National Geographic Learning's Reach for Reading Program: An Efficacy Study

Final Report

July 31, 2013

magnolia consulting

cultivating learning and positive change

www.magnoliaconsulting.org

Executive Summary

Cengage contracted with Magnolia Consulting, LLC, an independent evaluation consulting firm, to conduct an efficacy study of National Geographic Learning's Reach for Reading program in the third grade. Magnolia Consulting conducted this study in seven schools with 28 teachers and 580 students during the 2012–2013 school year. The purpose of this study was to evaluate the efficacy of Reach for Reading in increasing third-grade students' reading and writing skills. This study also included an examination of teachers' implementation of Reach for Reading and comparison reading and writing curricula.

Reach for Reading

Reach for Reading is a comprehensive K–5 Common Core reading program. The program features authentic, multicultural literature paired with content from National Geographic and real-world accounts from the National Geographic Explorers. The program was built around Common Core State Standards (CCSS). It includes eight units, each with four weeks of instructional plans for whole group, small group, and independent reading time. The program offers materials for reading and writing with special emphasis on academic vocabulary and academic talk.

Study Design and Methods

Evaluators used a randomized control trial design in which teachers were randomly assigned to treatment and comparison groups. Student measures for the efficacy study included (a) the Gates-MacGinitie Reading Test, Fourth Edition (GMRT-4) as an assessment of reading vocabulary and comprehension (b) the DIBELS Next Oral Language Fluency test, and (c) the Reach for Reading Common Core benchmark assessment. Teacher measures included weekly implementation logs, classroom observations, and interviews.

Program Implementation

KEY QUESTION:

Did teachers implement the curriculum according to the implementation guidelines and with a high level of fidelity?

Reach for Reading teachers demonstrated a high level of fidelity in implementing the required program components with an overall fidelity rating of 94%. As part of their implementation, teachers differentiated instruction with small groups, engaged students in academic talk, and used vocabulary and reading teaching routines during every lesson or most lessons, on average.

Study Results

KEY QUESTION:

Did treatment students in Reach for Reading classrooms demonstrate significant learning gains in reading achievement scores after one year of implementation?

Overall, teachers thought Reach for Reading supported them in addressing the Common Core State Standards to a great extent, and particularly in the areas of emphasizing academic language vocabulary and using more informational text. Treatment teachers reported during interviews that students' scores on the Common Core benchmark assessments were lower than expected. In general, they attributed this to the high bar set by the CCSS for writing and reading informational text skills. Based on teacher interviews, the majority of students had not experienced the level of writing rigor and stamina reflected in the standards prior to Reach for Reading.

Average treatment student scores on all portions of the benchmark assessments ranged from 45% to 60% correct at pretest and 49% to 76% correct at posttest. Student gains from pretest to posttest on the benchmark Reading test were statistically significant, $t(278) = 6.58$, $p < .001$. A decrease in student scores on the benchmark Writing test also was statistically significant, $t(278) = -3.78$, $p < .001$. Caution is warranted when interpreting these results because of assessment validity issues.

KEY QUESTION:

Did the Reach for Reading program significantly impact treatment students' reading achievement compared to comparison students' achievement after one year of implementation?

There was a statistically significant difference in treatment and comparison students' scores on the GMRT-4 Vocabulary and Total Reading tests. The difference in scores on the Comprehension test was not statistically significant. Effect sizes were 0.20 for Vocabulary (see Figure 1), 0.12 for Comprehension, and 0.14 for Total Reading (see Figure 2), which translates to the average treatment student scoring eight, five, and six percentile points higher than the average comparison student, respectively.

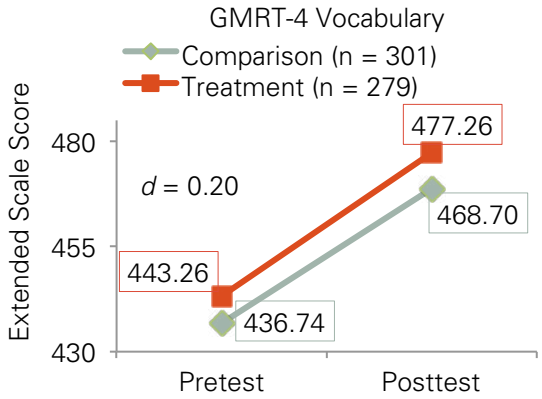

Figure 1. Pretest and posttest adjusted Vocabulary means by condition.

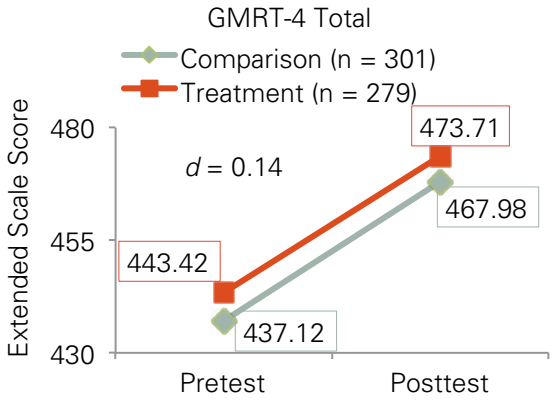

Figure 2. Pretest and posttest adjusted Total Reading means by condition.

There were no statistically significant differences in the number of treatment and comparison students demonstrating an oral reading fluency level at or above benchmark at the end of the study. The majority of students in each group met or exceeded the spring benchmark of reading 100 or more words correct per minute with 69% of treatment students and 60% of comparison students demonstrating fluency at this level. The odds of scoring at or above benchmark in spring were 1.41 times greater for Reach for Reading students than for comparison students. There were no statistically significant differences in oral reading fluency performance for subgroups of students qualifying for free- or reduced-price lunch (FRL) or Limited English Proficient (LEP) students.

KEY QUESTION:

Were there differential effects between treatment and comparison student subgroups?

Reach for Reading had a statistically significant and positive impact on FRL students' performances on the GMRT-4 Vocabulary test with an effect size of 0.28 and a percentile difference of 11 points between the average treatment student and the average comparison student (see

Figure 3). Although not statistically significant, the effect size for FRL students on the Comprehension test was 0.15 and 0.20 on the Total Reading test, which are considered substantively important positive effects by the U.S. Department of Education's What Works Clearinghouse (What Works Clearinghouse, 2008).

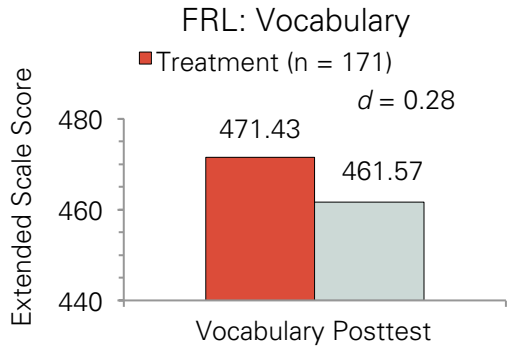

Figure 3. Posttest Vocabulary adjusted means for FRL students by condition.

For the LEP student subgroup, there were statistically significant differences in student performance on the GMRT-4 Vocabulary and Total Reading tests (see Figure 4 and Figure 5). Effect sizes for both of these tests were moderate with 0.57 for Vocabulary and 0.40 for Total Reading. Using an improvement index, this translates to the average LEP treatment student scoring 22 percentile points higher than the average comparison student on the Vocabulary test and 16 percentile points higher on the Total Reading test. The effect size of 0.23 on the Comprehension test is considered substantively important, although not statistically significant, and translates to a nine percentile-point difference between the average treatment and comparison student.

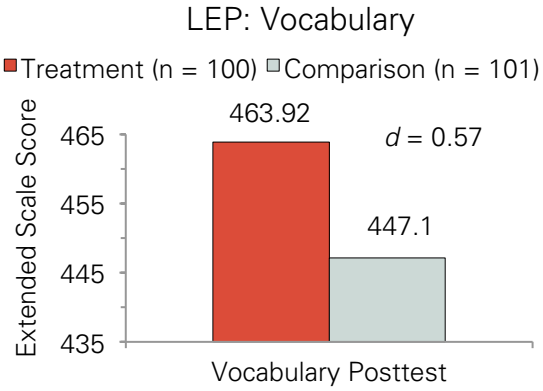

Figure 4. Posttest Vocabulary adjusted means for LEP students by condition.

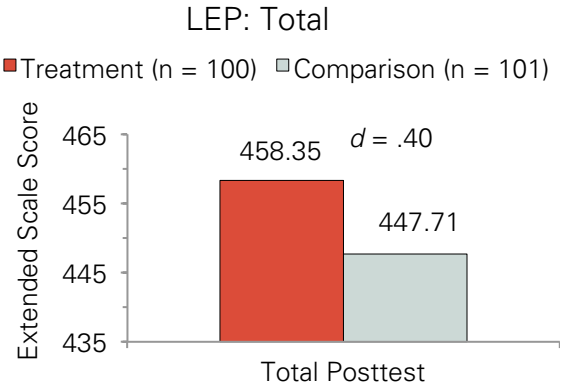

Figure 5. Posttest Total adjusted reading scores for LEP students by condition.

Overall. Through a rigorous, well-implemented randomized control trial, this study found that Reach for Reading has a statistically significant positive effect on student reading outcomes. This positive effect also is evident on reading outcomes for students with limited English proficiency. The program also positively impacts vocabulary outcomes for low-income students. Treatment students' oral reading fluency gains were comparable to those of comparison students.

Acknowledgements

The National Geographic Learning Reach for Reading evaluation represents the collaborative efforts of many individuals from Cengage and participating schools. We would like to express our deepest gratitude to the teachers and administrators who dedicated an extensive amount of time and effort to the study. We greatly valued your feedback and truly appreciated the opportunity to work with you. Second, we would like to thank the staff at Cengage for their ongoing support and understanding of evaluation, and especially Barbara Quincer-Coulter for her time, support, and management of the study. We would also like to express our gratitude to Candy Carro, who served as the Reach for Reading trainer and who gave her time to ensuring that teachers had the support they needed to implement the program with fidelity. Finally, we would like to extend our sincerest thanks to team members who contributed to making this a well-implemented study: Dr. deKoven Pelton, Beth Peery, Candace Rowland, and Beverly Bunch. Thank you also to Dr. Lisa Shannon for providing analytical assistance and to Sara Bullard for editing.

The authors,

Stephanie Wilkerson, Ph.D.
Monica Savoy, M.S.

Magnolia Consulting, LLC
5135 Blenheim Rd.
Charlottesville, VA 22902
(ph) 855.984.5540 (toll free)
http://www.magnoliaconsulting.org

Table of Contents

Tables

Figures

Introduction

Developed in 2010 by the National Governor's Association and the Council of Chief State School Officers, the Common Core State Standards (CCSS) reflect a state-led effort to establish a shared set of clear educational standards for English language arts and mathematics. The CCSS increase the rigor of literacy learning to make certain that high school graduates will have the skills and knowledge necessary for entering institutions of higher education and a globally competitive workforce (National Governors Association, Council of Chief State School Officers, & Achieve, 2008). The CCSS engage students in deep learning rather than shallow coverage of material (McTighe and Wiggins, 2012).

The CCSS for language arts involve vertical alignment or "staircasing" of skills and knowledge across K–12 to achieve the ultimate goals of college and career readiness in graduating students. For elementary schools, shifts in the new English language arts standards include, among others:

- a 50/50 split between narrative and informational texts;
- a focus on independent reading of high-quality, increasingly complex text;
- a focus on academic and domain-specific vocabulary with an emphasis on vertical alignment of vocabulary skills;
- use of text-dependent questions;
- providing text-based evidence of opinions
- the use of speaking and listening skills to communicate and collaborate;
- emphasis on disciplinary literacy that integrates literacy and content knowledge (Liebling & Meltzer, 2011).

