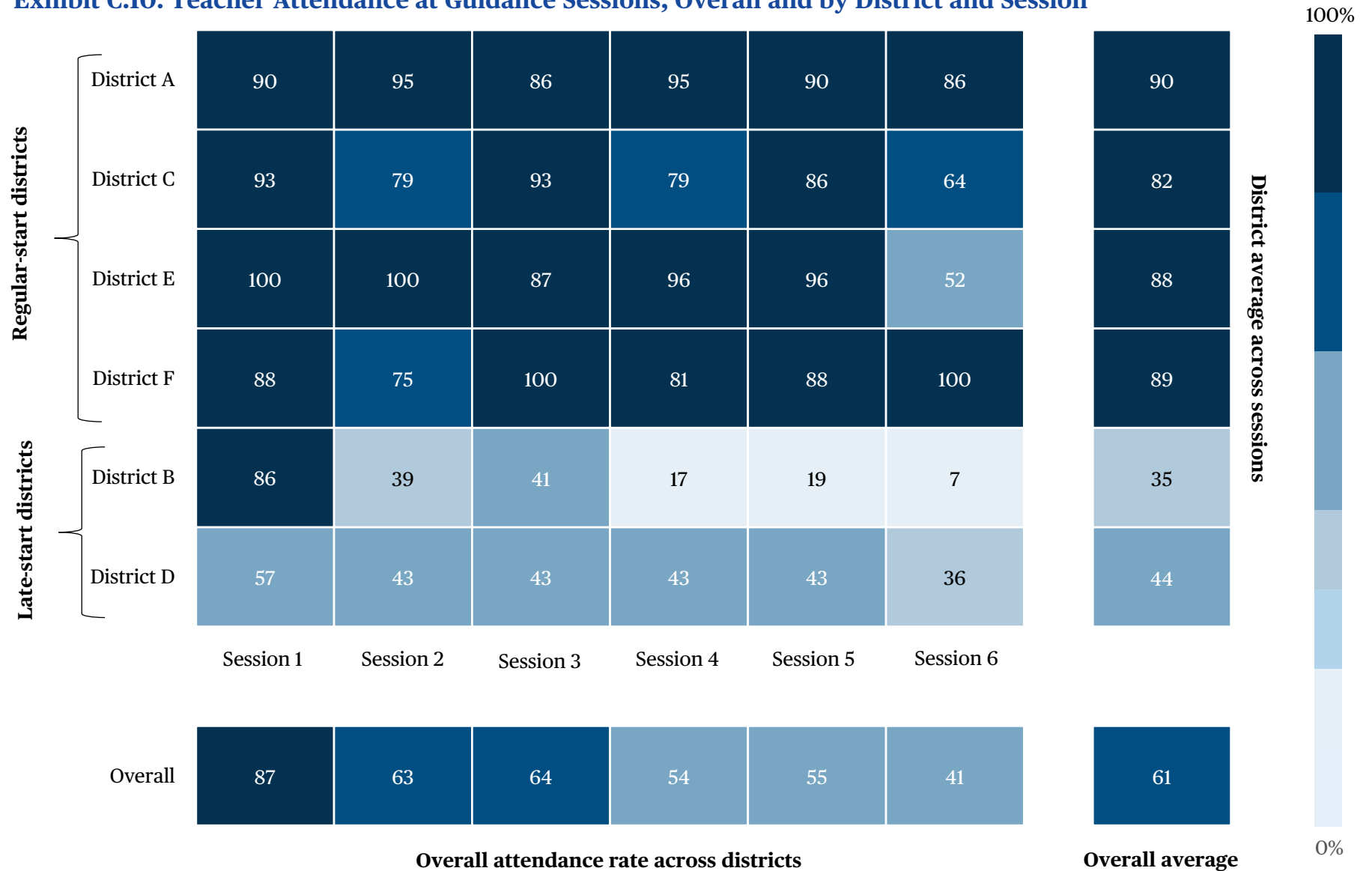
Exhibit C.9. Teacher Attendance at Initial Training, Overall and by District and Training Days



SOURCE: Authors' calculation based on teacher training attendance records collected throughout the 2017-2018 school year.

NOTES: Attendance rates are calculated based on all 157 program school teachers.
NA means that the training day was not provided.

Exhibit C.10. Teacher Attendance at Guidance Sessions, Overall and by District and Session



SOURCE: Authors' calculation based on teacher-scheduled coaching attendance records collected throughout the 2017-2018 school year.

NOTE: Attendance rates are calculated based on all 157 program school teachers.

Reflection Sessions: In addition, the program planned to provide six “reflection sessions” to teachers. These sessions were to be held one to two weeks after each guidance session. Teachers were expected to meet as a team with the coach to discuss their experience using the instructional components and practices in their classrooms. Due to two districts’ delayed program start and one other district’s lack of access to schools for coaches, the program provider could only offer all six reflection sessions in three of the six study districts. In the remaining three districts, the program provider was able to offer between one and four sessions. Therefore, the availability and attendance rate of these sessions varied by district, as shown in Exhibit C.11. The average teacher attendance rate for the reflection sessions was 28 percent. The attendance rate in each session varied from 14 percent to 52 percent.

Exhibit C.12 shows the percentage of program-school teachers receiving a given amount of training or support based on attendance data. Eighty-eight percent of program-school teachers received some initial training for one or two days. Ninety-two percent of the teachers attended at least one guidance session, and 56 percent attended four or more sessions. However, while 66 percent of the program-school teachers attended at least one reflection session, only 17 percent attended four or more sessions. In addition, 12 percent of the teachers did not show up for the initial training at all. Eight percent of teachers did not attend any guidance sessions and 34 percent of them did not attend any reflection sessions. Overall, 2.6 percent of all program-school teachers (about four teachers) did not attend any training and support events.

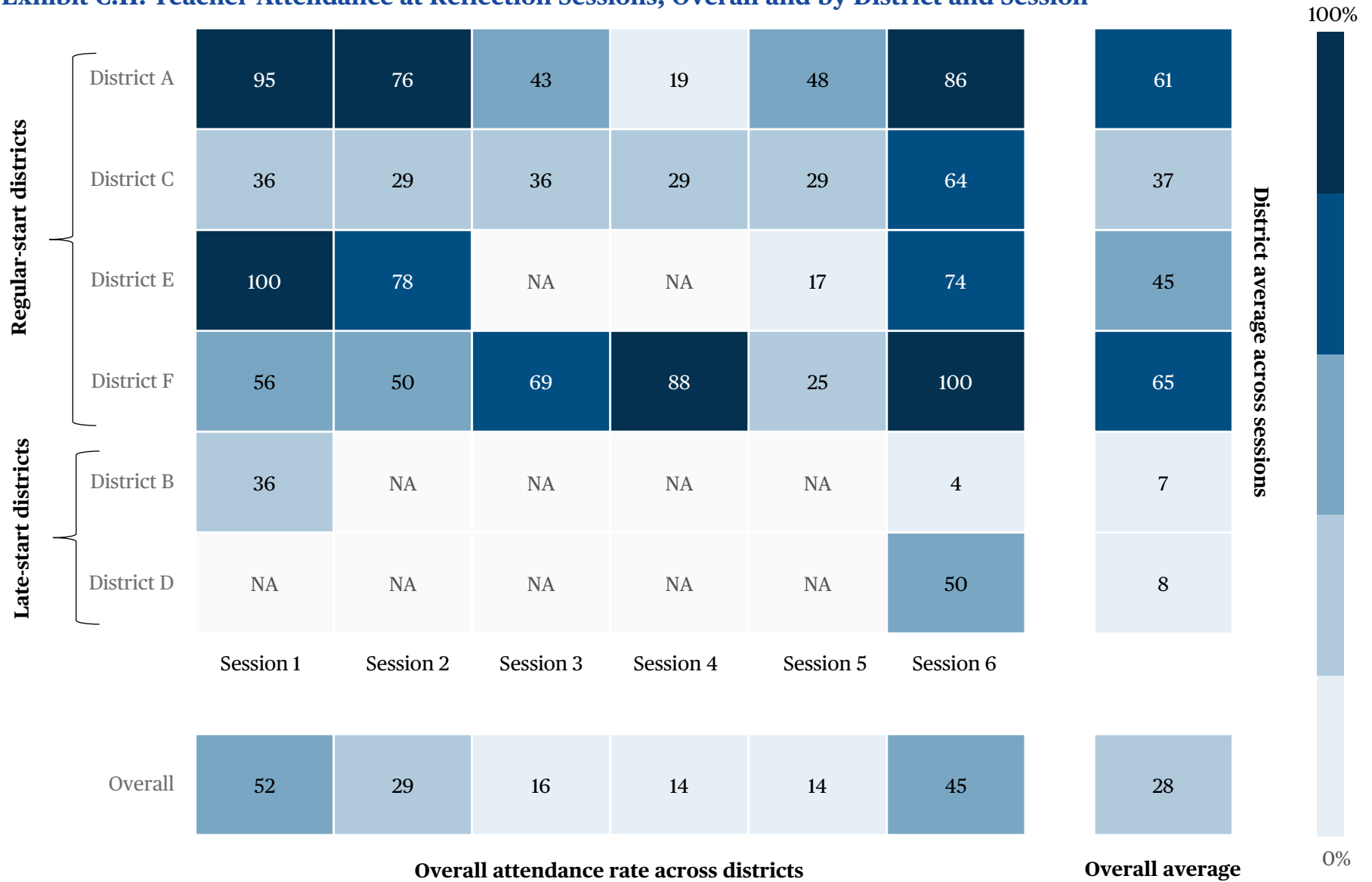
Receipt of Initial Training and Ongoing Support as Reported by Teachers

The teacher survey collected information on both program-related and non-program-related training and ongoing support for teachers in all study schools. The data include the number of hours of training teachers received by training type. The data also include the ongoing support score, which is the sum of the support activities teachers reported receiving. The total score consists of three kinds of support activities specific to the program and six types of non-program-related support. For each type of support activity, a teacher received a score of “1” if he or she received that type of support and a “0” if not. Exhibit C.13 presents the training and support received by the teachers and the contrasts between the program and non-program-school teachers.

Program-Related Training and Support: Based on survey responses from the program-school teachers, teachers in the average program school received about 10.5 hours of program-related training at the start of the program year. Of these teachers, 91 percent reported that they attended some guidance or reflection sessions during the program year. These findings are consistent with the results based on attendance records (see Exhibit C.12). Also, 35 percent of these teachers reported receiving one-on-one coaching, and 28 percent reported participation in the classroom observation of other program teachers (see Exhibit C.13).

Contrasts with Non-Program-School Teachers: Compared to teachers in the non-program schools, program-school teachers reported receiving about six fewer hours (not statistically significant) of training offered by their schools or districts that were not specific to the program (see Exhibit C.13). In particular, the program-provided initial training may have replaced some of the regular training on teaching English learners that districts provided for all teachers (5.7 hours for program-school teachers vs. 9.8 hours for non-program-school teachers, p-value = 0.013). However, teachers’ participation in five of the six types of non-program-related support activities were similar between these two groups of teachers. The only exception was that 39 percent of the program-school teachers reported receiving non-program-related one-on-one instructional coaching, compared to 23 percent of teachers in the non-program schools (p-value = 0.011). Therefore, with a few exceptions, the program-provided training and support did not seem to have replaced non-program training and support activities that the teachers would have participated in without the program.

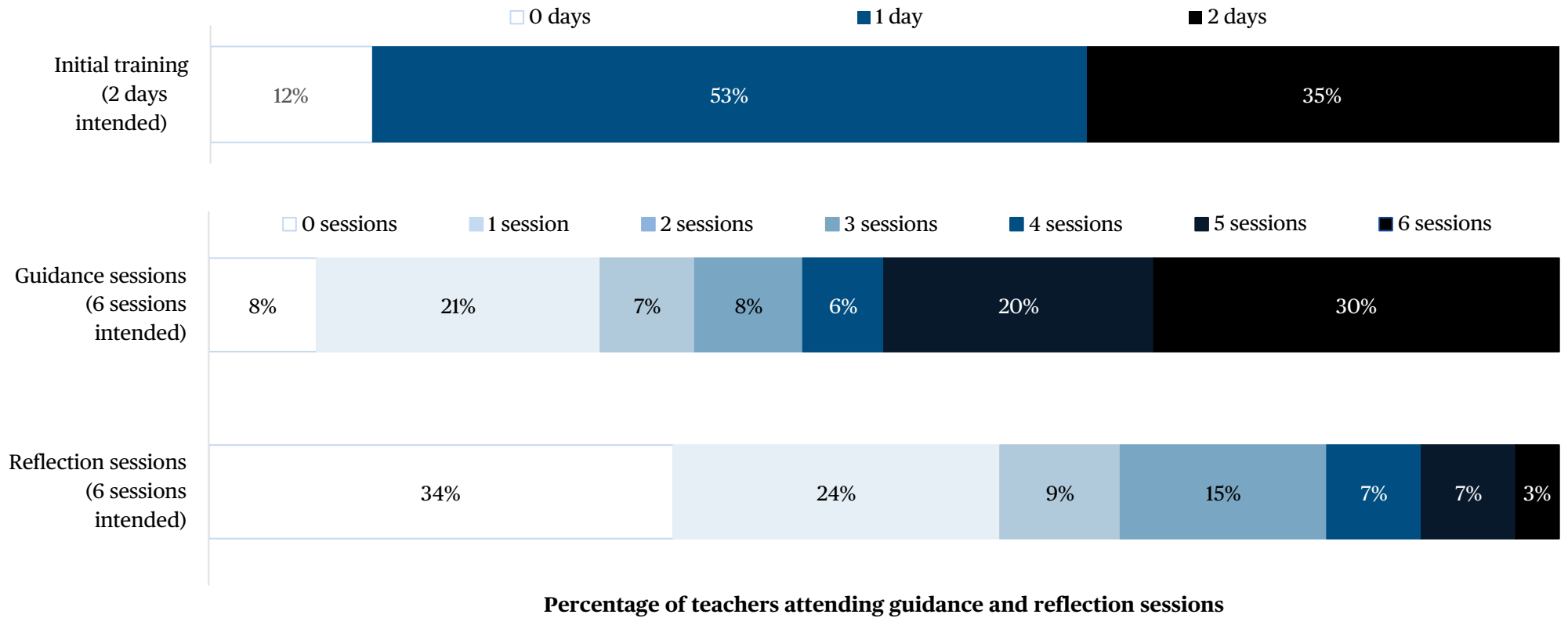
Exhibit C.11. Teacher Attendance at Reflection Sessions, Overall and by District and Session



SOURCE: Authors' calculation based on teacher-scheduled coaching attendance records collected throughout the 2017-2018 school year.