National Geographic Learning (NGL) developed Reach for Reading based on these Common Core shifts. The program provides students with reading instruction with structured and scaffolded opportunities that aim to equip students with the academic language, literacy, and writing skills they need. Cengage contracted with Magnolia Consulting, LLC, an independent evaluation consulting firm, to conduct an efficacy study of Reach for Reading in the third grade. Magnolia Consulting conducted this study in 7 schools with 28 teachers and 580 students during the 2012–2013 school year. The purpose of this study was to evaluate the efficacy of Reach for Reading in increasing third-grade students' reading and writing skills. This study also included an examination of teachers' implementation of Reach for Reading and comparison reading and writing curricula.

Evaluation Design

The purpose of this study was to evaluate the efficacy of the Reach for Reading materials in helping elementary students improve their reading and writing skills. The study focused on third-grade students in average performing schools across the country. The evaluation also assessed teachers' implementation of the Reach for Reading program. The evaluation study employed a randomized controlled trial (RCT) design in which evaluators randomly assigned teachers within the same school to either use the Reach for Reading program (treatment) or their current reading and writing program (comparison). This design allowed evaluators to make scientifically-based claims about the impact of Reach for Reading on student reading achievement.

The study addressed the following overarching evaluation questions:

1. Did teachers implement the curriculum according to the implementation guidelines and with a high level of fidelity?

2. Did treatment students in Reach for Reading classrooms demonstrate significant learning gains in reading achievement scores after one year of implementation?

3. Did the Reach for Reading program significantly impact treatment students' reading achievement compared to comparison students' achievement after one year of implementation?

4. Were there differential effects between treatment and comparison student subgroups?

Methodological Approach

Evaluators used a RCT in which teachers were randomly assigned to treatment and comparison groups. This design allows evaluators to make estimations of the difference between student performance in treatment and comparison classrooms and to determine if their difference is significant (Raudenbush, Spybrook, Liu, & Congdon, 2005). To further strengthen the validity of the study in making causal inferences, evaluators employed multiple student outcome measures during different time periods during the study.

Measures

A combination of quantitative and qualitative measures was included in the design to allow for a full understanding of (a) how the program impacted participating students compared to non-participating students; (b) whether the program resulted in desired outcomes; (c) how the Reach for Reading program was implemented with students; and d) how teachers perceived the quality and utility of the materials.

Student Measures

Student measures for the Reach for Reading efficacy study included (a) the Gates-MacGinitie Reading Test, Fourth Edition (GMRT-4) as an assessment of reading vocabulary and comprehension (b) the DIBELS Next Oral Language Fluency test, and (c) the Reach for Reading Common Core benchmark assessment.

GMRT-4, Level 3, Forms S and T

Teachers administered the GMRT-4 as a pre/post assessment of reading vocabulary and comprehension. The Level 3 subtests for third-grade students include Vocabulary and Comprehension. The 45-item Vocabulary subtest requires students to choose from a list of five words or phrases the one word whose meaning is closest to the test word. Students are given 20 minutes to complete the Vocabulary subtest. The Comprehension subtest consists of 39 items that require students to read stories and nonfiction passages each divided into short segments. Comprehension passages in Level 3 reflect various content categories, including fiction, social science, natural science and humanities, and are presented in both narrative and expository text formats. The student's task is to choose the picture that illustrates the reading segment or that answers a question about it. Students are given 35 minutes to complete this subtest. Scores on each sub-test of the GMRT-4 are converted to extended-scale scores that are derived from raw scores, and place achievement scores along an equal-unit scale.

DIBELS Next

Teachers also administered the DIBELS Next Oral Reading Fluency (DORF) measure at the beginning and end of the study period to all students. It measures advanced phonics and word attack skills and fluent reading of connected text. The ORF takes one minute to individually administer and less than five minutes to score and record.

Common Core State Standards Benchmark Assessment

As part of the Reach for Reading program, treatment teachers administered the Common Core State Standards benchmark assessment at the beginning, middle (optional), and end of the study period. The assessment is a measure of students' acquisition of the knowledge associated with the Common Core standards. Only treatment students took this assessment. The reading test includes 44 multiple-choice items and six constructed response items. Although untimed, it takes an estimated 90 minutes to administer, and teachers could divide administration time across two days. The writing test includes 12 multiple-choice and four writing prompts. Administration of the writing test also could be divided across multiple days based on the testing needs and conditions for students.

Teacher Measures

To measure program implementation and teacher perceptions, evaluators collected data through a combination of online and on-site data collection efforts with participating teachers. Treatment teachers completed weekly implementation logs, whereas comparison teachers completed a one-time implementation survey. The implementation data provide important information about the nature of teachers' reading instruction, their use of reading instructional

materials, and their perceptions of the effectiveness of their reading materials (whether treatment or comparison) in improving students' reading vocabulary, comprehension, and fluency. Evaluators conducted classroom observations and interviews during the fall of 2012 and the spring of 2013 with all treatment teachers and a sample of comparison teachers (spring only). Both the implementation logs and the comparison teacher survey were administered using Magnolia's SurveyGizmo survey software system. Together, these measures increase the validity of findings by (1) triangulating data through multiple data collection methods; (2) capturing the perspectives of various participants; and (3) collecting data throughout the project period (Erickson, 1986).

Teacher Implementation Log and Survey

Participating treatment teachers completed weekly online implementation logs that gauged the breadth and depth of their use of their reading materials and instructional practices. Logs took less than ten minutes to complete each week. Treatment teachers indicated (1) the frequency and extent to which they implemented specific Reach for Reading components and materials, (2) how often they used the program's additional resources, and (3) their perceptions about the Reach for Reading program, including its support of Common Core State Standards. The final implementation log at the end of the year included additional open-ended questions pertaining to (1) the classroom learning environment, including important characteristics of their school culture and student population that influence the learning context; (2) perceptions of program strengths and challenges; (3) changes in instructional practices; (4) perceptions of the professional development support they received; and (5) observations of student impacts (i.e., learning and motivation) during the study period. Data from the logs were aggregated at the end of the study period to arrive at a rating of teachers' level of implementation. Teachers' overall implementation ratings were used in the analysis of student performance on outcome measures as well as to describe teachers' fidelity of implementation. Teachers were encouraged to follow implementation guidelines in order to implement the program with high fidelity.

Through a one-time survey administered during the spring of 2013, comparison teachers responded to similar questions about the reading materials they used with students. These data allowed evaluators to document the materials, components, and instructional practices students received in comparison classrooms.

Classroom Observations and Interviews

During the fall of 2012 and spring of 2013, evaluators observed and interviewed treatment teachers. The purpose of the observations and interviews was to identify any implementation challenges, document the various ways teachers implemented the Reach for Reading program, and to provide an objective measure of implementation fidelity. Having two observation data points in the fall and spring allow for a more robust measure of fidelity. The fall site visit also allowed evaluators to identify whether treatment teachers needed additional professional development support to implement Reach for Reading with fidelity. Given the purpose of these observations, evaluators developed an observation protocol that aligned to the structure, components, and intended use of the Reach for Reading program. In order to ensure the protocol reflected best practices in reading instruction, evaluators referred to extant observation protocols that have been used in rigorous research studies. Interviews focused on

teachers' perceptions and experiences implementing Reach for Reading in order to provide contextual information regarding implementation and teacher capacity.

During the spring of 2013, evaluators observed and interviewed a sample of comparison teachers within each district. These observations and interviews helped triangulate data collected from the comparison teachers survey and gave contextualized examples of the reading programs that comparison students received.

Procedures

Magnolia Consulting worked with participants to ensure they completed all requested data collection activities and implemented all required program components with fidelity. This section presents the procedures used for site selection, data collection, and training and implementation.

Site Selection

With referrals from Cengage, Magnolia Consulting recruited and selected sites for the study. Cengage identified districts with high interest in using Reach for Reading, but that had not implemented it in schools. Cengage aimed to identify socio-economically diverse districts that were geographically distributed across the country. The recruitment process occurred through August of 2012. Once Cengage identified participating schools within a district, Magnolia Consulting vetted the site, secured a signed memorandum of understanding, and conducted the random assignment of teachers to conditions using IBM SPSS statistical software.

Data Collection Timeframe

Study sites began implementing the Reach for Reading program as soon as Cengage trained teachers, which spanned between late August to early September 2012. Implementation occurred through May of 2013 with data collection activities scheduled according to district and school schedules and the study's reporting deadline. Table 1 presents the timeline of data collection activities for each study task.

Table 1. Timeline of data collection activities

TASK AND ACTIVITY	August	September	October	November	December	January	February	March	April	May
Training, study orientation, study begins	→									
Administration of student measures		→							→	
Administration of implementation logs		◆	◆	◆	◆	◆	◆	◆	◆	
Administration of treatment teacher survey								◆		
Spring observations and interviews								→		
End study										◆

Implementation Fidelity

Magnolia Consulting monitored implementation throughout the study period and engaged in ongoing communication with teachers to ensure they were participating with fidelity based on implementation guidelines. At the beginning of the study, Cengage and Magnolia Consulting conducted onsite study orientations with each school and provided each participant with a study orientation folder containing a study schedule, instructions, and an informed consent form. Cengage also conducted half-day onsite trainings with participants in each school followed by another training visit 6–8 weeks after teachers started using Reach for Reading. The program training supported implementation fidelity by orienting teachers to the required program components. Magnolia Consulting monitored implementation fidelity using the log reports, allowing the program trainer to provide follow-up support to ensure that teachers progressed through the materials as expected.

Settings

This study took place in four school districts in the North-Central and Southeast regions of the country. Across the districts, seven schools participated. Table 2 presents overall demographic information for each of the four participating school districts. Districts represented a range of sizes and demographics, which enhanced the generalizability of findings to districts with similar demographics. District A is a mid-size city district comprised of mostly African American and Caucasian students and a 4% ELL population. District B is a mid-size suburban district made up of predominately low-income students and minority students, including 28% ELLS. District C is a mostly low-income, small suburb comprised of predominately Caucasian and Hispanic students and a 6% ELL population. District D is a small, ethnically diverse suburb made up of predominately Caucasian, Asian/Pacific Islander, and Hispanic students and has an ELL population of 32%.

Table 2. District characteristics

	District A	District B	District C	District D
Geographic location and city description*	Southeast City	North Central Suburb	North Central Suburb	North Central Suburb
Total student enrollment	43,654	16,462	2,750	3,537
Student/Teacher ratio	14.57	15.58	23.39	13.93
Percent qualifying as low-income	51.8%	62%	70.2%	47.4%
Ethnic breakdown				
Caucasian	44.0%	4.4%	54.5%	35.0%
African American	45.8%	16.3%	15.7%	4.5%
Asian/Pacific Islander	1.4%	1.6%	0.7%	35.1%
Hispanic	6.9%	75.6%	28.1%	23.5%
Other	1.9%	2.2%	1.0%	2.0%
English Language Learners	1,886	4,690	171	1,143

* City description as defined by the National Center for Educational Statistics (NCES) at http://nces.ed.gov/ccd/commonfiles/localedescription.asp#NewLocale

Teacher Participants

The study included 14 treatment teachers and 14 comparison teachers for a total of 28 participating teachers. The final student analysis sample for the study included 580 students (279 treatment and 301 comparison).

All treatment teachers participated in observations and interviews in the fall of 2012 and spring of 2013, and provided weekly log implementation data. A sample of seven comparison teachers across schools participated in an observation and interview only in the spring of 2013, and all 14 comparison teachers completed a one-time survey. As a benefit of study participation, all study treatment teachers received $300 and three program professional development sessions with a National Geographic Learning trainer. Comparison teachers received $200 for their participation in the study, and the study site coordinators in each district received $200 for assisting with assessment and product distribution. Before beginning the study, teachers and coordinators signed an informed consent form indicating their understanding of study requirements.

The majority of study teachers held a master's degree (53.6%) or bachelor's degree (42.9%) and had been teaching for an average of 9.64 years. Teachers had anywhere from 11 to 29 students in their classrooms, with an average of 22.21 students.

To ensure that teachers were comparable, researchers conducted *t*-tests to determine if treatment and control teachers' demographic information differed significantly. This analysis revealed that the treatment and comparison teachers were comparable with regard to the number of students per teacher, $t(26) = -.42$, $p = .677$. Additionally, there was no significant difference between treatment and comparison teachers for the total number of years teaching, $t(26) = .83$, $p = .413$, and the number of years at their current school, $t(26) = -1.33$, $p = .196$.

Student Participants

The following section describes attrition analyses in the overall student sample, presents student demographics in the analysis sample, and discusses group equivalence.

Attrition

Evaluators conducted two types of attrition analyses: overall sample attrition and differential attrition. Evaluators measured overall sample attrition by determining the number of students who began and completed the study, based on student classroom rosters and available student data. The overall sample attrition rate was 6.9%.

Evaluators measured differential attrition by calculating attrition rates for treatment and comparison samples and by conducting chi-square analyses to determine if these rates were statistically different from each other. The attrition rate for the treatment sample was 9.1%, and the attrition rate for the comparison sample was 4.7%. The differential attrition rate was 4.4%. A Chi-square analysis revealed that there was a statistically significant difference in attrition by condition $\chi 2$ (1, 623) = 4.34, p = .031, such that the treatment students had a significantly higher rate of attrition than the comparison students. However, because overall attrition was less than 10% and the differential attrition rate between treatment and comparison groups was less than 6%, the attrition for this study falls within acceptable levels based on the What Works Clearinghouse (WWC) standards (What Works Clearinghouse, 2011).

Student Demographics

The CONSORT model describes sample flow from pretest to posttest and shows the total number of students included in the analysis sample (see Appendix A). Evaluators included students in the analysis sample if the student had consented to participate, enrolled in school before the study cut-off (January 1, 2013), and was still enrolled at the end of the study. Based on these inclusion criteria the analysis sample consisted of 580 students (279 treatment and 301 comparison).

Table 3 details demographic information for students in the analysis sample. Approximately one-half of the students (50.5%) were male and one-half (49.5%) were female. Across treatment conditions, 39.3% were Caucasian, 11.2% of students were African American, 31.9% were Hispanic, 3.1% were Asian, 0.2% were Alaskan Native or American Indian, and 14.3% were categorized as either multiracial or other. In the total sample, 67.8% of students qualified for free or reduced-priced lunch and 34.7% were classified as Limited English Proficient (LEP). Additionally, 3.8% of the sample included special education students.

Table 3. Student demographics by group

Characteristics	*Comparison Students (n = 301)* Percent	*n*	*Treatment Students(n = 279)* Percent	*n*	*Total Students (n = 580)* Percent	*n*	*Chi-square Results* Value	Sig. (alpha = 0.05)
Gender								
Male	49.5%	149	49.5%	138	50.5%	287	0.00	.992
Female	50.5%	152	50.5%	141	49.5%	293		
Ethnicity								
African-American	8.0%	24	63.1%	41	11.2%	65	17.5	.004
Hispanic	30.2%	91	33.7%	94	31.9%	185		
Asian	1.7%	5	4.7%	13	3.1%	18		
Caucasian	42.9%	129	35.5%	99	39.3%	228		
Alaskan Native or American Indian	0%	0	0.4%	1	0.2%	1		
Other or Multiracial	17.3%	52	11.1%	31	14.3%	83		
Socio-economic status								
Free/Reduced Lunch	73.8%	222	61.3%	171	67.8%	393	10.2	.001
Non-FRL	26.2%	79	38.7%	108	32.2%	187		
English Proficiency								
LEP	33.6%	101	35.8%	100	34.7%	201	.33	.563
Non-LEP	66.4%	200	64.2%	179	65.3%	379		
Special Education								
Special Ed.	3.0%	9	4.7%	13	3.8%	22	1.1	.293
Non-Special Ed.	97%	292	95.3	266	96.2%	558		

Group Equivalency

To ensure the validity of the study's findings, it is important to demonstrate treatment and comparison-group equivalence regarding student demographic characteristics and pretest performance. Based on WWC recommendations, researchers conducted analyses to establish baseline equivalence of the analysis sample. Specifically, as shown in Table 4, evaluators conducted chi-square analyses to assess the equivalence between treatment and comparison groups in the analysis sample by examining differences in student demographic characteristics. These analyses demonstrated that males and females were equally likely to be in the treatment and comparison groups, as were students with LEP and students in special education. Students of various ethnicities were not equally likely to be in the treatment and comparison groups. Additionally, in the comparison group there was a greater percentage of students qualifying for free or reduced-price lunch. According to the chi-square analyses, there are statistically significant treatment group differences in ethnicity and free and reduced lunch. Evaluators also conducted HLM analyses to determine the equivalence between treatment and comparison groups in the analysis sample by examining differences in student pretest performance. These analyses revealed no statistically significant differences between groups on mean GMRT-4 Total Reading scores or DIBELS Next (see Table 4). To account for preexisting differences in demographics, evaluators used pretest achievement student-level

and school-level covariates in analyses (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007).

Table 4. Group equivalence at pretest

Outcome Measure	Coefficient	Standard Error	*t*-Value	Approx. df	*p*-Value
Pretest DIBELS	0.07	0.07	0.88	26	0.39
Pretest GMRT-4	6.18	7.50	0.82	26	0.42

Program Description

National Geographic Reach for Reading

Reach for Reading is a comprehensive K-5 Common Core reading program. The program features authentic, multicultural literature paired with content from National Geographic and real-world accounts from the National Geographic Explorers. The program was built around Common Core State Standards, which requires a shift in teaching practices to engage students. Reach for Reading addresses these necessary shifts by providing more informational text; shared responsibility for literacy across content areas; an emphasis on academic language and vocabulary; increased text complexity; text-dependent questions; and argumentation and text-based evidence. In addition to the National Geographic informational texts, students also have access to exclusive National Geographic videos and a complete digital library with images from National Geographic.

The program consists of eight units, each of which includes four weeks of instruction. Each unit within Reach for Reading is organized around a Big Question (e.g., What is so amazing about plants?) and based on a science or social studies theme aimed at developing students' knowledge of the content areas. The Reach for Reading instructional day balances whole group instruction, small group reading time, and independent reading practice. The reading selections have a balance of National Geographic informational text and multicultural literature that work together to build a coherent body of knowledge within and across grades.

Program materials include a Teacher Edition, student anthology, fiction and nonfiction leveled readers, National Geographic Explorer books, and teacher planning resources. The program also provides materials for learning stations including cross-curricular and language and literacy teamwork activities (flipcharts), and practice masters. To supplement instruction as needed Reach for Reading offers the Phonics Kit. NGReach.com offers a digital library, student eEdition, interactive whiteboard materials, a comprehension coach, and students' individual vocabulary notebooks. Program assessments include an oral reading test, progress monitoring assessments, the Common Core benchmark assessments as well as a variety of other assessment, scoring, and reporting tools.

Comparison Programs

Program A

Program "A" is a comprehensive core basal reading and writing program for grades K–6. This program teaches decoding, comprehension, inquiry and investigation, and writing. There are also applications for teaching spelling, vocabulary, grammar, usage, mechanics, penmanship, phonemic awareness, phonics, fluency, comprehension listening, speaking, and basic computer skills. Materials include a teacher edition containing lesson plans and suggestions for differentiation, pre-decodable and decodable texts, a student anthology, a reading and phonics package student workbook, blackline masters, unit assessments, multimedia support for students and teachers, intervention support, English language development support, and home connections for parents.

Program B

Program "B" is a comprehensive K–6 reading and writing program. Daily small groups provide an opportunity to meet the needs of different groups of students including struggling readers, advanced learners, and English language learners. The program features reading genres such as current fiction, poetry, and nonfiction. The program covers decoding skills, fluency, grammar, comprehension, spelling, phonics and vocabulary. Comprehensive writing instruction and practice are also integrated with the program literature. The writing aspects of the program include sentence fluency, persuasive writing, and informational writing. Students practice daily prewriting, drafting, and revising with writing prompts tied to the reading selections. Teachers use technology for planning and management, instruction, and student activities. Diagnostic, formative, and summative assessments are included. Program materials include leveled readers, vocabulary readers, decodable readers, instructional cards, work stations, student books, and practice books.

Program C

Program "C" is a K–8 reading curriculum that uses nonfiction and fiction read-aloud books to teach students comprehension strategies and social skills. This program is meant to replace or enhance the comprehension component of another literacy program through instruction for ELLs, assessments, and professional development. Teachers use read-aloud trade books, articles, selections for discussions, guided practice, and individualized daily reading. The program also includes a teacher's manual, program orientation materials, student response books, trade books, assessment materials, CD-ROMs, and additional grade-level-specific components. An optional vocabulary supplement teaches strategies for unlocking word meanings.

Program D

Program "D" is an online reading program that provides phonics, vocabulary, fluency, and comprehension instruction. The program incorporates small-group instruction with 15 to 25 minute sessions that begin with introducing a book, eliciting prior knowledge, and building background. The program contains several leveled readers spanning informational texts,

projectable books, serial books, decodable books, read-alouds, humor books, and trade books. The program books can also be multilevel and/or translated in Spanish, French, and British English. There are vocabulary books specially designed for ELLs. Some books have a supplemental Common Core lesson plans. In addition to the leveled books, the program includes worksheets, graphic organizers, online access for students, and assessment tools. Teacher resources include tips for differentiating instruction, planning tools, a video library, and live professional development webinars.

Program E

This reading program offers a collection of trade books, phonics components, ESL/Title 1 libraries, and other integrated language arts components. Teachers also have access to assessment materials and integrated technology designed to promote literacy. Comparison classrooms that used this program mostly used the student anthology and focused on reading comprehension through worksheets.

"Homegrown" Interventions

Homegrown interventions are various materials and practices intended to help students become better readers through balanced literacy, differentiated learning, and independent reading strategies. Homegrown interventions used in comparison schools focused on reading comprehension, language skills, grammar, vocabulary, and guided reading and writing. Many of the homegrown materials consisted of various worksheets and Internet resources.

Program Implementation and Perceptions

Evaluators measured program implementation during the 2012–2013 school year through weekly online logs, direct classroom observations, and interviews with treatment and comparison teachers. Evaluators conducted additional classroom observations for a sample of comparison teachers. When teachers did not implement Reach for Reading during a particular week because of testing or other instructional interruption, such as school closings, they indicated this on their logs. Those logs were not included in analyses.

Implementation of Reach for Reading Program in Treatment Classrooms

KEY QUESTION:

Did teachers implement the curriculum according to the implementation guidelines and with a high level of fidelity?

Evaluators examined teachers' use of the Reach for Reading program compared to implementation benchmarks based on typical program implementation and the implementation guidelines established for the study. Treatment teachers completed weekly implementation

logs comprised of questions about program use, program perceptions, and student engagement in the program. Within these logs, teachers provided feedback on their experiences with the Reach for Reading program. The 14 treatment teachers completed a total of 443 weekly logs for an average of 31.6 weekly logs per teacher and an overall response rate of 100%.