NOTES: Attendance rates are calculated based on all 157 program school teachers.

Exhibit C.12. Amount of Training and Support Teachers Received During the Program Year



SOURCES: Authors' calculation based on teacher-training attendance records collected in the summer and fall of 2017 and teacher-scheduled coaching attendance records collected throughout the 2017-2018 school year.

NOTE: Calculations based on all 157 program school teachers.

Exhibit C.13. Estimated Differences in the Amount of Training and Professional Development Reported by Teachers, Program Year

Measure	Program Schools	Non-Program Schools	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
<u>Total initial training (number of hours)</u>	25.35	21.41	3.94	3.33	0.15	0.242
Program-provided training (hours)	10.54	-	-	-	-	-
Non-program-provided training (hours)	15.30	21.47	-6.17	3.43	-0.23	0.077
Teaching English learners	5.74	9.78	-4.03*	1.57	-0.31	0.013
Teaching non-English learner struggling readers	3.77	4.67	-0.90	0.85	-0.12	0.292
Teaching language comprehension	3.20	4.35	-1.15	0.92	-0.16	0.215
Teaching word meaning	2.68	3.54	-0.86	0.74	-0.14	0.251
<u>Ongoing support total score</u>	4.24	2.73	1.50*	0.25	0.87	0.000
Program-provided ongoing support (item coverage, range = 0-3)	1.54	-	-	-	-	-
One-on-one, in-person coaching (proportion)	0.35	-	-	-	-	-
Observations of the instruction of other program teachers (proportion)	0.28	-	-	-	-	-
Guidance and reflection sessions (proportion)	0.91	-	-	-	-	-
Non-program-provided ongoing support (item coverage, range = 0-6)	2.99	2.76	0.22	0.21	0.13	0.307
Professional Learning Community (PLC) support in teaching particular student groups (proportion)	0.53	0.59	-0.06	0.07	-0.12	0.400
PLC support in classroom management (proportion)	0.41	0.31	0.10	0.06	0.20	0.098
PLC support in integrating other school subjects (proportion)	0.56	0.55	0.02	0.07	0.04	0.811
PLC support in subject matter content (proportion)	0.84	0.77	0.06	0.04	0.16	0.156
One-on-one instructional non-program coaching (proportion)	0.39	0.23	0.16*	0.06	0.35	0.011
Observations of the instruction of other teachers teaching the same subject (proportion)	0.42	0.41	0.01	0.10	0.02	0.915
Number of schools	32	26				
Number of classrooms	136	103				

(continued)

Exhibit C.13 (continued)

SOURCE: Teacher survey data collected in the 2017-2018 school year.

NOTES: The sample includes 239 fourth- and fifth-grade regular classroom teachers from 58 study schools. The number of observations for each row varies due to missing data.

The differences are estimated using linear models that account for the nested structure of the data, with teachers nested within schools. The models control for the blocking of random assignment. The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The effect sizes of the estimated differences are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

Available Instructional Days and Teacher Reported Unit Coverage

This program contains 12 curricular units to be delivered during a school year. Assuming a typical school year has 180 instructional days, the program-school teachers have 15 days on average to complete a unit. The study team calculated the available instructional days as the number of instructional days remaining in the program year from the date when teachers began teaching the program. Because teachers in each district started teaching the curricular units at different points in the school year, the number of instructional days for implementing the curricular units ranged from 119 days to 175 days across districts (Exhibit C.14). The program-school teachers reported that, on average, the last unit they covered during the program year was somewhere between unit 7 and unit 8 (average = 7.7), indicating that these teachers might have made it through about two-thirds of the total units. This teacher-reported coverage ranged from 6.5 to 9.5 units, but the variation was not statistically significant (p-value = 0.366).

Exhibit C.14. Number of Available Instructional Days and Number of Curricular Units Covered, Overall and by District, Program Year

District	Instructional Days Remaining After Program Start	Total School Year Instructional Days	Average of Last Program Unit Completed	Range of Program Units Completed		Average Instructional Days Per Unit
				Minimum	Maximum	
District A	166	180	9.6	5	12	17.4
District B	119	180	6.5	3	9	18.3
District C	175	179	8.4	6	12	20.8
District D	145	184	6.6	3	12	21.9
District E	162	170	7.2	4	11	22.6
District F	164	180	8.0	4	12	20.5
Overall	155	179	7.7	3	12	20.1

SOURCES: Program record data on implementation start dates and teacher survey data collected in the spring of 2018.

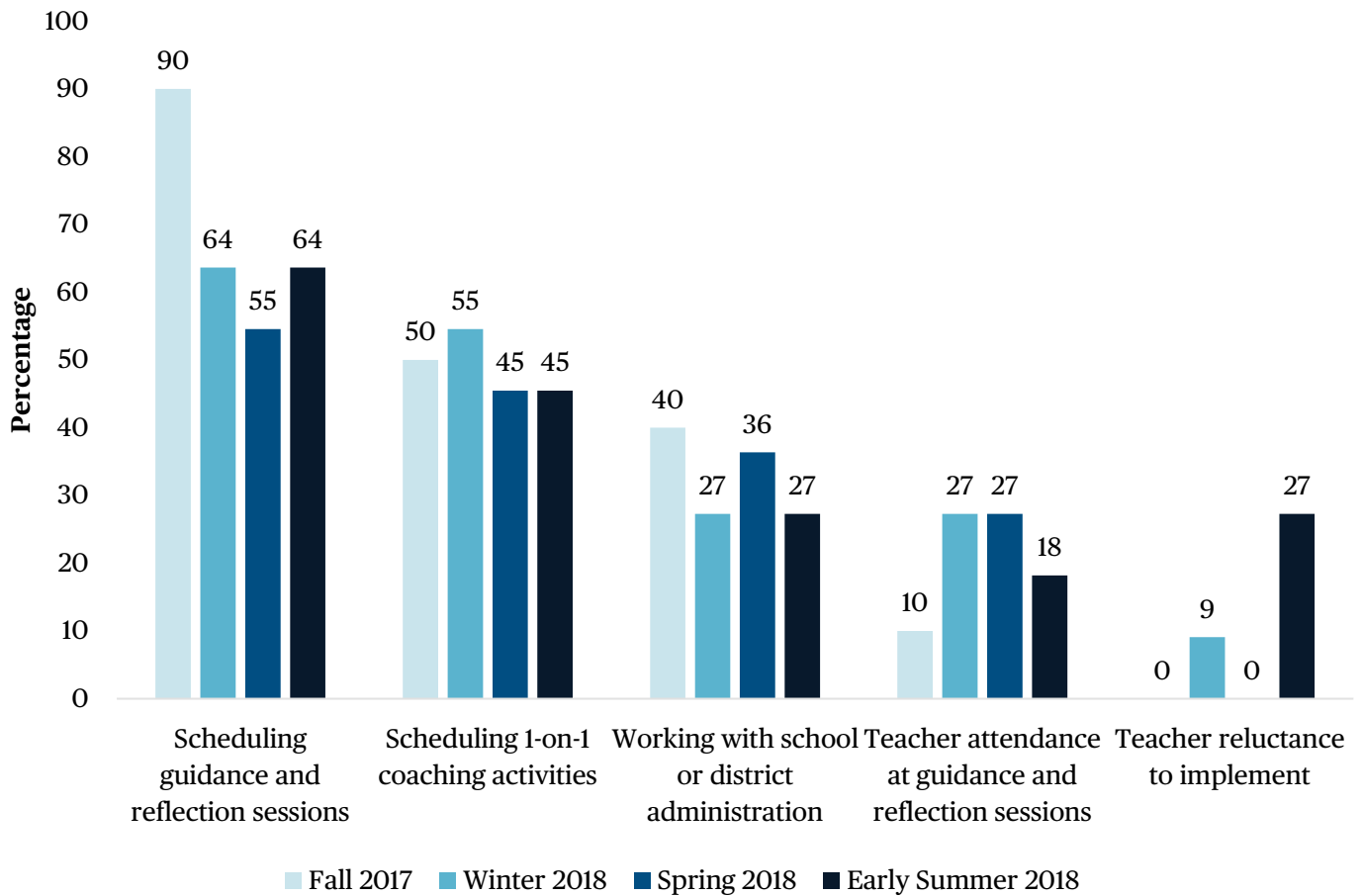
NOTE: Instructional days remaining after program start is calculated as the number of instructional days remaining in the program year from the date when teachers began implementing the program. Last program unit completed is based on the teacher-reported last unit covered.

Challenges to Implementation

The study team also collected information on what the program provider and coaches identified as challenges to the implementation of the program. The provider and coaches submitted periodic reports related to training, ongoing support, and program implementation.

Coaches faced scheduling obstacles. Coaches frequently reported difficulty scheduling support activities like small-group guidance and reflection sessions and one-on-one consultations with teachers throughout the implementation year (see Exhibit C.15).

Exhibit C.15. Top Five Implementation Challenges Reported by Coaches



SOURCES: Coach reports collected four times during the 2017-2018 school year.

NOTE: Each bar represents the percentage of coaches who reported a certain challenge for a given round of coach report collection.

Teachers had difficulty fitting the program into their schedule. In answers to an open-ended question about “the biggest challenge for teachers in implementing WordGen”, 9 of the 11 coaches identified the competing demands of other programs and state and local standardized assessments as the major challenge for teachers. This same issue emerged from the program provider’s report as well. These competing demands made it difficult for teachers to devote adequate time to the implementation of the program, consistently delivering it day to day, maintaining the pace of the lessons, and ensuring progress through the units.

Some teachers struggled with classroom management. The program provider and coaches also flagged challenges related to the management and culture in the classroom and the coaches reported that some teachers were struggling with students exhibiting challenging behaviors or with maintaining student attention. These challenges may account for the program provider’s report that teachers struggled with a shift to a more student- and discussion-centered classroom and commonly focused on teaching vocabulary over other practices that support student discourse.

IV. Additional Information for Systematic Review

Exhibits C.16-C.18 provide supplemental information about the estimation of the program effects on student outcomes that a systematic review might need to assess the quality of the study. It includes the summary statistics and the estimated effects for the impact findings presented in the report for the overall sample and the two key student subgroups.

Exhibit C.16. Supplemental Information on Student Baseline Reading and Math Achievement, by Analysis Sample

Analysis Sample	Program Schools				Non-Program Schools			
	Number of Individuals	Number of Clusters	Unadjusted Mean	Unadjusted Standard Deviation	Number of Individuals	Number of Clusters	Unadjusted Mean	Unadjusted Standard Deviation
<u>Baseline state ELA test standardized score</u>								
All students								
CALS-I sample	1,374	32	0.13	0.94	1,275	26	0.25	1.05
GMRT sample	1,333	31	0.14	0.94	1,252	26	0.26	1.04
State ELA test sample	3,466	32	-0.03	0.94	3,111	26	0.06	1.03
English learners								
CALS-I sample	338	28	-0.45	0.80	324	21	-0.38	0.86
GMRT sample	320	27	-0.42	0.81	316	21	-0.38	0.86
State ELA test sample	826	28	-0.57	0.81	811	22	-0.52	0.84
Students from economically disadvantaged backgrounds								
CALS-I sample	1,154	32	0.08	0.94	986	26	0.13	1.02
GMRT sample	1,123	31	0.10	0.94	965	26	0.15	1.03
State ELA test sample	2,988	32	-0.08	0.94	2,562	26	-0.04	1.00

(continued)

Exhibit C.16 (continued)

Analysis Sample	Program Schools				Non-Program Schools			
	Number of Individuals	Number of Clusters	Unadjusted Mean	Unadjusted Standard Deviation	Number of Individuals	Number of Clusters	Unadjusted Mean	Unadjusted Standard Deviation
<u>Baseline state math test standardized score</u>								
All students								
CALS-I sample	1,392	32	0.13	0.94	1,322	26	0.26	1.02
GMRT sample	1,354	31	0.14	0.93	1,301	26	0.27	1.02
State ELA test sample	3,551	32	-0.04	0.94	3,212	26	0.12	1.01
English learners								
CALS-I sample	356	28	-0.31	0.80	369	21	-0.21	0.91
GMRT sample	340	27	-0.27	0.82	362	21	-0.19	0.89
State ELA test sample	908	28	-0.43	0.81	906	22	-0.30	0.89
Students from economically disadvantaged backgrounds								
CALS-I sample	1,171	32	0.06	0.93	1,031	26	0.14	0.99
GMRT sample	1,141	31	0.08	0.92	1,012	26	0.16	0.99
State ELA test sample	3,056	32	-0.09	0.93	2,657	26	0.02	0.97

SOURCES: State standardized English Language Arts (ELA) and math test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: The baseline ELA and math test scores are standardized within each state and grade level, using the mean scores and the pooled standard deviations from all students with valid test scores within a state-by-grade cell.

CALS-I = Core Academic Language Skills Instrument, GMRT = Gates-MacGinitie Reading Test.

Exhibit C.17. Supplemental Information on Student Outcomes, by Outcome and Sample

Outcome	Program Schools					Non-Program Schools				
	Number of Individuals	Number of Clusters	Unadjusted Mean	Adjusted Mean	Unadjusted Standard Deviation	Number of Individuals	Number of Clusters	Unadjusted Mean	Adjusted Mean	Unadjusted Standard Deviation
All students										
CALS-I scaled score	1,634	32	496.58	496.05	27.90	1,484	26	501.54	498.04	32.76
GMRT scaled score	1,587	31	478.10	478.27	35.59	1,460	26	484.64	481.29	39.99
Percentage of students meeting proficiency standards on the state ELA test	3,984	32	27.51	26.51	44.66	3,468	26	30.77	28.01	46.16
English learners										
CALS-I scaled score	400	28	481.42	480.88	21.49	386	21	483.87	481.58	25.16
GMRT scaled score	382	27	462.68	462.72	28.32	379	21	464.40	465.25	33.45
Percentage of students meeting proficiency standards on the state ELA test	1,041	28	11.91	11.34	32.41	937	22	11.85	10.33	32.33

(continued)

Exhibit C.17 (continued)

Outcome	Program Schools					Non-Program Schools				
	Number of Individuals	Number of Clusters	Unadjusted Mean	Adjusted Mean	Unadjusted Standard Deviation	Number of Individuals	Number of Clusters	Unadjusted Mean	Adjusted Mean	Unadjusted Standard Deviation
Students from economically disadvantaged backgrounds										
CALS-I scaled score	1,238	32	495.78	495.68	27.24	1,045	26	497.91	496.85	30.68
GMRT scaled score	1,205	31	478.06	478.14	34.33	1,028	26	480.29	480.06	37.47
Percentage of students meeting proficiency standards on the state ELA test	3,253	32	26.19	25.36	43.97	2,715	26	26.56	25.84	44.17
Number of schools randomly assigned		36					34			

SOURCES: Core Academic Language Skill-Instrument (CALS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017, 2017-2018, and 2018-2019 school years.

NOTE: The adjusted means for the program and non-program schools were calculated based on a multilevel regression model that accounts for the random assignment blocks and student-level covariates, as well as the clustered structure of the data.

Exhibit C.18. Estimated Intraclass Correlation Coefficients and Explanatory Power of Covariates (R-Square) in Impact Estimation Model

Outcome	Intraclass Correlation Coefficient	School-Level R-Square	Student-Level R-Square
<u>All students</u>			
CALS-I scaled score	0.11	0.73	0.53
GMRT scaled score	0.10	0.78	0.40
Percentage of students meeting proficiency standards on the state ELA test	0.09	0.67	0.36
<u>English learners</u>			
CALS-I scaled score	0.11	0.73	0.41
GMRT scaled score	0.11	0.75	0.29
Percentage of students meeting proficiency standards on the state ELA test	0.10	0.55	0.17
<u>Students from economically disadvantaged backgrounds</u>			
CALS-I scaled score	0.10	0.68	0.58
GMRT scaled score	0.09	0.75	0.41
Percentage of students meeting proficiency standards on the state ELA test	0.08	0.63	0.36

SOURCES: Core Academic Language Skill-Instrument (CALS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: The intraclass correlation coefficient is estimated with a multilevel model that controls for the random assignment blocks. The school-level and student-level R-squares reflect the explanatory power of the following covariates: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores.

APPENDIX D. SUPPLEMENTAL FINDINGS

This appendix presents findings from additional analyses of the program’s impacts on student outcomes that are not discussed in the report but may help readers better understand or further interpret the main analyses. These analyses include sensitivity checks on whether the primary findings are robust to model specification, sample definition, and potential attrition bias. They also include impact findings on additional student outcomes and additional district or random assignment block-level subgroups and student subgroups. Lastly, the appendix presents supplementary findings from implementation and correlational analyses.

I. Sensitivity Checks of Program Impacts on Student Outcomes

This part of the appendix presents sensitivity check results on the primary impact findings shown in the report. It checks whether the findings are sensitive to model selection or sample definition. It also checks whether the impact findings are affected by the non-response or non-consent to study administered tests.

Sensitivity to Model Specification and Sample Definition

In concept, the program impact in a randomized controlled trial (RCT) is estimated by the difference in the average outcomes between the program and non-program groups. In practice, researchers usually use a regression model that controls for the baseline characteristics of the individuals in the sample to estimate such an impact. This regression approach helps reduce the random noise in the outcome measure and can often improve the precision of the impact estimation. The study team estimated the program impacts on key student outcomes and student samples with different sets of control variables and found that the findings presented in the report were robust to these variations in model specification (Exhibit D.1).

The study team also checked if the key findings were driven by the school sample used in the report. Specifically, the team dropped from the sample a school that was involved in a test cheating scandal in the year before the program year; a non-program school that had a much higher baseline reading achievement level than all other schools in the sample; and schools from one district that had the highest attrition rate and differential attrition rate. The estimated program impacts did not vary substantively across these different sample definitions (Exhibit D.2).

Sensitivity to Potential Attrition Bias

Due to the low rate of parental consent for their children to participate in testing for the study, the overall response rates for the two study-administered reading tests, CALS-I and GMRT, were below 50 percent. Even though the response rates did not differ between the program and non-program schools, it is important to assess whether the low response rates might affect the interpretation of the program effect findings.

The study team first assessed the internal validity of the analysis samples for the student outcomes. Exhibits B.12-B.20 showed that the samples of students included in the impact analyses on the three key student outcomes were similar between the program and non-program schools on a wide range of background characteristics, including their pre-program reading and math achievement levels. These findings provide supportive evidence for the internal validity of these samples.

However, there was evidence that the students who had non-missing CALS-I or GMRT scores (the “respondents”) and those who did not (the “non-respondents”) were different in terms of their background characteristics (see Exhibit D.3). The differences between the two groups are statistically significant for a range of available demographic characteristics and baseline achievement measures. These comparisons were based on five of the six study districts, or 52 of the 58 schools, because one study district did not provide the study with student records data for students whose parents or guardians did not provide consent.

Exhibit D.1. Model Specification Checks for Impacts on Student Outcomes for Program Year, Overall Sample and Subgroups

Measure	Model 1: No Covariates			Model 2: Demographic Covariates Only			Model 3: Baseline Test Covariates Only		
	Estimated Impact	Standard Error of Estimated Impact	P-Value of Estimated Impact	Estimated Impact	Standard Error of Estimated Impact	P-Value of Estimated Impact	Estimated Impact	Standard Error of Estimated Impact	P-Value of Estimated Impact
<u>Overall sample</u>									
CALS-I scaled score	-1.45	2.93	0.623	-1.97	2.41	0.419	-1.88	1.52	0.222
GMRT scaled score	-1.91	3.53	0.592	-2.70	2.94	0.364	-2.82	1.95	0.156
Percentage of students meeting proficiency standards on the state ELA test	-3.2	4.05	0.436	-3.0	3.37	0.372	-1.6	2.48	0.531
<u>English learners</u>									
CALS-I scaled score	0.39	2.92	0.894	0.77	2.71	0.777	-0.73	1.90	0.704
GMRT scaled score	-1.02	3.96	0.797	-0.55	3.51	0.877	-2.67	2.86	0.356
Percentage of students meeting proficiency standards on the state ELA test	0.2	3.54	0.953	0.5	3.05	0.875	0.9	2.95	0.773
<u>Students from economically disad- vantaged backgrounds</u>									
CALS-I scaled score	-0.45	2.77	0.873	-0.93	2.35	0.694	-1.01	1.66	0.546
GMRT scaled score	-1.23	3.46	0.724	-1.66	2.90	0.571	-2.00	2.18	0.362
Percentage of students meeting proficiency standards on the state ELA test	-1.7	3.85	0.663	-2.1	3.37	0.533	-0.3	2.51	0.916

(continued)

Exhibit D.1 (continued)

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data (sample size = 3,118 students) and the Gates-MacGinitie Reading Test (GMRT) data (sample size = 3,047 students) collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years (sample size = 7,452 students).

NOTES: The student sample includes all students with valid outcome measures. For the overall sample and students from economically disadvantaged backgrounds, the sample includes 58 study schools; for English learners, the sample includes 49 study schools.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. All models control for the blocking of random assignment and grade. Model 1 contains no covariates other than the random assignment block indicators and grade. Model 2 controls for student background characteristics such as age, gender, race and ethnicity, district-provided poverty indicator, English learner status, and Individualized Education Plan status. Model 3 controls for students' baseline standardized math and ELA test scores. All missing covariate values are imputed with zero and missing indicators for all covariates are also included in the model.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit D.2. Sample Specification Checks for Impacts on Student Outcomes, for the Overall Sample and Subgroups

Measure	Sample 1: Without Cheating School			Sample 2: Without Outlier School			Sample 3: Without High-Attrition District		
	Estimated Impact	Standard Error of	P-Value of	Estimated Impact	Standard Error of	P-Value of	Estimated Impact	Standard Error of	P-Value of
		Estimated Impact	Estimated Impact		Estimated Impact	Estimated Impact		Estimated Impact	Estimated Impact
Overall sample									
CALS-I scaled score	-2.56	1.52	0.100	-1.49	1.56	0.344	-0.26	1.85	0.890
GMRT scaled score	-3.25	1.81	0.080	-2.05	1.66	0.225	-1.20	1.97	0.547
Percentage of students meeting proficiency standards on the state ELA test	-2.9	1.96	0.154	-1.1	2.53	0.676	-1.4	3.17	0.660
Number of schools	57			57			39		
English learners									
CALS-I scaled score	-1.17	1.81	0.521	-0.89	2.01	0.660	0.31	2.20	0.890
GMRT scaled score	-2.82	2.81	0.323	-1.65	2.81	0.561	-0.02	3.22	0.994
Percentage of students meeting proficiency standards on the state ELA test	-0.5	1.65	0.741	1.9	2.63	0.474	2.1	3.88	0.596
Number of schools	48			48			31		

(continued)

Exhibit D.2 (continued)

Measure	Sample 1: Without Cheating School			Sample 2: Without Outlier School			Sample 3: Without High-Attrition District		
	Standard Error of		P-Value of	Standard Error of		P-Value of	Standard Error of		P-Value of
	Estimated Impact	Estimated Impact	Estimated Impact	Estimated Impact	Estimated Impact	Estimated Impact	Estimated Impact	Estimated Impact	Estimated Impact
<u>Students from economically disadvantaged backgrounds</u>									
CALS-I scaled score	-1.81	1.48	0.226	-0.79	1.63	0.633	0.20	1.94	0.919
GMRT scaled score	-2.39	1.88	0.210	-0.97	1.89	0.610	-0.89	2.15	0.683
Percentage of students meeting proficiency standards on the state ELA test	-1.8	1.91	0.338	0.2	2.47	0.939	0.1	3.12	0.973
Number of schools	57			57			39		

SOURCES: Core Academic Language Skills Instrument (CALS-I) data (sample size = 3,118) and the Gates-MacGinitie Reading Test (GMRT) data (sample size = 3,047 students) collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years (sample size = 7,452 students).

NOTES: Sample 1 excludes one school involved in a test cheating scandal during the baseline year. Sample 2 excludes one school with a much higher baseline achievement level than other schools. Sample 3 excludes schools from one district that had the most unbalanced attrition rate among all study districts.

ELA = English Language Arts.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit D.3. Background Characteristics Comparison of Students With and Without CALS-I or GMRT Scores

Measure	Respondents	Non-respondents	Estimated Difference	Standard Error of Estimated Difference	Effect Size of Estimated Difference	P-Value of Estimated Difference
Age (years)	8.93	9.00	-0.06 *	0.01	-0.09	0.000
Female (%)	51.6	47.7	3.8 *	1.24	0.08	0.002
Race/ethnicity (%)						
Hispanic	66.2	65.8	0.4	1.00	0.01	0.722
Black, non-Hispanic	12.2	13.1	-0.8	0.82	-0.02	0.308
White, non-Hispanic	11.1	11.0	0.1	0.62	0.00	0.869
Asian	8.4	8.4	0.1	0.50	0.00	0.878
Other	2.0	1.8	0.3	0.29	0.02	0.367
Missing	0.3	0.1	0.2	0.13	0.04	0.098
Students in Grade 4 in 2017-2018 (%)	52.1	50.3	1.8	1.24	0.04	0.152
Students with low-income status (%)	80.5	82.3	-1.7 *	0.88	-0.05	0.046
English learners (%)	25.9	37.7	-11.8 *	1.16	-0.25	0.000
Students with special education status (%)	11.6	19.4	-7.7 *	0.95	-0.21	0.000
Students meeting proficiency standards on the state ELA test (%)	32.0	22.3	9.8 *	1.15	0.22	0.000
State ELA test standardized score	0.18	-0.11	0.30 *	0.03	0.30	0.000
Students meeting proficiency standards on the state math test (%)	32.0	22.7	9.2 *	1.11	0.21	0.000
State math test standardized score	0.19	-0.10	0.29 *	0.02	0.30	0.000

SOURCES: Core Academic Language Skills Instrument (CALS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. District records data for the 2016-2017 school year.