Evaluators calculated an implementation fidelity score for each treatment teacher by examining data from the weekly logs and classroom observations. To calculate an overall implementation percentage score for each Reach for Reading teacher, evaluators compared teachers' weekly log reports of program implementation to an established implementation benchmark for 14 log items. Implementation percentages from the teacher logs were allowed to be greater than 100% to account for teachers who implemented some aspects of the program above and beyond the expected benchmarks. The implementation percentage from the teacher logs was averaged with the total implementation percentage from classroom observations to create an overall implementation fidelity score for each teacher.

Overall, treatment teachers met implementation fidelity requirements in both the classroom observations and the weekly logs, with an overall combined average implementation fidelity score of 94% (See Table 5). Treatment teachers scored high across all observation categories, with an overall average classroom observation score of 92%. Treatment teachers scored an average of 96% on their weekly logs. On the logs, teachers most often had slight difficulties meeting with students reading significantly below grade level (77%), as well as implementing structured response teaching routines (77%). Treatment teacher implementation fidelity level did not have a statistically significant impact on treatment student reading gains during the study period.

Table 5. Reach for Reading program implementation levels

Implementation Level	Overall
High (90-100%)	9 teachers
Moderate (80%–89%)	4 teachers
Low (70%-79%)	1 teacher

Reach for Reading Implementation

Reach for Reading Units

The Reach for Reading curriculum includes eight units for teachers to cover over the course of the school year. Evaluators did not ask teachers to complete a minimum of units during the study, but did ask teachers to complete the units in numerical order. The weekly logs showed that the majority of teachers were able to complete all units up to Unit 6 (see Table 6). Several teachers began Unit 7 (64%) but did not complete the unit. None of the teachers were able to start the final unit, Unit 8.

Table 6. Percentage of treatment teachers who completed Reach for Reading units

Unit	Percentage of teachers completing the unit
1. Happy to Help	100%
2. Nature's Balance	100%
3. Life in the Soil	100%
4. Let's Work Together	100%
5. Mysteries of Matter	100%
6. From Past to Present	93%
7. Blast! Crash! Splash!	0%
8. Getting There	0%

Reach for Reading Instruction

On average, treatment teachers used the Reach for Reading program 3.84 days per week (range 0–5 days). On a typical day of instruction, 65% of the weekly logs indicated that teachers spent *more than 90 minutes* implementing Reach for Reading, while 25% of the logs indicated teachers spent *90 minutes* on instruction, and 10% of the logs indicated teachers spent *less than 90 minutes* on instruction (Figure 6). Additionally, teachers reported spending an average of 108.94 minutes to plan and prepare for their Reach for Reading lessons each week (range 0–650 minutes).

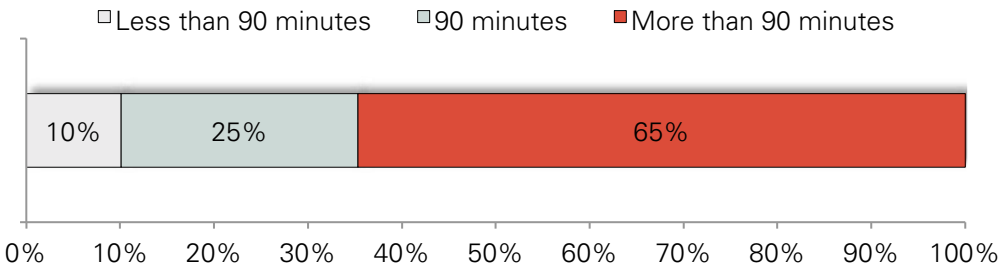

Figure 6. Treatment teachers' amount of time spent daily on Reach for Reading instruction (n = 443).

Instructional Materials

Throughout the week, teachers used several materials to support classroom instruction. Most often teachers used the teacher edition (4.15 days per week), student anthology (3.49 days per week) and small group reading books (3.08 days per week). Teachers least often used Reach into Phonics (0.40 days per week), cross-curricular teamwork activities (0.88 days per

week), and language and literacy teamwork activities (1.03 days per week), all of which are supplemental materials and not part of core instruction (see Figure 7).

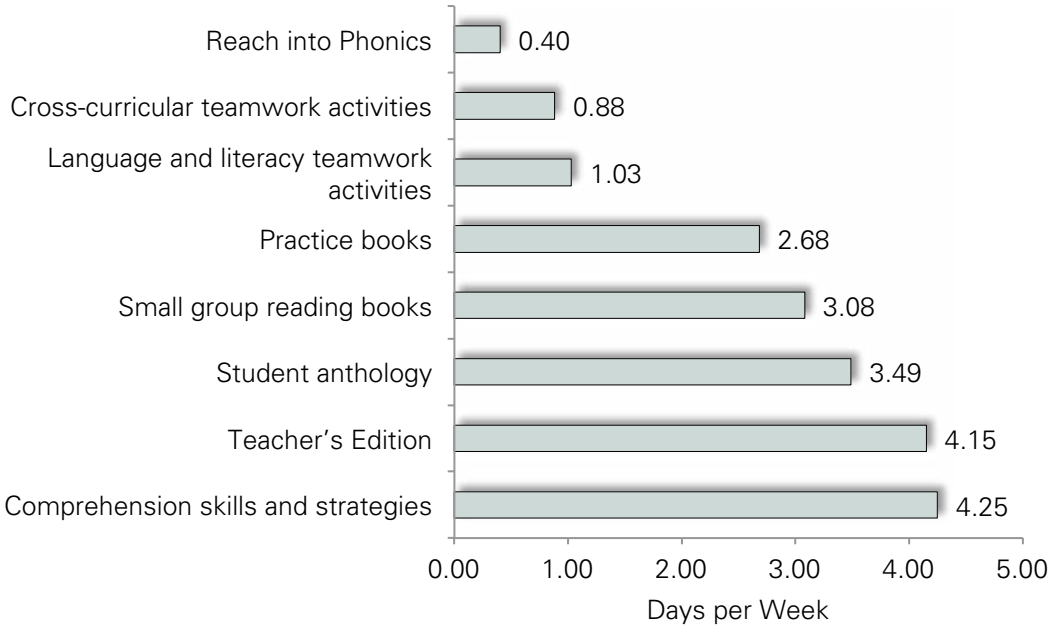

Figure 7. Average days per week treatment teachers used instructional materials (n = 392–399 logs).

Teachers indicated whether they used supplemental materials during each week that they implemented Reach for Reading. In 13% of the logs, teachers reported supplementing the Reach for Reading program with additional materials. These materials included:

1. SRA decodable books *(n = 10)*
2. Additional lesson resources (worksheets, notecards, discussion cards, etc.) *(n = 9)*
3. Comprehension activities *(n = 4)*
4. Went outdoors to enhance lessons related to nature *(n = 2)*
5. Additional literary examples *(n = 2)*
6. Online materials *(n = 2)*
7. Science kit *(n = 1)*
8. Music *(n = 1)*
9. Writing program *(n = 1)*

On the weekly logs, teachers indicated how helpful various Reach for Reading components were in reinforcing and extending key learning concepts for students (see Figure 8). Of all of the weekly logs, 38% stated the practice masters were *very helpful*, 56% stated NGReach.com was *very helpful*, 69% stated the small group reading books were *very helpful*, 19% stated the language and literacy flip charts were *very helpful* and 15% stated the cross-curricular flip charts were *very helpful*. Both the language and literacy flip charts and the cross-curricular flip charts were supplemental teamwork activities that teachers could use as learning

center activities; therefore, as stated previously, not all teachers implemented these materials consistently.

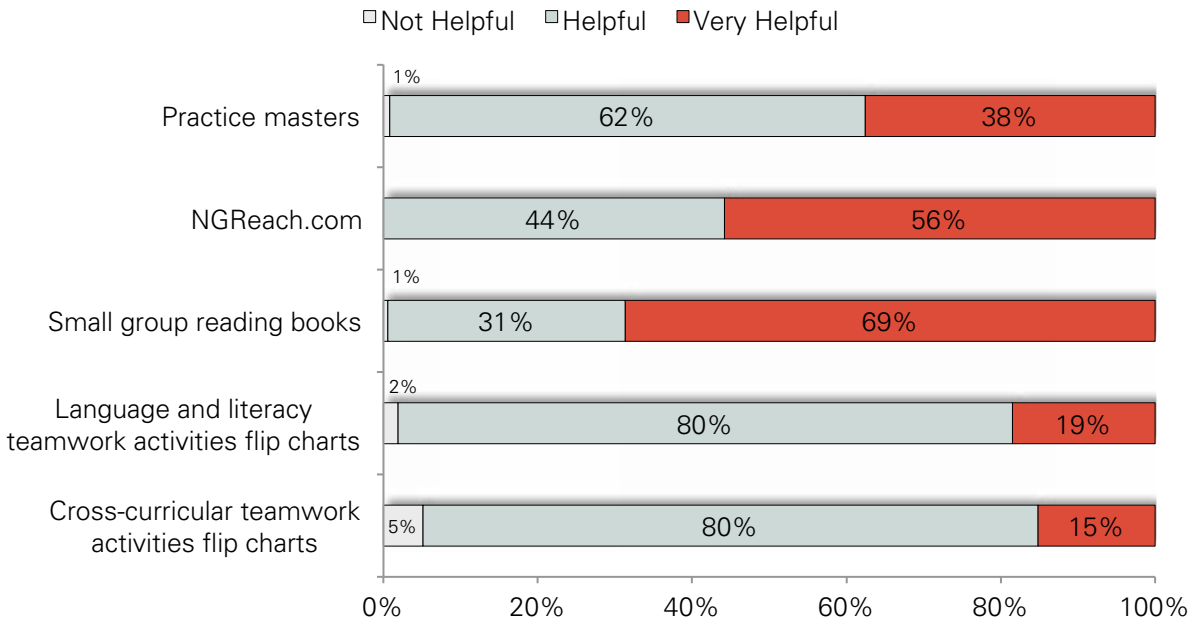

Figure 8. Teachers' perceptions of the helpfulness of Reach for Reading components in reinforcing and extending key learning concepts (n = 392–399).

One teacher reflected on usefulness of the Reach for Reading materials:

> *"We are able to use visual, auditory, and kinesthetic modalities every day. Students who need more practice and/or time for writing, reading, or hands-on activities can be identified and assisted and monitored."*
> (Treatment teacher interview, spring 2013)

Digital Instruction

In addition to the print instructional materials of the Reach for Reading program, many teachers also indicated using digital resources. With the exception of one teacher, almost all reported using a digital resource at least once over the course of the school year. Five teachers did not have digital whiteboards in their classrooms, and were therefore unable to use many of the digital materials as part of whole class instruction. Table 7 displays the percentage of weekly logs where teachers indicated using a particular digital resource.

Table 7. Percentage of treatment teacher digital resource use (n = 399 logs).

Digital Resource	Percent of Logs Indicating Use
Teacher's eEdition	42%
Student eEdition	35%
Digital library	22%
Build background video	17%
Comprehension Coach	10%
My Vocabulary Notebook	7%
Vocabulary games	27%
Online lesson planner	23%
Interactive whiteboard lessons	22%
Other	4%

In 4% of logs, teachers indicated using "other" digital resources during instruction. The "other" digital resources included:

1. Learning stations (video, web article) *(n = 7)*
2. eVisuals *(n = 4)*
3. Phonics games *(n = 1)*
4. Student resources *(n = 1)*
5. Mark up models *(n = 1)*
6. Word builder *(n = 1)*
7. mp3 recording *(n = 1)*
8. Read with Me *(n = 1)*

A teacher reflected on the Reach for Reading technology:

> "Reach for Reading *is highly motivating to students. I like how technology is fully incorporated into the program. Even if I could not utilize much of the technology in my teaching, students get the practice they need to read and navigate a blog or website. This is important for students to learn at this young age." [Treatment Teacher log]*

Small Group Instruction

Small group student instruction is an essential component of the Reach for Reading curriculum. During small group instruction, teachers differentiate instruction using leveled readers with homogenous ability groups. On average, teachers met with small groups between two and three days per week (see Figure 9). Teachers grouped students based on their reading level as two grades below grade level, one grade below grade level, on grade level, or above grade level. For some teachers, small group instruction was a new classroom strategy for meeting the reading needs of students with different reading abilities.