NOTES: The sample includes all eligible Grades 4 and 5 students in 52 of the 58 study schools who had district records data for spring 2018. One district is excluded from this analysis because it did not provide records data for students whose parents or guardians did not provide consent. The numbers of observations vary by baseline characteristic due to missing values; the numbers of observations range from 5,998 to 7,943.

The estimated differences are regression-adjusted using hierarchical linear models to account for the nested structure of the data with students nested within schools. The models control for indicators of random assignment blocks. Rounding may cause slight discrepancies in calculating sums and differences.

ELA = English Language Arts.

A two-tailed t-test with a null of zero difference is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was used to determine whether there is a systematic difference between the respondents and nonrespondents, with respect to the characteristics included in this table. The p-value for this test is 0.00003.

Therefore, the study team conducted sensitivity checks to assess whether the impact findings based on the respondent sample could be generalized to the broader sample of all enrolled students. These checks were only possible for five of the six study districts that provided records data for non-consented students. Specifically, the team estimated the program effects on students' state English language arts (ELA) test performance for the respondents and non-respondents separately. The results show that the estimated program impacts on state ELA test performance did not differ by students' respondent status (Exhibit D.4). These findings suggest that the program did not appear to have affected the respondents and non-respondents differently, even though they differed in their background characteristics.

II. Supplemental Findings of Program Impacts on Student Outcomes

This section presents program impact findings on additional student outcomes. It also provides impact findings on additional student subgroups and district or random assignment block-level subgroups.

Program Impacts on Other Student Outcomes

To provide context to the key impact findings presented in the report, the study team estimated the program impacts on other student reading outcomes collected through study-administered tests or district records. The study team first estimated the program effects on the eight tasks included in the CALS-I tests. These tasks measure students' core academic language skills at the word level, sentence level, and in discourse (see Exhibit B.8 for descriptions of these tasks). Exhibit D.5 shows that the program did not affect students' performance on any of these tasks.

The study team then estimated the program impact on the standardized ELA test scores. In the report, students' performance on the state ELA tests was reported as the percentage of students scoring at or above the state proficiency level. This percentage is often used for school accountability purposes and thus is a policy-relevant measure. Alternatively, students' performance could be measured by their actual scores, standardized within district and grade so that results can be pooled together for the overall sample (see Appendix B for the description of the standardization). Exhibit D.6 presents the estimated program impacts on these alternative state ELA test measures. The findings are consistent with the primary impact findings presented in the report.

Finally, the study team estimated the program impact on students' performance in state math tests. Even though the current program targeted students' academic language and reading skills, one hypothesis is that improved academic language skills could help students better absorb content and context in other subject areas. Thus, the study team looked at whether the program had any effect on students' math performance. Exhibit D.7 presents these findings and shows that, after one year of implementation, the program did not affect students' math test scores in either the program year or the follow-up year.

Program Impacts for Other Student Subgroups

The study team assessed possible heterogeneity in program impacts across different student populations. To do so, the study team estimated program impacts for student-level subgroups based on students' English learner status, family income status, gender, grade level, and reading performance level prior to the program year. Exhibits D.8-D.12 present findings from these subgroup analyses. By and large, the estimated program impacts did not vary by these student background characteristics, with one exception. As shown in Panel A of Exhibit D.11, the program appeared to negatively affect fifth-graders' CALS-I and GMRT test scores. In addition, the estimated impacts for fifth-graders seemed to differ from those for fourth-graders. Further exploratory analyses revealed that this result might be driven by one district that had one school with a high-achieving fifth-grade cohort in the program year. The overall difference in estimated impacts between fourth and fifth grades disappeared if this district was excluded from the impact estimation (see Panel B of Exhibit D.11). These subgroup impact findings show no compelling evidence for impact variation across subgroups defined by student background characteristics.

Exhibit D.4. Estimated Impacts on Percentages of Students Meeting Proficiency Standards in State ELA Tests, by Respondent Status, Sample, and Year

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	95 Percent Confidence Interval		Program Sample Size	Non-Program Sample Size
							Lower Bound	Upper Bound		
Overall sample										
Program year										
Respondents	33.6	36.5	-2.9	2.85	-0.06	0.311	-8.5	2.7	1,443	1,399
Nonrespondents	23.9	24.4	-0.5	2.75	-0.01	0.863	-5.9	4.9	2,375	2,023
Estimated difference			-2.4	1.92	-0.05	0.208	-6.2	1.3	3,818	3,422
Follow-up year										
Respondents	36.1	38.2	-2.1	2.72	-0.04	0.440	-7.4	3.2	1,264	1,272
Nonrespondents	22.1	22.5	-0.5	2.80	-0.01	0.868	-5.9	5.0	1,999	1,659
Estimated difference			-1.6	2.40	-0.03	0.495	-6.3	3.1	3,263	2,931
English learners										
Program year										
Respondents	14.3	13.2	1.0	3.39	0.02	0.758	-5.6	7.7	337	352
Nonrespondents	9.5	8.6	0.9	3.15	0.02	0.772	-5.3	7.1	629	557
Estimated difference			0.1	3.14	0.00	0.967	-6.0	6.3	966	909
Follow-up year										
Respondents	16.4	16.7	-0.4	3.47	-0.01	0.914	-7.2	6.4	315	336
Nonrespondents	12.6	8.0	4.6	3.63	0.09	0.207	-2.5	11.7	581	456
Estimated difference			-5.0	4.09	-0.10	0.225	-13.0	3.1	896	792
Students from economically disadvantaged backgrounds										
Program year										
Respondents	32.4	34.2	-1.7	2.92	-0.04	0.549	-7.5	4.0	1,158	1,032
Nonrespondents	22.5	22.2	0.3	2.82	0.01	0.904	-5.2	5.9	1,939	1,639
Estimated difference			-2.1	2.11	-0.04	0.321	-6.2	2.0	3,097	2,671
Follow-up year										
Respondents	35.9	36.6	-0.7	2.70	-0.01	0.798	-6.0	4.6	1,006	934
Nonrespondents	21.2	20.6	0.7	2.89	0.01	0.818	-5.0	6.3	1,591	1,322
Estimated difference			-1.4	2.74	-0.03	0.620	-6.7	4.0	2,597	2,256

(continued)

Exhibit D.4 (continued)

SOURCES: State standardized English Language Arts (ELA) test data and district records data for the 2016-2017, 2017-2018, and 2018-2019 school years.

NOTES: The sample includes all eligible Grades 4 and 5 students in 52 of the 58 study schools who had district records data for spring 2018. One district is excluded from this analysis because it did not provide records data for students whose parents or guardians did not provide consent.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit D.5. Estimated Impacts on CALS-I Subscales

Outcome	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	95 Percent Confidence Interval	
							Lower Bound	Upper Bound
CALS-I section scores								
Connecting ideas	4.31	4.57	-0.26	0.14	-0.09	0.068	-0.54	0.02
Tracking themes	2.32	2.37	-0.05	0.06	-0.03	0.465	-0.17	0.08
Organizing texts	3.88	3.95	-0.08	0.14	-0.03	0.584	-0.35	0.20
Breaking works	6.25	6.32	-0.07	0.17	-0.02	0.699	-0.42	0.28
Comprehending sentences	2.43	2.44	-0.01	0.07	-0.01	0.859	-0.15	0.13
Identifying definitions	2.08	2.10	-0.02	0.06	-0.02	0.720	-0.13	0.09
Sure/unsure	3.99	4.07	-0.08	0.12	-0.05	0.487	-0.32	0.16
Understanding responses	2.50	2.60	-0.10	0.11	-0.08	0.336	-0.32	0.11
Sample size								
Number of students	1,634	1,484						
Number of schools	32	26						

SOURCES: Core Academic Language Skills Instrument (CALS-I) data (sample size = 3,118 students) collected in the spring of 2018. District records data for the 2016-2017 school year.

NOTES: The student sample includes all students with valid outcome measures.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and English Language Arts (ELA) test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit D.6. Estimated Impacts on State ELA Test Standardized Scores, by Sample and Year

Outcome	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	95 Percent Confidence Interval	
							Lower Bound	Upper Bound
Overall sample								
Program year	-0.03	0.01	-0.04	0.07	-0.04	0.585	-0.17	0.10
Follow-up year	-0.07	-0.04	-0.03	0.07	-0.03	0.620	-0.17	0.10
English learners								
Program year	-0.48	-0.45	-0.03	0.09	-0.03	0.710	-0.22	0.15
Follow-up year	-0.43	-0.45	0.03	0.09	0.03	0.762	-0.15	0.20
Students from economically disadvantaged backgrounds								
Program year	-0.06	-0.04	-0.03	0.07	-0.03	0.704	-0.16	0.11
Follow-up year	-0.07	-0.06	-0.01	0.07	-0.01	0.906	-0.15	0.13

SOURCES: State standardized English Language Arts (ELA) test data and district records data for the 2016-2017, 2017-2018, and 2018-2019 school years.

NOTES: The overall sample includes all Grades 4 and 5 students with state ELA test scores (sample size = 7,452 students). English learners are identified by their status in the 2016-2017 school year. Schools from one district are not included in this subgroup analysis because no English learners participated in the program (sample size = 1,978 students).

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

State test scores are standardized by subtracting the non-program sample mean within each state and grade level, and then dividing by the non-program sample standard deviations within each state and grade level.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit D.7. Estimated Impacts on State Math Test Scores, by Sample and Year

Outcome	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	95 Percent Confidence Interval	
							Lower Bound	Upper Bound
<u>Percentage of students meeting proficiency standards on state math test</u>								
Overall sample								
Program year	21.6	23.6	-2.0	2.22	-0.04	0.377	-6.5	2.5
Follow-up year	22.0	23.1	-1.1	1.82	-0.03	0.532	-4.8	2.5
English learners								
Program year	10.4	11.5	-1.1	3.04	-0.03	0.713	-7.3	5.0
Follow-up year	12.4	11.1	1.2	2.80	0.03	0.664	-4.4	6.9
Students from economically disadvantaged backgrounds								
Program year	20.9	21.6	-0.7	2.34	-0.02	0.761	-5.4	4.0
Follow-up year	20.6	20.8	-0.1	2.01	0.00	0.943	-4.2	3.9
<u>State math test standardized score</u>								
Overall sample								
Program year	-0.08	-0.02	-0.06	0.06	-0.06	0.352	-0.19	0.07
Follow-up year	-0.08	-0.06	-0.02	0.06	-0.02	0.718	-0.13	0.09
English learners								
Program year	-0.39	-0.37	-0.02	0.09	-0.02	0.831	-0.20	0.16
Follow-up year	-0.34	-0.40	0.06	0.08	0.06	0.486	-0.11	0.22
Students from economically disadvantaged backgrounds								
Program year	-0.10	-0.06	-0.04	0.07	-0.04	0.581	-0.18	0.10
Follow-up year	-0.11	-0.12	0.01	0.06	0.01	0.876	-0.11	0.13

(continued)

Exhibit D.7 (continued)

SOURCES: State standardized English Language Arts (ELA) test data and district records data for the 2016-2017, 2017-2018, and 2018-2019 school years.

NOTES: The overall sample includes all fourth- and fifth-grade students with state math test scores (sample size = 7,615 students). English learners are identified by their status in the 2016-2017 school year. Schools from one district are not included in this subgroup analysis because no English learners participated in the program (sample size = 2,020 students). Students from economically disadvantaged backgrounds are identified by their family income status in the 2016-2017 school year (sample size = 5,997 students).

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

State test scores are standardized by subtracting the non-program sample mean within each state and grade level, and then dividing by the non-program sample standard deviations within each state and grade level.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit D.8. Estimated Impacts on Student Outcomes in the Program Year, by English Learner Status

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference	Number of Observations
CALS-I scaled score							1.70	0.367	
English learners	480.66	481.37	-0.71	2.12	-0.02	0.738			
Other students	504.71	507.11	-2.41	1.92	-0.07	0.210			
Sample size									2,385
GMRT scaled score							3.57	0.200	
English learners	462.48	464.63	-2.15	2.48	-0.05	0.386			
Other students	486.04	491.75	-5.71 *	2.11	-0.14	0.007			
Sample size									2,338
Percentage of students meeting proficiency standards on the state ELA test							0.8	0.770	
English learners	11.3	11.2	0.1	3.19	0.00	0.973			
Other students	34.6	35.2	-0.6	3.05	-0.01	0.833			
Sample size									5,926

(continued)

Exhibit D.8 (continued)

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: English learners are identified by their status in the 2016-2017 school year. Schools from one district are not included in this subgroup analysis because no English learners participated in the program.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, district-provided poverty indicator, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was applied to the difference in the estimated impacts between the two subgroups.