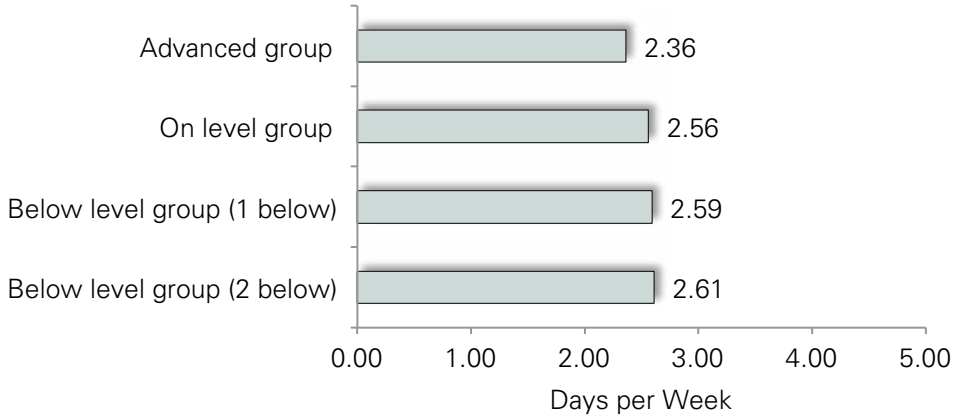

Figure 9. Average days per week treatment teachers met with small groups (n = 298–339 logs).

Teachers reflected on the adequacy of Reach for Reading in meeting the needs of each of the small groups of students (see Figure 10). Overall, 92% of the weekly logs indicated that Reach for Reading *adequately* or *very adequately* addressed the needs of above level students; 81% indicated it *adequately* or *very adequately* addressed the needs of on level students; 67% indicated the program *adequately* or *very adequately* addressed the needs of students reading one grade level below; and 59% indicated it *adequately* or *very adequately* addressed the needs of students reading two grade levels below.

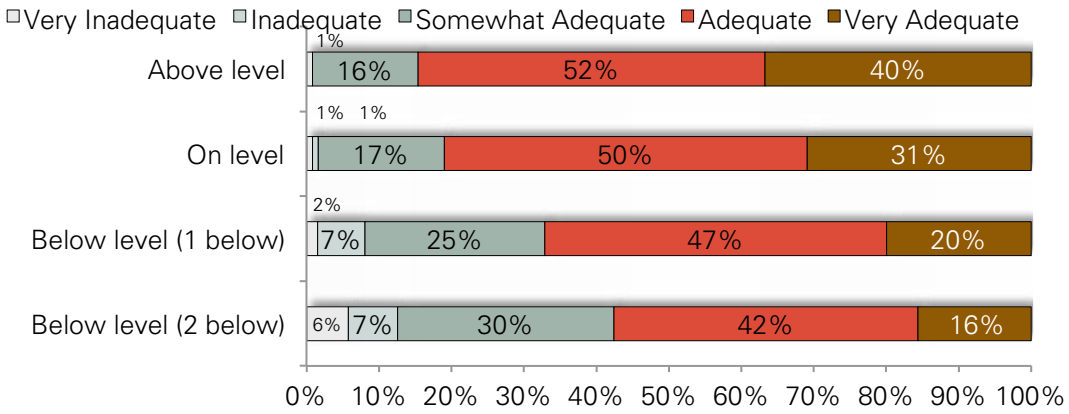

Figure 10. Adequacy of Reach for Reading in meeting the needs of small groups (n = 295–379).

One teacher stated:

> *"Having* Reach for Reading *every day has added to my understanding of differentiated instruction and how important that is for my ELLs and lower students….It has given me a broader sense and understanding of what they need in order to be able to succeed in the classroom and beyond."*
>
> [Treatment teacher interview, spring 2013]

Teaching Routines

Teachers indicated how often they implemented the key Reach for Reading instructional routines. Across all weekly logs, teachers reported that they used reading (71%), vocabulary (64%), writing (54%), and structured response teaching routines (52%) for every lesson or most lessons each week. Teachers were less likely to use cooperative learning (38%) and technology routines (7%) in every lesson or most lessons each week (Figure 11).

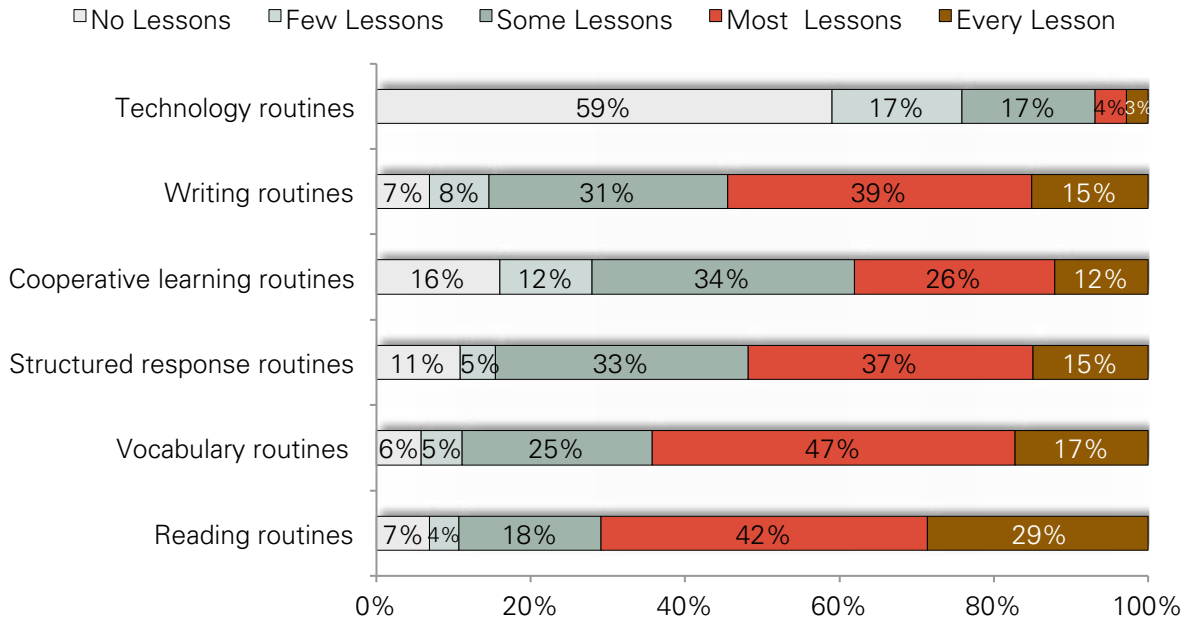

Figure 11. How frequently teachers implemented teaching routines with students (n = 388–395).

Assessments and Reteaching Tools

Teachers rated the effectiveness of Reach for Reading assessments. The majority of teachers' log ratings indicated that Reach for Reading assessments were *effective* or *very effective* in helping teachers identify learner differences (72%); use data to guide instruction (70%); assess how well students organize and learn new content (68%); assess how well students meet the Common Core State Standards (71%); and measure students' application of new strategies (71%). (Figure 12).

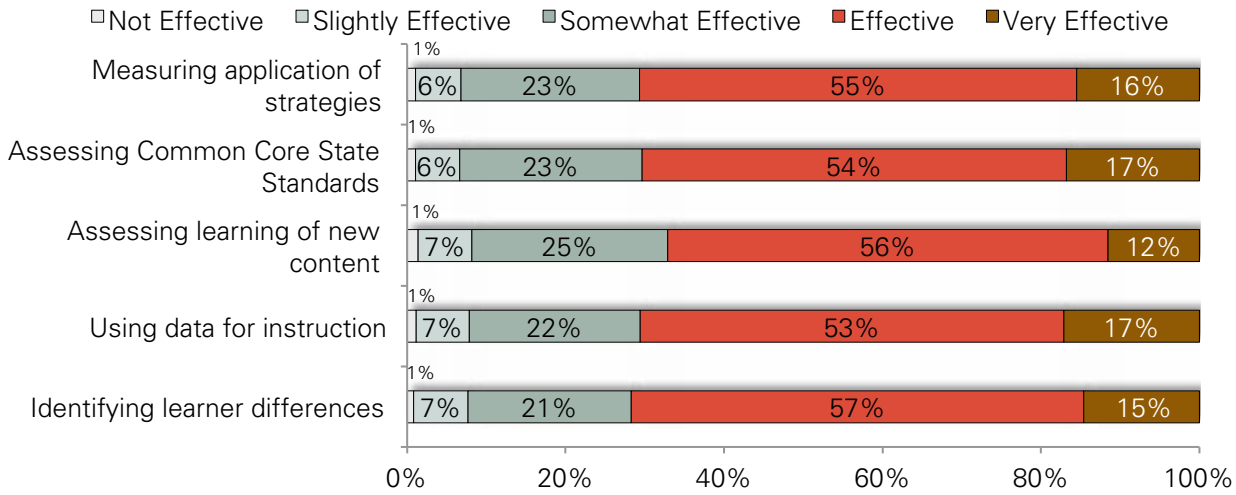

Figure 12.Treatment teachers' perceptions of the effectiveness of the Reach for Reading assessment and reteaching tools (n = 364–368 logs).

Student Learning

Overall, teachers indicated that the Reach for Reading curriculum was *effective* or *very effective* in improving student learning in several academic areas. In all areas, 66%–78% of teacher log responses indicated that Reach for Reading was *effective* or *very effective* (Figures 13–16). Teacher ratings for writing were slightly lower, on average, than the other academic areas.

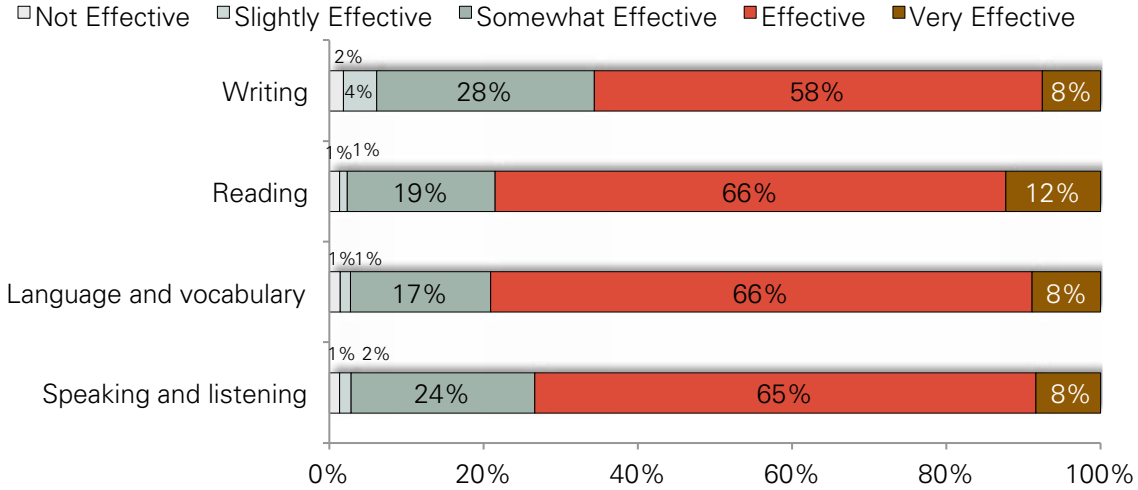

Figure 13. Teachers' perceptions of the effectiveness of the Reach for Reading program in improving student learning (n = 391–393 logs).