Exhibit D.9. Estimated Impacts on Student Outcomes in the Program Year, by Economic Background

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference	Number of Observations
CALS-I scaled score							-2.28	0.457	
Students from economically disadvantaged backgrounds	495.68	496.92	-1.24	1.53	-0.04	0.417			
Students not from economically disadvantaged backgrounds	506.94	505.91	1.04	3.19	0.03	0.745			
Sample size									2,810
GMRT scaled score							-0.13	0.978	
Students from economically disadvantaged backgrounds	478.14	479.92	-1.78	1.75	-0.04	0.310			
Students not from economically disadvantaged backgrounds	483.02	484.67	-1.65	4.54	-0.04	0.716			
Sample size									2,750

(continued)

Exhibit D.9 (continued)

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference	Number of Observations
Percentage of students meeting proficiency standards on the state ELA test							1.80	0.612	
Students from economically disadvantaged backgrounds	25.4	25.9	-0.5	2.50	-0.01	0.832			
Students not from economically disadvantaged backgrounds	35.2	37.5	-2.3	3.94	-0.05	0.558			
Sample size									7,069

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: Students from economically disadvantaged backgrounds are identified by their family income status in the 2016-2017 school year.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, gender, race and ethnicity, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

Exhibit D.10. Estimated Impacts on Student Outcomes in the Program Year, by Student Gender

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
CALS-I scaled score							-0.57	0.716
Female	498.68	500.67	-1.99	1.66	-0.06	0.231		
Male	494.65	496.06	-1.42	1.66	-0.04	0.394		
Sample size								2,946
GMRT scaled score							1.50	0.516
Female	482.57	484.65	-2.09	2.03	-0.05	0.303		
Male	474.82	478.41	-3.59	2.02	-0.09	0.076		
Sample size								2,875
Percentage of students meeting proficiency standards on the state ELA test							3.4	0.107
Female	32.9	32.8	0.2	2.77	0.00	0.951		
Male	20.1	23.3	-3.3	2.65	-0.07	0.217		
Sample size								7,451

(continued)

Exhibit D.10 (continued)

SOURCES: Core Academic Language Skills Instrument (CALs-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: Student gender is based on information from the 2016-2017 school year district records data.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was applied to the difference in the estimated impacts between the two subgroups.

Exhibit D.11. Estimated Impacts on Student Outcomes in the Program Year, by Grade Level

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
Overall sample								
CALS-I scaled score							-4.31	0.006 †
Grade 4	491.28	490.89	0.39	1.69	0.01	0.819		
Grade 5	502.19	506.11	-3.93 *	1.73	-0.12	0.023		
Sample size								3,118
GMRT scaled score							-3.85	0.092
Grade 4	473.32	474.31	-0.99	2.00	-0.02	0.620		
Grade 5	484.05	488.88	-4.84 *	2.07	-0.12	0.020		
Sample size								3,047
Percentage of students meeting proficiency standards on the state ELA test							-1.9	0.376
Grade 4	28.3	28.8	-0.5	2.66	-0.01	0.854		
Grade 5	24.8	27.1	-2.3	2.72	-0.05	0.387		
Sample size								7,452
Panel B: excluding the district with outlying values								
CALS-I scaled score							-2.61	0.103
Grade 4	491.29	491.69	-0.40	1.75	-0.01	0.819		
Grade 5	502.79	505.79	-3.01	1.79	-0.09	0.094		
Sample size								2,900

(continued)

Exhibit D.11 (continued)

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
GMRT scaled score							0.86	0.705
Grade 4	472.23	474.51	-2.29	1.93	-0.06	0.235		
Grade 5	484.26	485.68	-1.42	2.00	-0.04	0.477		
Sample size								2,842
Percentage of students meeting proficiency standards on the state ELA test							-1.0	0.547
Grade 4	28.3	28.4	-0.1	2.70	0.00	0.959		
Grade 5	26.4	27.5	-1.2	2.71	-0.02	0.663		
Sample size								7,240

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: Student grade level is based on information from the 2016-2017 school year district records data.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was applied to the difference in the estimated impacts between the two subgroups. Statistical significance is indicated by (†) when the p-value is less than 0.05.

Exhibit D.12. Estimated Impacts on Student Outcomes in the Program Year, by Pre-Program Reading Level

Measure	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
CALS-I scaled score							-2.09	0.324
Students who were proficient readers in spring 2017	522.56	524.71	-2.16	2.20	-0.07	0.328		
Students who were not proficient readers in spring 2017	489.26	489.32	-0.06	1.45	0.00	0.964		
Sample size								2,636
GMRT scaled score							0.94	0.770
Students who were proficient readers in spring 2017	509.30	509.53	-0.23	3.19	-0.01	0.942		
Students who were not proficient readers in spring 2017	469.86	471.04	-1.17	1.92	-0.03	0.541		
Sample size								2,571
Percentage of students meeting proficiency standards on the state ELA test							-5.3	0.1
Students who were proficient readers in spring 2017	72.7	77.9	-5.1	3.3	-0.1	0.1		
Students who were not proficient readers in spring 2017	13.6	13.5	0.2	2.0	0.0	0.9		
Sample size								6,582

(continued)

Exhibit D.12 (continued)

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: Students' pre-program reading level is based on their performance in the state ELA test in the spring of 2017.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was applied to the difference in the estimated impacts between the two subgroups.

Program Impacts for District or Random Assignment Block Subgroups

The study team also explored whether the program impacts varied by students' experience with the program. The report explored the association between program implementation features and student outcomes through a multi-level, multi-variate regression framework (see Section III in Appendix B for a description of the approach and Exhibits C.7 and C.8 for the findings). Another approach is to divide the study sample into two subgroups based on the implementation measures' level or the magnitude of contrasts in these measures. The districts or random assignment blocks with higher than median level values of a given feature are in the "high" implementation group. Those with values below the median level are in the "low" implementation group. Assessing and comparing the estimated impacts between the high and low implementation subgroups for each feature could provide indications of whether and how certain implementation features might be associated with the program impacts.

There are a couple of issues worth noting when interpreting the results from this analysis. First, this kind of subgroup analysis cannot be considered as experimental because the values of the implementation features were determined after random assignment. As a result, some unobserved factors associated with both implementation and student outcome levels could potentially bias the findings. Examples of such factors include school leadership or teacher quality. Second, the study team examined 14 sets of subgroups and 28 individual subgroups for each student outcome. This amount of hypothesis testing greatly increases the likelihood of concluding that a given test is statistically significant when such impact does not exist (this is known as a type I error or a false positive). In particular, one would expect to see one false positive for every 20 hypothesis tests conducted when $p < 0.05$ is selected as the criterion for statistical significance. Therefore, findings reported here need to be interpreted with caution.

Exhibits D.13-D.18 present the estimated impacts for the subgroups defined at the district level or the random assignment block level for the three key student outcomes, respectively. By and large, there were no consistent and systematic patterns identifying any single implementation factor that might be associated with the program impacts.

III. Supplemental Information About Contrasts in Program Implementation

This section provides supplementary findings on how the program affected teachers' use of measured instructional practices important for academic language development, how such practices were associated with student outcomes, and how the program affected teachers' self-reported attitudes and beliefs about training and teaching.

Detailed Program Impacts on Use of Instructional Practices

As discussed earlier, the program produced a difference in teachers' use of instructional practices important for academic language development, and this overall difference was largely driven by teachers' increased use of word knowledge instruction. Exhibit D.19 expands those findings by showing the estimated difference in each specific practice included in the three core components. It shows that the program changed teachers' use of some practices but not others within each core component.

The study team further dissected the contrasts in teachers' use of these instructional practices by district subgroups. In particular, the team estimated such differences separately for the districts that started implementing the program on time and the districts that started implementation late and compared the estimated difference for these two groups. As shown in Exhibits C.9-C.11, due to time constraints, teachers from the late-start districts received less initial training and ongoing support than those from the districts that began implementation at the start of the program year. Specifically, in four study districts that started the program on time at the beginning of the program year, the provider delivered two days of initial training, and about 74 percent of teachers participated in both days of training. In two districts that started the program late in the fall of the program year, the program provided an adapted single day of condensed training that covered the same topics, and about 87 percent of teachers in these two districts participated in the training. In addition, due to two districts' delayed program start and one other district's lack of access to schools for coaches, the program provider could only offer all six reflection sessions in three of the six study districts. In the remaining three districts, the program provider was able to offer between one and four sessions. Therefore, the availability and attendance rate of these sessions varied by district (Exhibit C.11).

Exhibit D.13. Estimated Impacts on CALS-I Scores, by District-Level Subgroup

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Use of program-specific practices</u>								
Overall							-7.40	0.017 †
High	494.34	499.39	-5.05 *	2.11	-0.15	0.021		
Low	498.55	496.20	2.36	2.13	0.07	0.274		
Word knowledge instruction							-2.50	0.433
High	492.58	495.65	-3.07	2.50	-0.09	0.226		
Low	498.43	499.00	-0.57	1.93	-0.02	0.769		
Academic skill instruction							-7.40	0.017 †
High	494.34	499.39	-5.05 *	2.11	-0.15	0.021		
Low	498.55	496.20	2.36	2.13	0.07	0.274		
Provision of practice opportunities							-0.25	0.936
High	497.73	499.48	-1.75	2.10	-0.05	0.409		
Low	493.60	495.09	-1.50	2.34	-0.05	0.527		
<u>Contrast in the use of general teaching practices related to academic language instruction</u>								
Overall							-2.50	0.433
High	492.58	495.65	-3.07	2.50	-0.09	0.226		
Low	498.43	499.00	-0.57	1.93	-0.02	0.769		
Word knowledge instruction							-2.50	0.433
High	492.58	495.65	-3.07	2.50	-0.09	0.226		
Low	498.43	499.00	-0.57	1.93	-0.02	0.769		
Academic skill instruction							0.29	0.931
High	493.61	495.35	-1.75	2.73	-0.05	0.525		
Low	497.52	499.56	-2.03	1.89	-0.06	0.289		
Provision of practice opportunities							-2.50	0.433
High	492.58	495.65	-3.07	2.50	-0.09	0.226		
Low	498.43	499.00	-0.57	1.93	-0.02	0.769		

(continued)

Exhibit D.13 (continued)

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Contrast in received training</u>								
Overall							-0.29	0.931
High	497.52	499.56	-2.03	1.89	-0.06	0.289		
Low	493.61	495.35	-1.75	2.73	-0.05	0.525		
Program training							1.25	0.679
High	495.37	496.64	-1.27	2.20	-0.04	0.568		
Low	496.52	499.04	-2.52	2.04	-0.08	0.224		
Non-program training							-0.29	0.931
High	497.52	499.56	-2.03	1.89	-0.06	0.289		
Low	493.61	495.35	-1.75	2.73	-0.05	0.525		
<u>Last curricular unit reached</u>								
Overall							1.70	0.587
High	495.53	496.41	-0.89	2.52	-0.03	0.727		
Low	496.37	498.95	-2.59	1.80	-0.08	0.159		
<u>Available instruction days</u>								
Overall							1.70	0.587
High	495.53	496.41	-0.89	2.52	-0.03	0.727		
Low	496.37	498.95	-2.59	1.80	-0.08	0.159		
<u>By implementation start time</u>								
Overall							-1.08	0.715
On time	495.25	496.40	-1.16	2.02	-0.04	0.569		
Late start	496.97	499.21	-2.24	2.16	-0.07	0.305		

(continued)

Exhibit D.13 (continued)

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data collected in the spring of 2018. District records data for the 2016-2017 school year.

NOTES: For each implementation feature, districts were divided into two groups: districts with higher than median value in the given feature and districts with lower than median value in the feature.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and English Language Arts test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was applied to the difference in the estimated impacts between the two subgroups. Statistical significance is indicated by (†) when the p-value is less than 0.05.

Exhibit D.14. Estimated Impacts on CALS-I Scores, by Random Assignment Block Level Subgroup

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Use of program-specific practices</u>								
Overall							0.50	0.878
High	496.63	498.33	-1.70	1.89	-0.05	0.374		
Low	493.27	495.47	-2.20	2.59	-0.07	0.401		
Word knowledge instruction							4.19	0.233
High	495.92	496.73	-0.81	1.75	-0.02	0.646		
Low	494.91	499.91	-5.00	2.99	-0.15	0.101		
Academic skill instruction							3.45	0.265
High	496.32	496.93	-0.61	1.96	-0.02	0.756		
Low	494.12	498.19	-4.07	2.34	-0.12	0.090		
Provision of practice opportunities							3.15	0.331
High	498.49	499.06	-0.57	2.06	-0.02	0.782		
Low	490.52	494.24	-3.72	2.45	-0.11	0.137		
<u>Contrast in the use of general teaching practices related to academic language instruction</u>								
Overall							0.20	0.950
High	493.89	495.41	-1.52	2.16	-0.05	0.486		
Low	498.50	500.22	-1.72	2.24	-0.05	0.448		
Word knowledge instruction							-2.59	0.410
High	492.01	495.15	-3.14	2.04	-0.10	0.133		
Low	500.93	501.48	-0.55	2.35	-0.02	0.817		
Academic skill instruction							3.58	0.246
High	496.21	496.47	-0.26	1.89	-0.01	0.890		
Low	494.49	498.33	-3.84	2.38	-0.12	0.114		
Provision of practice opportunities							-3.40	0.267
High	495.04	498.54	-3.49	2.00	-0.11	0.088		
Low	496.07	496.16	-0.09	2.27	0.00	0.969		

(continued)

Exhibit D.14 (continued)

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Contrast in received training</u>								
Overall							0.46	0.888
High	497.60	499.33	-1.72	1.99	-0.05	0.391		
Low	493.76	495.95	-2.18	2.59	-0.07	0.403		
Program training							0.91	0.807
High	494.13	495.93	-1.80	1.74	-0.05	0.307		
Low	500.03	502.75	-2.71	3.28	-0.08	0.413		
Non-program training							0.46	0.888
High	497.60	499.33	-1.72	1.99	-0.05	0.391		
Low	493.76	495.95	-2.18	2.59	-0.07	0.403		
<u>Last curricular unit reached</u>								
Overall							2.94	0.346
High	496.21	497.23	-1.01	1.84	-0.03	0.586		
Low	493.78	497.74	-3.95	2.47	-0.12	0.117		
<u>Available instruction days</u>								
Overall							3.58	0.246
High	496.21	496.47	-0.26	1.89	-0.01	0.890		
Low	494.49	498.33	-3.84	2.38	-0.12	0.114		

SOURCES: Core Academic Language Skills Instrument (CALIS-I) data collected in the spring of 2018. District records data for the 2016-2017 school year.

NOTES: For each implementation feature, random assignment blocks were divided into two groups: blocks with higher than median value in the given feature, and blocks with lower than median value in the feature.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and English Language Arts test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was applied to the difference in the estimated impacts between the two subgroups.

Exhibit D.15. Estimated Impacts on GMRT Scores, by District-Level Subgroup

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Use of program-specific practices</u>								
Overall							-9.54	0.010 †
High	476.46	483.24	-6.78*	2.55	-0.17	0.011		
Low	480.77	478.01	2.76	2.48	0.07	0.271		
Word knowledge instruction							-4.14	0.269
High	480.38	485.41	-5.02	2.95	-0.13	0.096		
Low	476.74	477.62	-0.88	2.23	-0.02	0.696		
Academic skill instruction							-9.54	0.010 †
High	476.46	483.24	-6.78*	2.55	-0.17	0.011		
Low	480.77	478.01	2.76	2.48	0.07	0.271		
Provision of practice opportunities							-4.22	0.247
High	476.78	481.23	-4.46	2.42	-0.11	0.072		
Low	480.33	480.57	-0.23	2.66	-0.01	0.930		
<u>Contrast in the use of general teaching practices related to academic language instruction</u>								
Overall							-4.14	0.269
High	480.38	485.41	-5.02	2.95	-0.13	0.096		
Low	476.74	477.62	-0.88	2.23	-0.02	0.696		
Word knowledge instruction							-4.14	0.269
High	480.38	485.41	-5.02	2.95	-0.13	0.096		
Low	476.74	477.62	-0.88	2.23	-0.02	0.696		
Academic skill instruction							1.15	0.762
High	479.74	481.98	-2.24	3.14	-0.06	0.480		
Low	477.34	480.73	-3.39	2.10	-0.08	0.114		
Provision of practice opportunities							-4.14	0.269
High	480.38	485.41	-5.02	2.95	-0.13	0.096		
Low	476.74	477.62	-0.88	2.23	-0.02	0.696		

(continued)

Exhibit D.15 (continued)

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Contrast in received training</u>								
Overall							-1.15	0.762
High	477.34	480.73	-3.39	2.10	-0.08	0.114		
Low	479.74	481.98	-2.24	3.14	-0.06	0.480		
Program training							2.60	0.471
High	479.03	480.57	-1.54	2.58	-0.04	0.555		
Low	477.72	481.85	-4.13	2.47	-0.10	0.101		
Non-program training							-1.15	0.762
High	477.34	480.73	-3.39	2.10	-0.08	0.114		
Low	479.74	481.98	-2.24	3.14	-0.06	0.480		
<u>Last curricular unit reached</u>								
Overall							-0.48	0.901
High	478.32	481.95	-3.63	3.12	-0.09	0.252		
Low	478.23	481.38	-3.15	2.20	-0.08	0.160		
<u>Available instruction days</u>								
Overall							-0.48	0.901
High	478.32	481.95	-3.63	3.12	-0.09	0.252		
Low	478.23	481.38	-3.15	2.20	-0.08	0.160		
<u>By implementation start time</u>								
Overall							0.53	0.884
On time	479.67	482.94	-3.27	2.45	-0.08	0.189		
Late start	476.57	479.32	-2.75	2.63	-0.07	0.302		

(continued)

Exhibit D.15 (continued)

SOURCES: The Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. District records data for the 2016-2017 school year.

NOTES: For each implementation feature, random assignment blocks were divided into two groups: blocks with higher than median value in the given feature, and blocks with lower than median value in the feature.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and English Language Arts test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was applied to the difference in the estimated impacts between the two subgroups. Statistical significance is indicated by (†) when the p-value is less than 0.05.

Exhibit D.16. Estimated Impacts on GMRT Scores, by Random Assignment Block Level Subgroup

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Use of program-specific practices</u>								
Overall							-5.74	0.131
High	477.53	482.00	-4.46 *	2.20	-0.11	0.049		
Low	479.50	478.23	1.27	3.00	0.03	0.674		
Word knowledge instruction							0.58	0.887
High	478.96	481.32	-2.36	2.02	-0.06	0.249		
Low	475.12	478.06	-2.94	3.52	-0.07	0.408		
Academic skill instruction							3.24	0.341
High	477.82	478.90	-1.09	2.17	-0.03	0.620		
Low	477.88	482.21	-4.32	2.56	-0.11	0.099		
Provision of practice opportunities							-1.48	0.693
High	477.90	481.13	-3.24	2.39	-0.08	0.183		
Low	477.75	479.51	-1.76	2.85	-0.04	0.540		
<u>Contrast in the use of general teaching practices related to academic language instruction</u>								
Overall							0.95	0.791
High	480.66	482.88	-2.21	2.50	-0.06	0.381		
Low	475.36	478.53	-3.16	2.56	-0.08	0.222		
Word knowledge instruction							-1.16	0.741
High	477.71	480.90	-3.19	2.33	-0.08	0.180		
Low	478.37	480.40	-2.03	2.59	-0.05	0.438		
Academic skill instruction							1.30	0.730
High	480.32	482.73	-2.41	2.30	-0.06	0.301		
Low	473.95	477.65	-3.70	2.94	-0.09	0.215		
Provision of practice opportunities							-8.67	0.016 †
High	480.18	486.53	-6.35 *	2.29	-0.16	0.008		
Low	474.55	472.23	2.32	2.58	0.06	0.374		

(continued)

Exhibit D.16 (continued)

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Contrast in received training</u>								
Overall							-0.31	0.934
High	477.54	480.52	-2.98	2.22	-0.07	0.187		
Low	479.24	481.91	-2.67	3.00	-0.07	0.378		
Program training							2.61	0.547
High	477.45	479.76	-2.31	2.02	-0.06	0.260		
Low	479.55	484.48	-4.92	3.79	-0.12	0.202		
Non-program training							-0.31	0.934
High	477.54	480.52	-2.98	2.22	-0.07	0.187		
Low	479.24	481.91	-2.67	3.00	-0.07	0.378		
<u>Last curricular unit reached</u>								
Overall							-1.93	0.563
High	478.58	482.21	-3.62	1.97	-0.09	0.073		
Low	476.23	477.92	-1.69	2.67	-0.04	0.531		
<u>Available instruction days</u>								
Overall							1.30	0.730
High	480.32	482.73	-2.41	2.30	-0.06	0.301		
Low	473.95	477.65	-3.70	2.94	-0.09	0.215		

SOURCES: The Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. District records data for the 2016-2017 school year.

NOTES: For each implementation feature, random assignment blocks were divided into two groups: blocks with higher than median value in the given feature, and blocks with lower than median value in the feature.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and English Language Arts test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

An F-test was applied to the difference in the estimated impacts between the two subgroups. Statistical significance is indicated by (†) when the p-value is less than 0.05.

Exhibit D.17. Estimated Impacts on State ELA Test Performance, by District-Level Subgroup

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Use of program-specific practices</u>								
Overall							-3.3	0.478
High	26.6	29.5	-2.9	3.25	-0.06	0.371		
Low	26.4	26.0	0.4	3.30	0.01	0.909		
Word knowledge instruction							-4.8	0.366
High	20.5	24.4	-3.9	4.22	-0.08	0.363		
Low	30.6	29.7	0.9	3.11	0.02	0.771		
Academic skill instruction							-3.3	0.478
High	26.6	29.5	-2.9	3.25	-0.06	0.371		
Low	26.4	26.0	0.4	3.30	0.01	0.909		
Provision of practice opportunities							2.1	0.680
High	29.5	29.7	-0.3	3.41	-0.01	0.940		
Low	22.2	24.5	-2.3	3.59	-0.05	0.523		
<u>Contrast in the use of general teaching practices related to academic language instruction</u>								
Overall							-4.8	0.366
High	20.5	24.4	-3.9	4.22	-0.08	0.363		
Low	30.6	29.7	0.9	3.11	0.02	0.771		
Word knowledge instruction							-4.8	0.366
High	20.5	24.4	-3.9	4.22	-0.08	0.363		
Low	30.6	29.7	0.9	3.11	0.02	0.771		
Academic skill instruction							-2.9	0.585
High	22.9	25.9	-2.9	4.39	-0.06	0.507		
Low	28.7	28.7	0.0	3.02	0.00	0.998		
Provision of practice opportunities							-4.8	0.366
High	20.5	24.4	-3.9	4.22	-0.08	0.363		
Low	30.6	29.7	0.9	3.11	0.02	0.771		

(continued)

Exhibit D.17 (continued)

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Contrast in received training</u>								
Overall							2.9	0.585
High	28.7	28.7	0.0	3.02	0.00	0.998		
Low	22.9	25.9	-2.9	4.39	-0.06	0.507		
Program training							3.7	0.432
High	25.2	24.5	0.7	3.43	0.02	0.834		
Low	27.4	30.4	-3.0	3.25	-0.07	0.357		
Non-program training							2.9	0.585
High	28.7	28.7	0.0	3.02	0.00	0.998		
Low	22.9	25.9	-2.9	4.39	-0.06	0.507		
<u>Last curricular unit reached</u>								
Overall							3.2	0.525
High	26.2	25.5	0.7	4.10	0.02	0.866		
Low	26.7	29.2	-2.5	2.76	-0.05	0.375		
<u>Available instruction days</u>								
Overall							3.2	0.525
High	26.2	25.5	0.7	4.10	0.02	0.866		
Low	26.7	29.2	-2.5	2.76	-0.05	0.375		
<u>By implementation start time</u>								
Overall							-0.6	0.897
On time	24.1	25.3	-1.2	3.37	-0.02	0.734		
Late start	29.2	31.0	-1.8	3.34	-0.04	0.599		

(continued)

Exhibit D.17 (continued)

SOURCES: State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: For each implementation feature, districts were divided into two groups: districts with higher than median value in the given feature, and districts with lower than median value in the feature.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was applied to the difference in the estimated impacts between the two subgroups.

Exhibit D.18. Estimated Impacts on State ELA Test Performance, by Random Assignment Block Level Subgroup

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Use of program-specific practices</u>								
Overall							5.7	0.283
High	28.9	28.7	0.2	3.15	0.00	0.950		
Low	19.6	25.1	-5.5	4.16	-0.12	0.195		
Word knowledge instruction							2.1	0.699
High	26.5	27.2	-0.7	2.89	-0.02	0.799		
Low	26.3	29.1	-2.9	4.59	-0.06	0.537		
Academic skill instruction							4.9	0.321
High	28.2	27.7	0.5	3.15	0.01	0.863		
Low	23.3	27.6	-4.4	3.72	-0.09	0.248		
Provision of practice opportunities							5.1	0.316
High	30.0	29.1	0.9	3.36	0.02	0.789		
Low	20.0	24.2	-4.2	3.78	-0.09	0.270		
<u>Contrast in the use of general teaching practices related to academic language instruction</u>								
Overall							-5.2	0.305
High	21.8	25.5	-3.7	3.52	-0.08	0.296		
Low	31.9	30.4	1.5	3.54	0.03	0.682		
Word knowledge instruction							-6.5	0.201
High	22.6	26.3	-3.7	3.31	-0.08	0.269		
Low	31.9	29.1	2.8	3.75	0.06	0.461		
Academic skill instruction							1.8	0.713
High	25.1	25.9	-0.8	3.11	-0.02	0.800		
Low	28.2	30.8	-2.6	3.61	-0.06	0.483		
Provision of practice opportunities							-3.3	0.503
High	24.1	27.1	-3.0	3.32	-0.07	0.370		
Low	29.6	29.3	0.3	3.55	0.01	0.937		

(continued)

Exhibit D.18 (continued)

Subgroup	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Effect Size of Estimated Impact	P-Value of Estimated Impact	Estimated Subgroup Difference	P-Value of Estimated Subgroup Difference
<u>Contrast in received training</u>								
Overall							4.6	0.386
High	28.7	27.8	1.0	3.15	0.02	0.764		
Low	22.9	26.5	-3.6	4.15	-0.08	0.389		
Program training							-4.2	0.476
High	25.0	27.1	-2.1	2.83	-0.05	0.461		
Low	30.6	28.5	2.1	5.09	0.05	0.685		
Non-program training							4.6	0.386
High	28.7	27.8	1.0	3.15	0.02	0.764		
Low	22.9	26.5	-3.6	4.15	-0.08	0.389		
<u>Last curricular unit reached</u>								
Overall							3.3	0.530
High	26.3	26.7	-0.4	3.21	-0.01	0.896		
Low	26.1	29.8	-3.7	4.02	-0.08	0.365		
<u>Available instruction days</u>								
Overall							1.8	0.713
High	25.1	25.9	-0.8	3.11	-0.02	0.800		
Low	28.2	30.8	-2.6	3.61	-0.06	0.483		

SOURCES: State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: For each implementation feature, random assignment blocks were divided into two groups: blocks with higher than median value in the given feature, and blocks with lower than median value in the feature.

The impacts are estimated using two-level hierarchical linear models to account for the nested structure of the data, with students nested within schools. The models control for the blocking of random assignment and for the following baseline variables: gender, grade, age, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

An F-test was applied to the difference in the estimated impacts between the two subgroups.