Speaking and Listening

Overall, teachers rated the Reach for Reading program as *effective* or *very effective* in improving students' academic talk (77% of log responses) and active listening skills (71% of log responses). (Figure X).

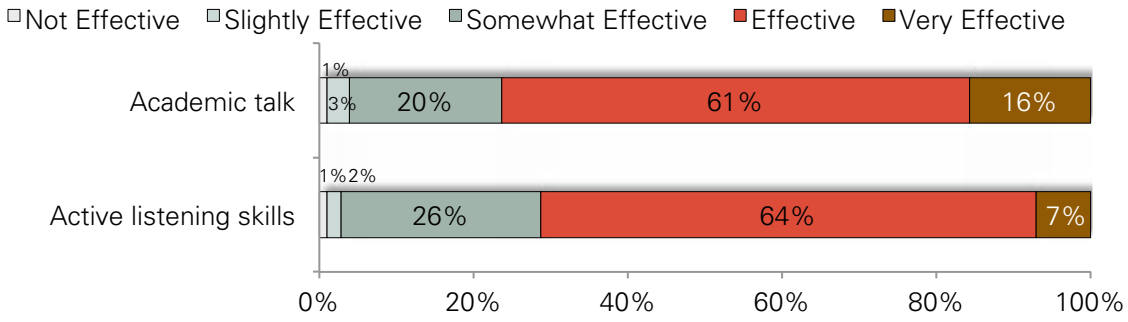

Figure 14.Teachers' perceptions of the effectiveness of the Reach for Reading program in improving student learning in speaking and listening areas (n = 381 logs).

Language and Vocabulary

Overall, teachers rated the Reach for Reading program as *effective* or *very effective* in improving student content vocabulary (78% of log responses), academic vocabulary (81% of log responses), grammar skills (65% of log responses), and spelling and word work (75% of log responses). (Figure X).

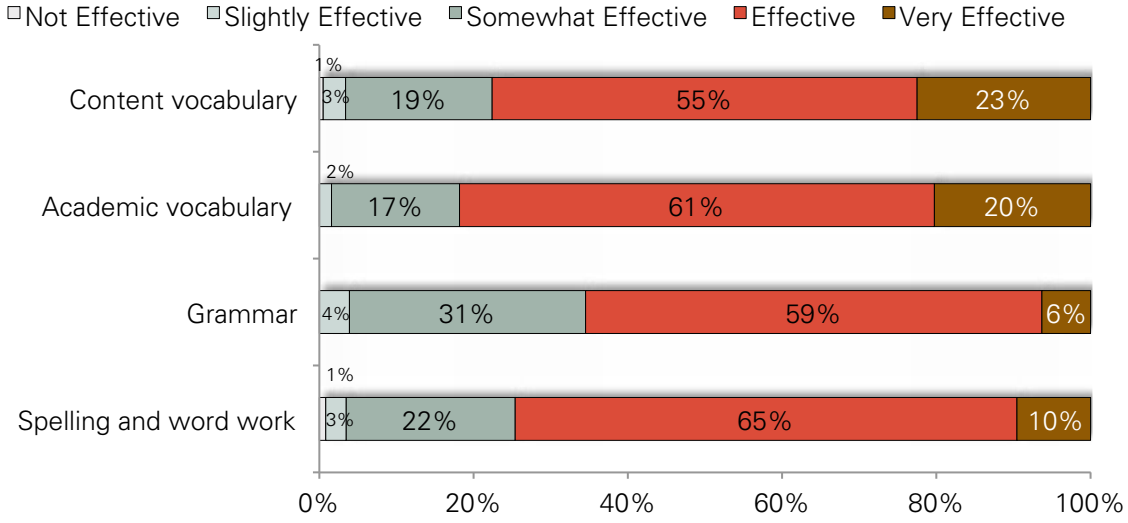

Figure 15. Teachers' perceptions of the effectiveness of the Reach for Reading program in improving student learning in vocabulary areas (n = 375-381 logs).

Reading and Writing

Teachers indicated through 57% and 75% of their log responses that Reach for Reading was *effective* or *very effective* in improving fluency and comprehension, respectively.

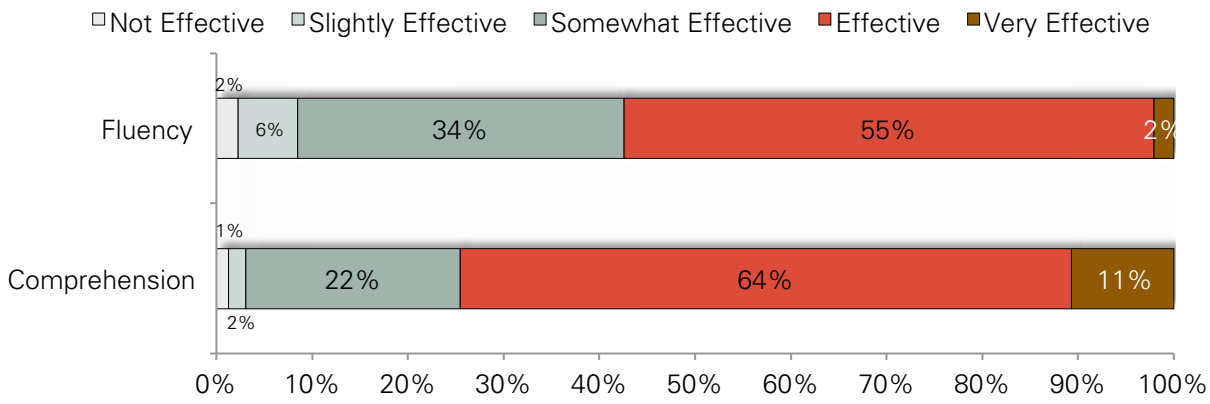

Figure 16. Teachers' perceptions of the effectiveness of the Reach for Reading program in improving student reading fluency and comprehension (n = 390-393 logs).

One teacher commented on the student impacts of Reach for Reading:

> *"I felt that the integration of the academic vocabulary across the curriculum was a huge bonus for me as a teacher as well as for my students. I did not have to spend time scrounging around for extra materials to supplement my teaching. I feel that the way the writing was integrated into the reading made my students more comfortable with the writing process as it became a daily expectation for them to respond to what they had read. It was a natural transition for them to begin to write about what they were talking about."* [Treatment teacher log]

Program Perceptions

In their weekly logs, and fall and spring interviews, teachers offered feedback on their perceptions of the Reach for Reading program implementation, materials, and ability to meet student academic needs, as well as their perceptions of student engagement.

Teacher-Related Perceptions

Over time, most teacher logs indicated that teachers thought the Reach for Reading program contained more material than they could cover (74%). (Figure 17). This is to be expected given that the program offers comprehensive and extensive resources to meet teachers' particular needs and therefore is not intended to be implemented in its entirety. No teacher indicated that there was *not enough* material to cover. When asked about pacing, 60% of logs indicated that the Reach for Reading program was *reasonably paced*, 30% of logs indicated that the program was *fast paced,* and 10% indicated slow paced (see Figure 18). These results reflect teachers' use of the program during the first year of implementation.

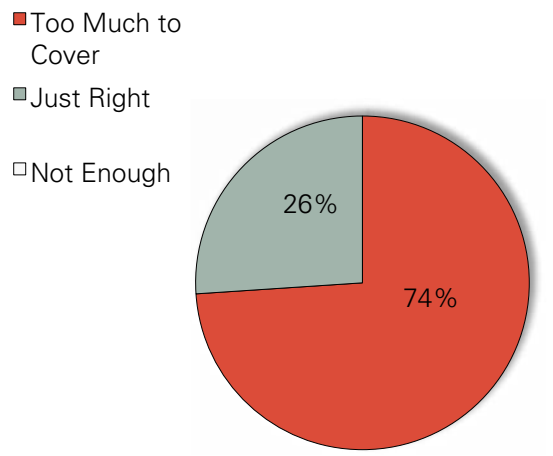

Figure 18. Treatment teachers' perceptions of the amout of program material (n = 393).

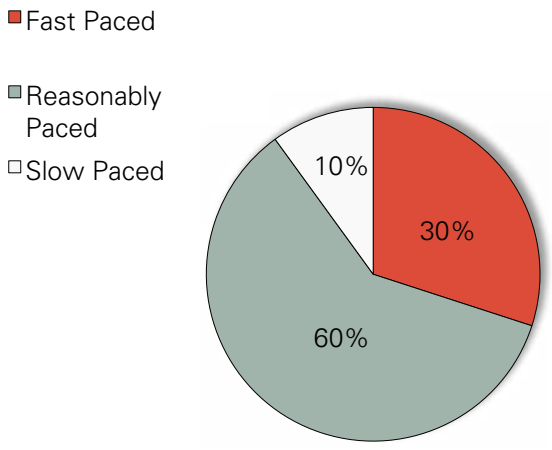

Figure 17. Treatment teachers' perceptions about the pace of the program (n = 382).

Common Core State Standards

On the final weekly log, treatment teachers indicated the extent to which Reach for Reading supported them in implementing various aspects of the Common Core State Standards (see Figure 19). Overall, teachers reported that Reach for Reading supported them in implementing the Common Core State Standards. The logs indicated that the program was effective *to a great extent* at supporting the use of informational text (86%), emphasizing academic language vocabulary (71%), promoting literacy development across content areas (57%), asking students to respond to text-dependent questions (57%), promoting critical thinking (57%), and increasing text complexity (43%).

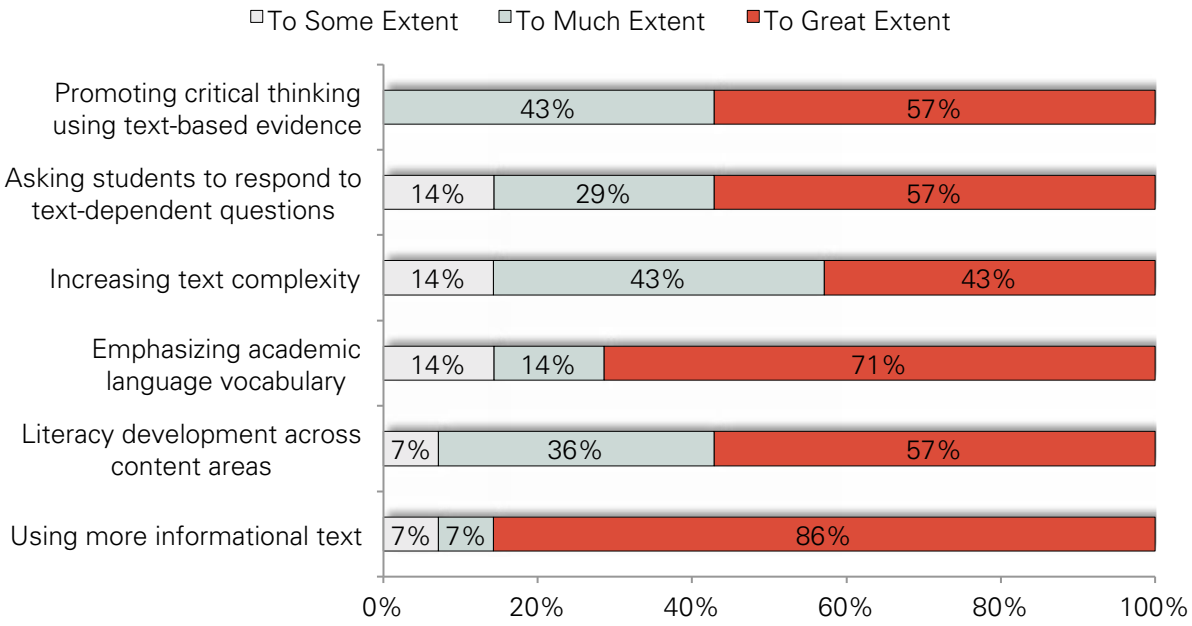

Figure 19. Extent to which Reach for Reading provided support in implementing various aspects of the Common Core State Standards (n = 14).