Exhibit D.19. Program Effects on Teachers' Use of Core Instructional Practices, by Item

Item	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Estimated Impact in Effect-Size Units	P-Value of Estimated Impact
<u>Word knowledge instruction</u>	2.00	1.41	0.59 *	0.15	0.64	0.000
Teacher introduced, reviewed, or called attention to the use of vocabulary word(s)	0.85	0.69	0.16 *	0.05	0.41	0.002
Vocabulary word(s) were visually displayed	0.76	0.44	0.32 *	0.06	0.74	0.000
Teacher referred to/prompted students to use visual display or graphic organizer of vocabulary word(s)/definitions (with skills listed)	0.39	0.28	0.12	0.07	0.30	0.093
<u>Academic skill instruction</u>	2.52	2.43	0.09	0.13	0.08	0.489
Teacher introduced learning objective of the lesson or guiding question	0.73	0.86	-0.12 *	0.05	-0.42	0.015
Students followed norms or teacher reminded students of norms	0.79	0.76	0.03	0.05	0.07	0.590
Teacher reminded students that they must provide reasons and evidence to support their positions	0.68	0.55	0.13 *	0.06	0.30	0.034
Closure was established by summarizing key points or interesting observations	0.33	0.27	0.06	0.04	0.18	0.201
<u>Provision of practice opportunities</u>	5.60	5.47	0.13	0.25	0.07	0.595
All students had access to workbook, textbook, or other curricular material used for learning	0.97	0.86	0.11 *	0.04	0.41	0.005
Teacher prompted students to read text passage (silently or aloud), or teacher read aloud text passage	0.75	0.81	-0.06	0.06	-0.20	0.321

(continued)

Exhibit D.19 (continued)

Item	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Estimated Impact in Effect-Size Units	P-Value of Estimated Impact
Teacher instructed students to read independently (silently, without read-aloud support)	0.42	0.50	-0.08	0.06	-0.19	0.218
IF TEXT READ ALOUD: All students could see text while listening to the teacher reading aloud	0.66	0.68	-0.02	0.06	-0.07	0.686
Teacher modeled/prompted students to use comprehension strategies during reading	0.50	0.69	-0.19 *	0.07	-0.47	0.008
Students participated in a classroom discussion	0.85	0.74	0.11 *	0.05	0.28	0.047
Teacher asked two or more open-ended questions	0.77	0.70	0.07	0.05	0.16	0.192
At least two opinions or viewpoints were represented, reviewed or discussed	0.69	0.50	0.20 *	0.05	0.46	0.000
Total	10.13	9.31	0.82 *	0.39	0.27	0.040
Number of schools	32	25				
Number of classes	93	72				

SOURCE: Classroom observation data collected in the 2017-2018 school year.

NOTES: The sample includes 165 fourth- and fifth-grade regular classrooms from 57 study schools.

The impacts are estimated using linear models that account for the nested structure of the data, with classrooms nested within schools. The models control for the blocking of random assignment. The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test was applied to each estimated impact. Statistical significance is indicated by (*) when the p-value is less than 0.05.

It is of interest to see if the differences in the amount of training and support led to differences in whether and how much teachers adopted these core instructional practices in their classrooms.

Exhibit D.20 presents the estimated contrasts in teachers' use of the instructional practices important for academic language development for these two groups of districts. It also provides the difference in the estimated contrasts between the two groups. These results show that, in the districts that started the program on time, the program increased teachers' overall use of these practices, particularly their use of word knowledge instruction. In contrast, the program did not affect teachers' overall use of such practices, nor did it affect teachers' use of the three core components in the late-start districts. There were, however, statistically significant differences in the estimated contrasts between the two sets of districts for the overall score and the provision of practice opportunities. It is worth pointing out that the levels of usage among the program teachers did not differ much between the on-time districts and the late-start ones across these practices. Therefore, a large part of the difference in the estimated contrasts between them seemed to be driven by the different usage levels among the non-program teachers between these two groups of districts. Regardless of these observed differences in the use of the practices, as mentioned earlier (in Exhibits D.13-D.18), there was no difference in estimated program impacts on student outcomes between the districts that started the program on time and those that started late.

Additional Correlation Analyses

Appendix C presented the estimated relationships between teacher reported professional development (PD) amount and their use of program-specific instructional practices (Exhibit C.7) and the estimated relationships between teachers' instructional practices and student outcomes (Exhibit C.8). The study team also explored other associations among program components to supplement those presented earlier in the report. The same multi-level multi-variate regression framework outlined in Appendix B is used for the analysis presented below.

Associations Between Professional Development Received by Teachers and Student Outcomes

The study team examined the relationship between the amount of professional development teachers reported receiving and student outcomes as measured by their scores on the CALS-I, GMRT, and state reading achievement tests. Findings from this exercise are presented in Exhibit D.21. The first panel of this exhibit presents the estimated relationship between the amount of WordGen-related PD received by the teachers and student outcomes in program schools, since only program-school teachers received WordGen-related PD during the implementation year. The second panel presents the estimated relationship between teacher reported overall PD receipt (both WordGen- and non-WordGen-related) and student outcomes for all study schools. Overall, these findings suggest that there was no clear pattern of association between the amount of PD teachers received and student outcome levels.

Associations Between Use of Instructional Practices and Student Outcomes

The study team explored the relationship between teachers' use of instructional practices generally considered important for academic language development (not specific to the program tested) and student outcomes. The set of practices examined here were not explicitly tied to the program curriculum but were considered generally beneficial for students' academic language development and student outcomes (see Appendix B for details about these two sets of practices). Because these practices could be used and measured in both program and non-program schools, this analysis used classroom observation data from all study schools. Exhibit D.22 shows that there were no significant associations between the level of usage of these practices and the CALS-I and GMRT test scores, but there appeared to be a statistically significant association between the overall use of these practices and students' performance on the state ELA tests. This association seemed to be driven by the use of academic skill instruction.

Exhibit D.20. Program Effects on Teachers' Use of Core Instructional Practices, by Site Starting Time, by Item

Item	Districts that Started the Program on Time				Districts that Started the Program Late				Difference in Impacts Between On-Time and Late-Start Districts	P-Value for Difference in Impacts Between On-Time and Late-Start Districts
	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact		
<u>Word knowledge instruction</u>	2.29	1.44	0.85	0.000 *	1.70	1.43	0.27	0.243	0.58	0.058
Teacher introduced, reviewed, or called attention to the use of vocabulary word(s)	0.85	0.68	0.17	0.032 *	0.84	0.69	0.15	0.019 *	0.02	0.689
Vocabulary word(s) were visually displayed	0.87	0.44	0.43	0.000 *	0.64	0.48	0.17	0.073	0.26	0.059
Teacher referred to/prompted students to use visual display or graphic organizer of vocabulary word(s)/definitions (with skills listed)	0.57	0.32	0.25	0.005 *	0.21	0.26	-0.05	0.626	0.30	0.026 †
<u>Academic skill instruction</u>	2.39	2.23	0.16	0.410	2.66	2.66	-0.01	0.964	0.17	0.585
Teacher introduced learning objective of the lesson or guiding question	0.71	0.79	-0.08	0.227	0.76	0.93	-0.18	0.024 *	0.10	0.413
Students followed norms or teacher reminded students of norms	0.73	0.66	0.07	0.363	0.84	0.88	-0.03	0.460	0.10	0.229

(continued)

Exhibit D.20 (continued)

Item	Districts that Started the Program on Time				Districts that Started the Program Late				Difference in Impacts Between On-Time and Late-Start Districts	P-Value for Difference in Impacts Between On-Time and Late-Start Districts
	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact		
Teacher reminded students that they must provide reasons and evidence to support their positions	0.61	0.52	0.09	0.203	0.74	0.57	0.17	0.095	-0.08	0.450
Closure was established by summarizing key points or interesting observations	0.34	0.26	0.08	0.252	0.31	0.28	0.03	0.574	0.05	0.740
<u>Provision of practice opportunities</u>	5.48	4.84	0.64	0.085	5.73	6.25	-0.51	0.065	1.15	0.017 †
All students had access to workbook, textbook, or other curricular material used for learning	0.96	0.84	0.12	0.041 *	0.99	0.89	0.10	0.049 *	0.02	0.842
Teacher prompted students to read text passage (silently or aloud), or teacher read aloud text passage	0.76	0.73	0.03	0.735	0.74	0.91	-0.17	0.037 *	0.20	0.099

(continued)

Exhibit D.20 (continued)

Item	Districts that Started the Program on Time				Districts that Started the Program Late				Difference in Impacts Between On-Time and Late-Start Districts	P-Value for Difference in Impacts Between On-Time and Late-Start Districts
	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact		
Teacher instructed students to read independently (silently, without read-aloud support)	0.45	0.51	-0.06	0.464	0.39	0.49	-0.10	0.323	0.04	0.705
IF TEXT READ ALOUD: All students could see text while listening to the teacher reading aloud	0.70	0.63	0.07	0.438	0.61	0.75	-0.14	0.117	0.21	0.109
Teacher modeled/prompted students to use comprehension strategies during reading	0.53	0.56	-0.03	0.710	0.47	0.85	-0.39	0.000 *	0.36	0.012 †
Students participated in a classroom discussion	0.80	0.60	0.19	0.033 *	0.90	0.90	0.00	0.929	0.19	0.067
Teacher asked two or more open-ended questions	0.68	0.59	0.10	0.239	0.86	0.83	0.02	0.585	0.08	0.477
At least two opinions or viewpoints were represented, reviewed, or discussed	0.61	0.38	0.23	0.001 *	0.78	0.62	0.16	0.099	0.07	0.438

(continued)

Exhibit D.20 (continued)

Item	Districts that Started the Program on Time				Districts that Started the Program Late				Difference in Impacts Between On-Time and Late-Start Districts	P-Value for Difference in Impacts Between On-Time and Late-Start Districts
	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact	Program Schools	Non-Program Schools	Estimated Impact	P-Value of Estimated Impact		
Total	10.16	8.51	1.65	0.007 *	10.09	10.34	-0.25	0.537	1.90	0.012 †
Number of schools	17	14			15	11				
Number of classes	48	40			45	32				

SOURCE: Classroom observation data collected in the 2017-2018 school year.

NOTES: The sample includes 165 fourth- and fifth-grade regular classrooms from 57 study schools.

The impacts are estimated using linear models that account for the nested structure of the data, with classrooms nested within schools. The models control for the blocking of random assignment. The values in the column labeled "program schools" are the weighted average of the observed district means for students from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test was applied to each estimated impact. Statistical significance is indicated by (*) when the p-value is less than 0.05.

An F-test was applied to the difference in the estimated impacts between the two subgroups. Statistical significance is indicated by (†) when the p-value is less than 0.05.

Exhibit D.21. Relationship Between the Amount of Professional Development Teachers Reported and Student Outcomes

	Academic Language Skills (CALS-I Score)		Reading Comprehension Skills (GMRT Score)		Reading Achievement (Percentage at or Above State Proficiency Level)	
	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value
<u>Program schools</u>						
WordGen training (teacher report)						
Initial training	-0.67	0.031 *	-0.91	0.079	0.51	0.393
Ongoing support	0.76	0.695	-0.63	0.755	-0.44	0.864
<u>Program and non-program schools</u>						
Total training (teacher report)						
Initial training	-0.10	0.055	-0.08	0.147	0.00	0.980
Ongoing support	-0.44	0.427	-1.18	0.077	-0.32	0.747

SOURCES: Teacher survey data collected in the spring of 2018. Core Academic Language Skills Instrument (CALS-I) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: The sample used in the analysis for the top panel includes 30-31 program schools. The sample used in the analysis for the bottom panel includes 56-57 study schools. The student sample includes all fourth- and fifth-graders with non-missing values for respective outcomes and all explanatory variables in a given model.

The correlations between the explanatory variables and each outcome are estimated with a multilevel, multivariate regression with students nested within schools. The regression models also control for the blocking of random assignment and for student background characteristics such as grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The estimated correlation reflects the amount of outcome change (in the unit of the outcome measure) that is associated with one unit of change in a given explanatory variable, controlling for all covariates in the model.

A two-tailed t-test was applied to each estimated impact. Statistical significance is indicated by (*) when the p-value is less than 0.05.

Exhibit D.22. Relationship Between Teachers' Use of Core Instructions and Student Outcomes

	Academic Language Skills		Reading Comprehension Skills		Reading Achievement (Percentage at or Above State Proficiency Level)	
	(CALSI Score)		(GMRT Score)			
	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value	Estimated Coefficient	P-Value
<u>Program and non-program schools</u>						
Total	-0.13	0.782	0.10	0.807	1.32	0.026 *
Word knowledge instruction	-0.45	0.691	-0.55	0.646	1.90	0.223
Academic skill instruction	1.84	0.258	1.27	0.408	3.98	0.028 *
Provision of practice opportunities	-0.58	0.413	0.19	0.770	1.47	0.123

SOURCES: Classroom observation data collected in the spring of 2018. Core Academic Language Skills Instrument (CALSI) data and the Gates-MacGinitie Reading Test (GMRT) data collected in the spring of 2018. State standardized English Language Arts (ELA) test data and district records data for the 2016-2017 and 2017-2018 school years.

NOTES: The sample used in this analysis includes 56-57 study schools. The student sample includes all fourth- and fifth-graders with non-missing values for respective outcomes and all explanatory variables in a given model.

The correlations between the explanatory variables and each outcome are estimated with a multilevel, multivariate regression with students nested within schools. The regression models also control for the blocking of random assignment and for student background characteristics such as grade, age, gender, race and ethnicity, district-provided poverty indicator, English learner status, Individualized Education Plan status, and baseline standardized math and ELA test scores. All missing values in these covariates are imputed with zero and missing indicators for all covariates are also included in the model.

The estimated correlation reflects the amount of outcome change (in the unit of the outcome measure) that is associated with one unit of change in a given explanatory variable, controlling for all covariates in the model.

A two-tailed t-test was applied to each estimated impact. Statistical significance is indicated by (*) when the p-value is less than 0.05.