One teacher gave an overall perception of the Reach for Reading program:

> *"I enjoyed using the* Reach for Reading *materials. I loved teaching with authentic literature. Science, social studies and math are embedded into the curriculum. The themes tied everything together. The students loved the grammar games! The materials get the students motivated to read, even the students with lower reading abilities. The students learned a lot of vocabulary, which is very important in learning new concepts and overall reading comprehension. I liked having many materials at my finger tips to teach with. R4R includes ESL strategies that are helpful with our ESL students. The program includes technology that our children need to know how to navigate if they are to be competitive in our society." [*Treatment teacher log]

Student-Related Perceptions

In each log, teachers provided observations of student engagement during reading instruction. Teachers classified students as illustrating *high engagement, average engagement,* or *low engagement* with program materials (Table 8). The majority of log entries reported students exhibited *high engagement* (74.9%), followed by *average engagement* (20.7%), and *low engagement* (7.3%).

Table 8. Percentage of student engagement across treatment teacher logs (n = 308–397)

Engagement level	Mean percentages across logs
High Engagement	75%
Average Engagement	21%
Low Engagement	7%

Note: For each log, the percentages added up to 100%; however, the above data represent data across multiple teachers and logs, and as a result, might not add to 100%.

A teacher reflected on students' engagement:

> *"They are excited for reading. They come in ready. Never seen such a group of kids ready to read and ready to learn. I've seen more engagement than I ever have before. They are excited to get through the week, because they are excited about what is coming next. They want to read so much more, especially in guided reading they beg for small groups every day."*
> [Treatment teacher interview, Spring]

Comparison Teacher Implementation

Teachers in comparison classrooms continued to use their typical reading and writing programs and materials and reported on their use of these materials in a one-time survey

during the spring of 2013. Each of the 14 comparison teachers completed the spring survey resulting in a 100% response rate.

Program Implementation

On average, comparison teachers provided reading instruction either four days (14%) or five days (86%) per week. On average, comparison teachers provided writing instruction three days (21%), four days (36%), or five days (43%) per week. Teachers also reported on their average daily reading and writing instruction time. Half of the comparison teachers reported spending *more than 90 minutes* on reading instruction, and the majority (85%) spent *less than 90 minutes* on writing instruction (Figure 20 and Figure 21).

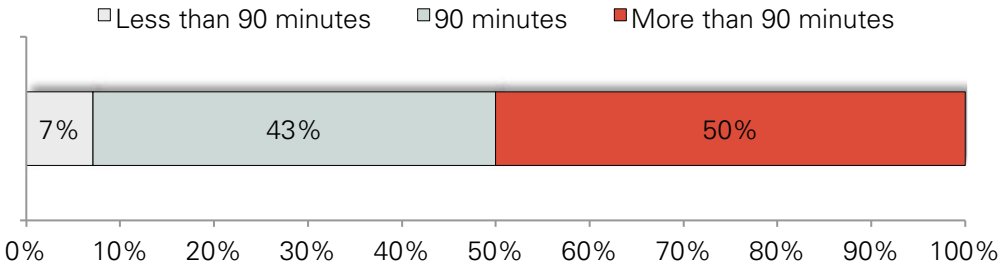

Figure 20. Comparison teacher's amount of time spent daily on reading instruction (n =14).

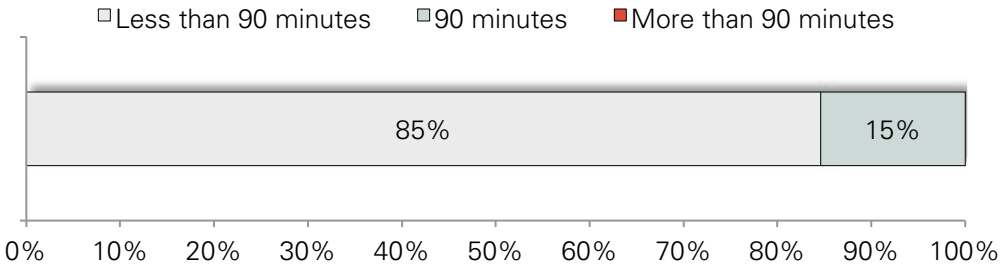

Figure 21. Comparison teacher's amount of time spent daily on writing instruction (n =14).

On average, comparison teachers most often used the student anthology (3.93) and the teacher edition (3.43) during the week. Comparison teachers least often used practice books (2.14) and worksheets (2.43) during the week (Figure 22).

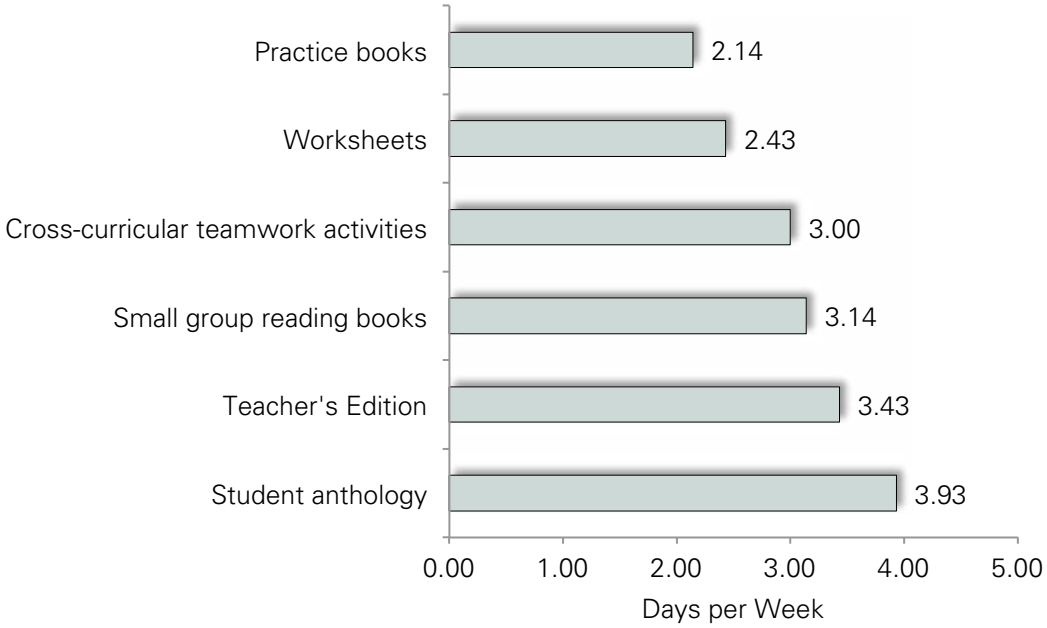

Figure 22. Days per week comparison teachers reported using instructional materials (n = 14).

Digital Instruction

Over half of the comparison teachers (57%) indicated that they typically integrate digital reading materials into instruction. Table 9 shows that among a variety of digital resources, teachers most often use background videos (36%) and interactive whiteboard lessons (29%).

Table 9. Percentage of comparison teacher digital resource use (n = 14)

Digital Resource	Percent of Teachers Indicating Use
Teacher's eEdition	21%
Student eEdition	21%
Digital library	7%
Build background video	36%
Vocabulary games	14%
Online lesson planner	7%
Interactive whiteboard lessons	29%
Other	14%

Of the 14% of comparison teachers that indicated using "other" digital resources, those resources included academic websites, flipboards, and power point presentations.

Small group instruction

All comparison teachers indicated that they typically provided small group reading instruction to students. Teachers most often met with the groups who are two grade levels below (4.27 days per week) and least often met with the advanced group of students (1.85 days per week). (Table 10).

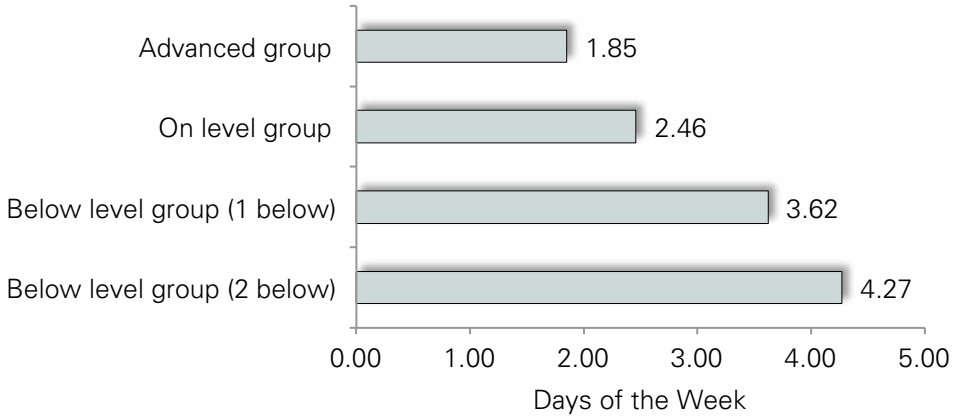

Table 10. Average days per week comparison teachers met with small groups (n = 11–13).

Teachers reflected on the adequacy of their reading materials to meet the needs of each of the small groups of students. Overall, 66% of teachers indicated that the materials *adequately* or *very adequately* addressed the reading needs of above level students, 58% indicated the materials *adequately* or *very adequately* addressed the reading needs of on level students, 21% indicated the materials *adequately* or *very adequately* addressed the reading needs of students one grade level below, and 15% indicated the materials *adequately* or *very adequately* addressed the reading needs of students two grade levels below (Figure 23).

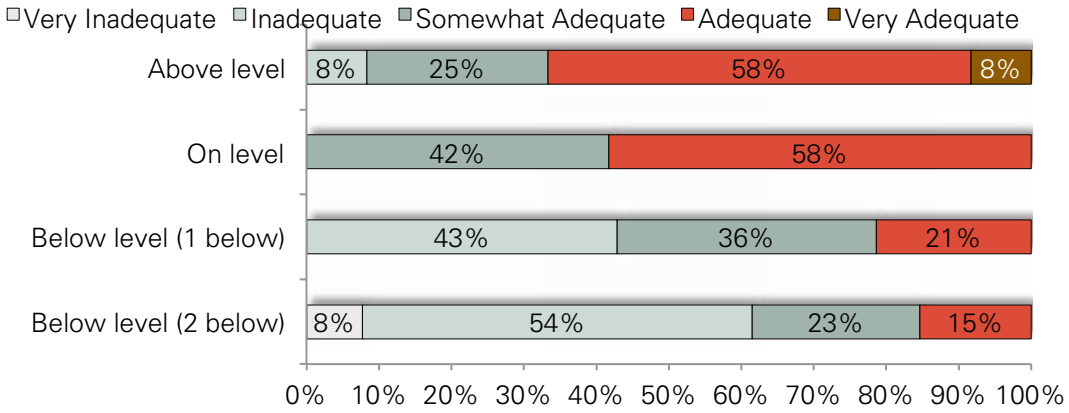

Figure 23. Comparison teachers' perceptions of adequacy of reading materials in meeting the needs of small groups (n = 14).

As shown in Figure 24, 44% of teachers rated their writing materials as *adequately* or *very adequately* addressing the needs of student writing above level, 40% gave these ratings

for students writing on level, 8% for students writing one grade below level, and 9% for students writing two grades below level.