Program Impacts on Teacher Reported Practices, Attitudes, and Challenges

The study team administered two rounds of surveys in all study schools to collect information from teachers related to their perspectives on program implementation. The study team combined data from these two rounds of surveys for this analysis. For continuous measures, the study team took the average of the responses from the two rounds of surveys. For a binary measure, a new dichotomous variable was created to be equal to one if the response from either round of the survey was one, and zero otherwise. Exhibit D.23 presents the estimated program impacts on teacher-reported usage of instructional practices considered supportive of academic language development. Overall, teachers from both program and non-program schools reported high levels of usage for such practices: for all but three listed practices, over 80 percent of teachers said that they used such practices. About 72 percent to 73 percent of them reported modeling how to use text cues to interpret text. Just below 80 percent of the teachers reported that they modeled how to generate questions and evaluate predictions about the text. In contrast, 57 percent of teachers from both groups reported using culturally appropriate materials and activities in classrooms. There were no differences in the usage of such practices between the program and non-program schools.

The survey also asked teachers about their attitudes regarding training and support in their schools and the challenges they faced in teaching English learners and struggling readers who were native English speakers. The top panel of Exhibit D.24 shows that teachers in the program schools, in general, were more positive about the training and support than teachers in the non-program schools, with the only exception being teachers' perception of the school administration's behavior toward the staff where the estimated difference was the smallest and not statistically significant. This overall pattern indicates that the program might have positively affected teachers' attitudes toward their teaching environment. The second panel presents findings for teacher-perceived challenges in teaching English learners and struggling readers. Teachers in both program and non-program group schools identified the need to modify activities or work to accommodate the needs of these students as the top challenge. With few exceptions, there were no differences in the proportion of teachers choosing a given challenge between the two groups of schools.

Exhibit D.23. Estimated Impacts on Teachers' Self-Reported Use of Instructional Practices that Support Academic Language Development, by Practice

Survey item (%)	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Estimated Impact in Effect-Size Units	P-Value of Estimated Impact
Prompt students to consider different perspectives (for example, ask students to explain different understandings of an event)	88.9	80.2	8.7	4.6	0.23	0.061
Model how to use titles, headers, figures, and other text cues to interpret text	72.6	71.9	0.7	5.0	0.02	0.889
Refer to or elicit students' personal experiences to engage them in a new topic or illustrate a new point	88.9	81.4	7.5	4.4	0.2	0.094
Facilitate classroom discussion by asking students to explain each other's responses and respond directly to each other's claims	86.7	89.9	-3.2	3.6	-0.1	0.385
Develop content-driven class discussions between you and your students or among students to build deeper knowledge	82.2	87.6	-5.4	3.6	-0.17	0.135
Introduce and define key academic and disciplinary language and terms	89.6	89.6	0.0	3.5	0	0.994
Incorporate culturally appropriate materials and activities in the classroom	56.9	57.0	-0.2	8.0	0	0.984
Ask students to define words, use words in a sentence, or state synonyms	90.4	84.4	5.9	3.8	0.17	0.126
Use sentence starters or templates to help students organize their thoughts for writing	90.4	87.0	3.4	2.8	0.1	0.223
Remind students to provide reasons and evidence during classroom discussion	87.2	91.3	-4.2	4.7	-0.16	0.380
Focus on the intended meaning in student talk or writing, not primarily on conventional correctness	82.2	79.6	2.7	5.0	0.07	0.594

(continued)

Exhibit D.23 (continued)

Survey item (%)	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Estimated Impact in Effect-Size Units	P-Value of Estimated Impact
Model how to generate questions and evaluate predictions about the text	77.8	80.9	-3.1	4.3	-0.09	0.469
Ask students questions requiring inferences based on text	92.6	90.8	1.8	3.3	0.06	0.586
Use think-alouds or role plays to model skills and processes (for example, how to use text clues to interpret text)	84.4	85.8	-1.4	4.6	-0.04	0.769
None of the above	1.8	1.5	0.3	1.4	0.03	0.833
Number of schools	31	26				
Number of teachers	135	100				

SOURCE: Teacher survey data collected in the 2017-2018 school year.

NOTES: The sample includes 235 fourth- and fifth-grade regular classroom teachers from 57 study schools. The number of teachers varies by practices due to missing responses. Teacher survey respondents can select multiple instructional practices. Thus, the percentages across practices may add up to more than 100 percent. The number of observations varies by item due to missing values.

The impacts are estimated using linear models that account for the nested structure of the data, with teachers nested within schools. The models control for the blocking of random assignment. The values in the column labeled "program schools" are the weighted average of the observed district means for teachers from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

None of the differences between the program and non-program schools are statistically significant at the 0.05 level.

Exhibit D.24. Estimated Impacts on Teachers' Self-Reported Attitudes and Perceived Challenges

Item	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Estimated Impact in Effect-Size Units	P-Value of Estimated Impact
Teacher attitudes (1- to 4-point scale)						
To what extent do you agree with the following statements?						
I am adequately trained to teach students in my classroom who are ELs.	3.25	2.96	0.30*	0.1	0.41	0.004
Inclusion of ELs in my class has worked well.	3.39	3.12	0.27*	0.09	0.39	0.002
The school administrator knows what kind of school he/she wants and has communicated it to the staff.	3.34	3.02	0.32*	0.14	0.33	0.022
Teachers in this school are continually learning and seeking new ideas.	3.45	3.09	0.36*	0.11	0.45	0.002
Most of the ELs I teach are capable of learning the material I am supposed to teach them.	3.27	2.94	0.33*	0.11	0.40	0.003
The school administration's behavior toward the staff is supportive and encouraging.	3.22	3.04	0.18	0.16	0.19	0.278
Teacher-reported challenges (%)						
Challenges in providing effective instruction to English learners						
Language barriers (for example, different language, dialect, speaking nonstandard English) between myself and the student	46.2	57.7	-11.5	7.3	-0.23	0.120
Need to modify classroom activities or work to accommodate ELs' needs	62.2	64.3	-2.1	7.5	-0.04	0.775
Lack of a formal policy or procedures for instructing ELs	31.1	33.6	-2.5	6.7	-0.05	0.708
Lack of training in instructional strategies for improving ELs' reading and writing	34.5	38.7	-4.3	6.3	-0.09	0.499
Lack of support from administration for meeting ELs' needs	26.1	23.9	2.2	6.5	0.05	0.740
Other staff members who do not share similar ideas about how to teach ELs	10.9	13.6	-2.7	4.7	-0.08	0.567
Other	5.9	13.9	-8.1*	3.8	-0.24	0.040
None of the above	26.9	16.1	10.8	5.9	0.28	0.072

(continued)

Exhibit D.24 (continued)

Item	Program Schools	Non-Program Schools	Estimated Impact	Standard Error of Estimated Impact	Estimated Impact in Effect-Size Units	P-Value of Estimated Impact
Challenges in providing effective instruction to non-EL struggling readers						
Language barriers (for example, different dialect, speaking nonstandard English) between myself and the student	19.4	13.5	5.8	5.3	0.17	0.273
Need to modify classroom activities or work to accommodate struggling readers' needs	54.3	63.4	-9.1	6.2	-0.18	0.147
Lack of a formal policy or procedures for instructing struggling readers	22.5	26.9	-4.4	5.1	-0.1	0.384
Lack of training in instructional strategies for improving struggling readers' reading and writing	26.4	39.6	-13.3*	6.2	-0.27	0.038
Lack of support from administration for meeting struggling readers' needs	19.4	22.8	-3.4	5.9	-0.08	0.565
Other staff members who do not share similar ideas about how to teach struggling readers	7.8	14.6	-6.8	5.0	-0.19	0.178
Other	8.5	8.7	-0.2	3.3	-0.01	0.949
None of the above	38.0	26.3	11.6	7.2	0.26	0.111
Number of schools	31	26				
Number of teachers	131	97				

SOURCE: Teacher survey data collected in the 2017-2018 school year.

NOTES: The sample includes 228 fourth- and fifth-grade regular classroom teachers from 57 study schools. The number of teachers varies by practices due to missing responses. Teacher survey respondents can select multiple challenges. Thus, the percentages across the perceived challenges may add up to more than 100 percent. The number of observations varies by item due to missing values.

EL = English learner.

The impacts are estimated using linear models that account for the nested structure of the data, with teachers nested within schools. The models control for the blocking of random assignment. The values in the column labeled "program schools" are the weighted average of the observed district means for teachers from the program schools (using number of program schools in each district as weight). The non-program schools' values are calculated by subtracting the estimated impacts from the program school averages. Rounding may cause slight discrepancies in calculating sums and differences.

The estimated impacts' effect sizes are calculated as a proportion of the standard deviation of the full non-program school analysis sample.

A two-tailed t-test with a null of zero impact is reported, and statistical significance (rejection of the null) is indicated by (*) when the p-value is less than 0.05.

ENDNOTES

¹Pianta, Hamre, and Mintz (2012a).

²Uccelli et al. (2015a); Uccelli et al. (2015b).

³See Nagy and Townsend (2012); National Research Council (2010).

⁴Uccelli et al. (2015a); Barr and Uccelli (2016).

⁵Uccelli et al. (2015a); Uccelli et al. (2015b); Barr, Uccelli, and Galloway (2019).

⁶For example, Johnson (2005) and McCabe (2005) reported a test-re-test reliability of above 0.88. Ulhrich and Swalm (2007) showed a reliability of 0.93.

⁷Most teachers in the sample were observed two separate times. However, many teachers in one of the later starting districts were observed only once in spring 2018.

⁸Pianta, Hamre, and Mintz (2012a).

⁹Kane and Staiger (2012); Grossman, Loeb, Cohen, and Wyckoff (2013); Pianta, Hamre, and Mintz (2012a).

¹⁰Pianta, Hamre, and Mintz (2012b) reports that average inter-rater reliabilities for exact plus adjacent percent agreement is 68 percent to 95 percent across domains. Across the three studies, internal consistencies for emotional support has an alpha value of 0.88; Classroom organization has an alpha value of 0.88; and Instructional support has an alpha value of 0.90.

¹¹Puma, Olsen, Bell, and Price (2009).

REFERENCES

- Barr, Christopher D., and Paola Uccelli. 2016. *CALS-I Psychometric Report*. Unpublished internal report prepared for the IES-funded Catalyzing Comprehension through Discussion and Debate research project.
- Barr, Christopher D., Paola Uccelli, and Emily Phillips Galloway. 2019. "Specifying the Academic Language Skills That Support Text Understanding in the Middle Grades: The Design and Validation of the Core Academic Language Skills Construct and Instrument." *Language Learning* 69, 4: 978-1021. DOI: [10.1111/lang.12365](https://doi.org/10.1111/lang.12365).
- Grossman, Pamela, Susanna Loeb, Julia Cohen, and James Wyckoff. 2013. "Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." *American Journal of Education* 119, 3: 445-470.
- Johnson, Kathleen M. 2005. "Test Review of Gates-MacGinitie Reading Tests(r), Fourth Edition, Forms S and T." Pages 4-8 in Robert A. Spies and Barbara S. Plake (eds.), *The Sixteenth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute Inc.
- Kane, Thomas J., and Douglas O. Staiger. 2012. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Seattle, WA: Bill & Melinda Gates Foundation.
- McCabe, Patrick P. 2005. "Test Review of Gates-MacGinitie Reading Tests(r), Fourth Edition, Forms S and T." Pages 8-12 in Robert A. Spies and Barbara S. Plake (eds.), *The Sixteenth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute Inc.
- Nagy, William, and Dianna Townsend. 2012. "Words as Tools: Learning Academic Vocabulary as Language Acquisition." *Reading Research Quarterly* 47, 1: 91-108. DOI: [10.1002/RRQ.011](https://doi.org/10.1002/RRQ.011).
- National Research Council. 2010. *Language Diversity, School Learning, and Closing Achievement Gaps: A Workshop Summary*. Washington, DC: The National Academies Press. DOI: [10.17226/12907](https://doi.org/10.17226/12907).
- Pianta, Robert C., Bridget K. Hamre, and Susan L. Mintz. 2012a. *Classroom Assessment Scoring System: Upper Elementary Manual*. Charlottesville, VA: Teachstone.
- Pianta, Robert C., Bridget K. Hamre, and Susan L. Mintz. 2012b. "Upper Elementary and Secondary CLASS Technical Manual." Website: https://cdn2.hubspot.net/hubfs/336169/Technical_Manual.pdf.
- Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price. 2009. *What to Do When Data Are Missing in Group Randomized Controlled Trials (NCEE 2009-0049)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Uccelli, Paola, Christopher D. Barr, Christina L. Dobbs, Emily Phillips Galloway, Alejandra Meneses, and Emilio Sánchez. 2015a. "Core Academic Language Skills: An Expanded Operational Construct and a Novel Instrument to Chart School-Relevant Language Proficiency in Preadolescent and Adolescent Learners." *Applied Psycholinguistics* 36, 5: 1077-1109. DOI: [10.1017/S014271641400006X](https://doi.org/10.1017/S014271641400006X).
- Uccelli, Paola, Emily Phillips Galloway, Christopher D. Barr, Alejandra Meneses, and Christina L. Dobbs. 2015b. "Beyond Vocabulary: Exploring Cross-Disciplinary Academic-Language Proficiency and its Association with Reading Comprehension." *Reading Research Quarterly* 50, 3: 337-356. DOI: [10.1002/rrq.104](https://doi.org/10.1002/rrq.104).
- Uhrich, Tabatha A., and Ricky L. Swalm. 2007. "A Pilot Study of a Possible Effect from a Motor Task on Reading Performance." *Perceptual and Motor Skills* 104, 3: 1035-1041. DOI: [10.2466/pms.104.3.1035-1041](https://doi.org/10.2466/pms.104.3.1035-1041).