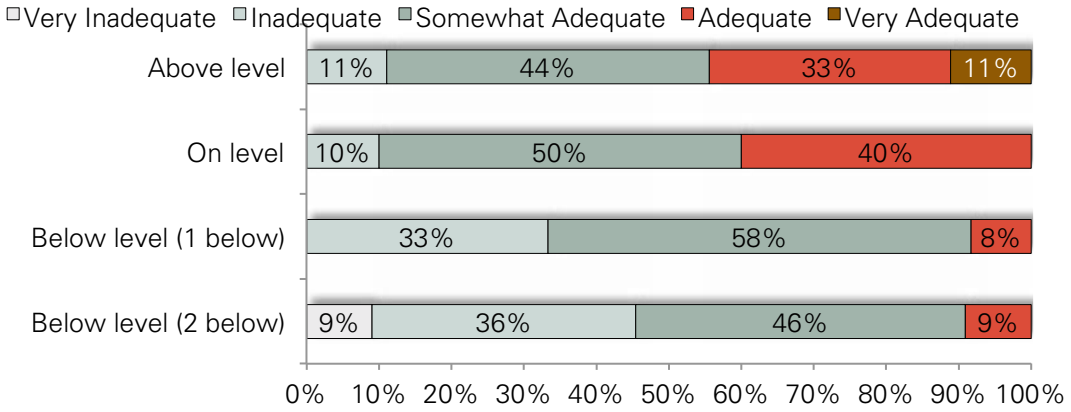

Figure 24. Comparison teachers' perceptions of adequacy of writing materials in meeting the needs of small groups (n = 13-14).

Assessment and Reteaching

All comparison teachers indicated that they used assessment tools as a part of their reading and writing instruction. The majority of teachers' ratings indicated that their assessments were *somewhat effective* to *effective* in helping teachers identify learner differences (57%), use data to guide instruction (72%), assess how well students organize and learn new content (71%), assess how well students meet the Common Core State Standards (58%), and measure students' application of new strategies (86%). (Figure 25).

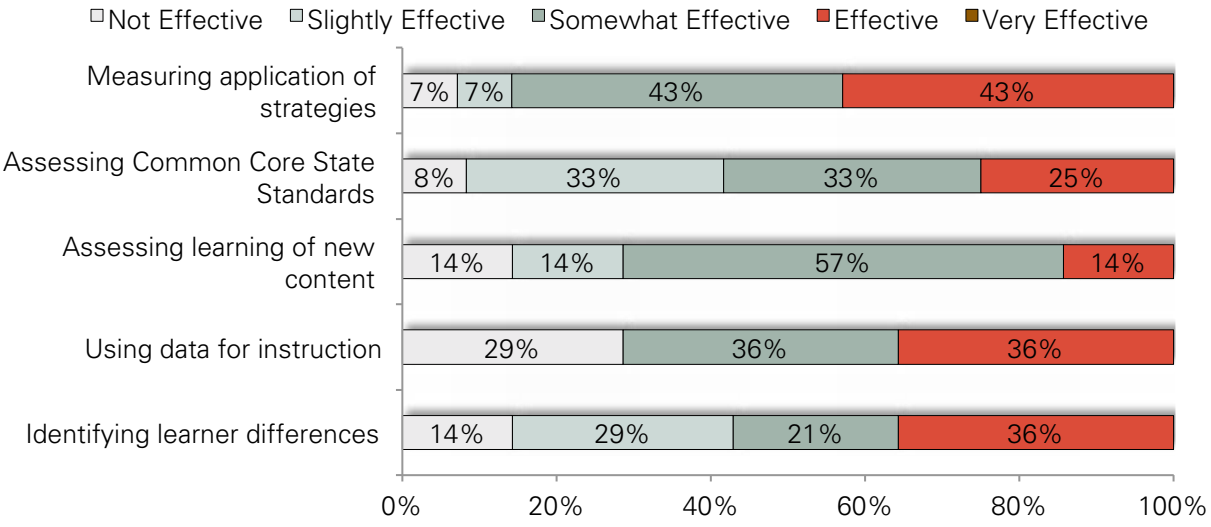

Figure 25. Comparison teachers' perceptions of the effectiveness of assessment and reteaching tools on achieving goals (n = 12–14).

Student Learning

Overall, comparison teachers indicated that their instructional materials were *somewhat effective* or *effective* in improving student learning in several academic areas and subareas. There were few instances where teachers indicated that their instructional materials were *very effective* in improving student learning. In all areas and subareas, teachers responded that their instructional materials were 21% - 50% *effective* or *very effective* (Figures 26 – 29).

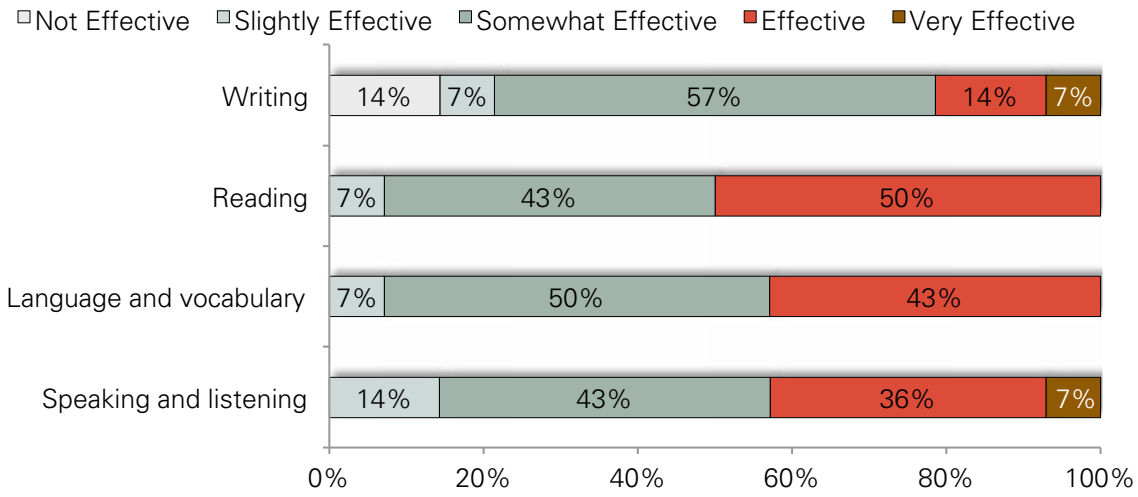

Figure 26. Comparison teachers' perceptions of the effectiveness of instructional materials in improving student learning (n = 14).

Speaking and Listening

Fifty percent of comparison teachers rated their instructional materials as *effective* or *very effective* in improving student academic talk and 36% of teachers rated their materials as *effective* or *very effective* in improving students' active listening skills.

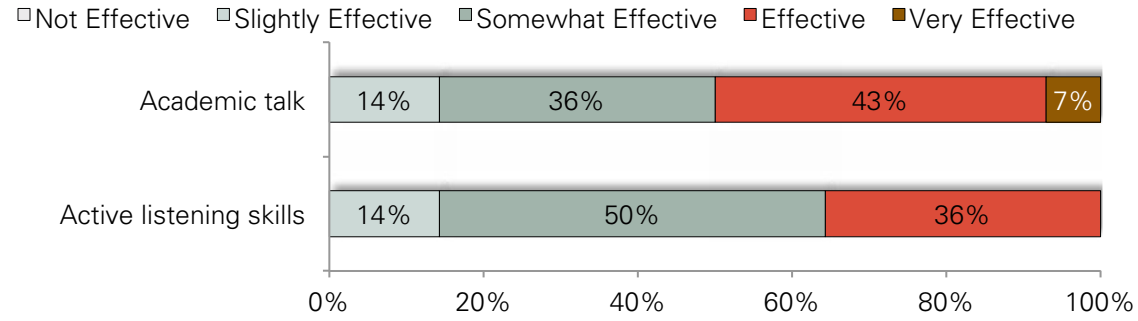

Figure 27. Comparison teachers' perceptions of the effectiveness of instructional materials in improving student learning in speaking and listening areas (n = 14).

Language and Vocabulary

Overall, 21% of teachers rated their instructional materials as *effective* in improving student content vocabulary, 43% rated their materials as *effective* in improving student academic vocabulary, 43% rated their materials as *effective* or *very effective* in improving students' grammar skills and 43% rated their materials as *effective* in improving student spelling and word work.

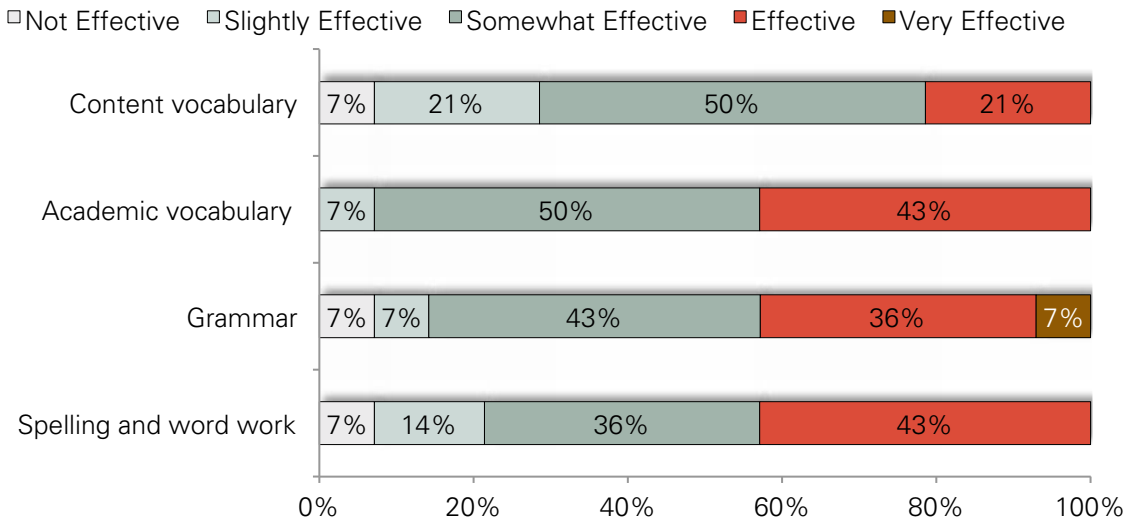

Figure 28. Comparison teachers' perceptions of the effectiveness of instructional materials in improving student learning in vocabulary areas (n = 14).

Reading Fluency and Comprehension

When asked about the effectiveness of their instructional materials in improving student fluency and comprehension, 50% rated their materials as *effective* in improving fluency and 64% rated their materials as *effective* in improving comprehension.

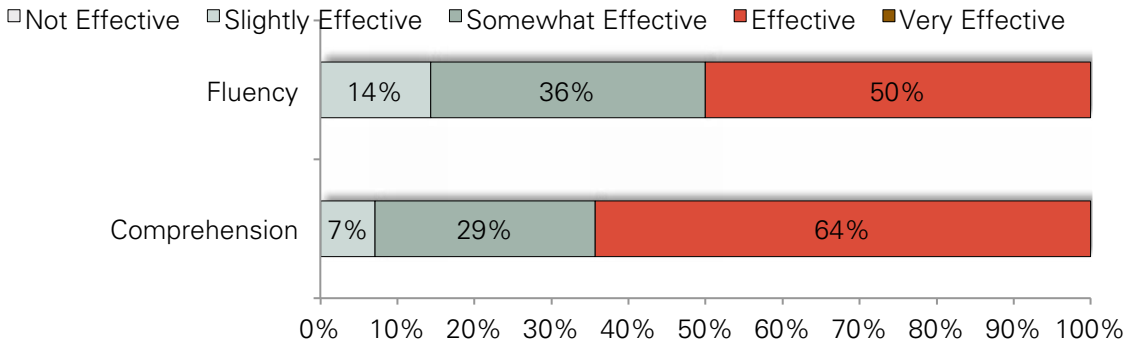

Figure 29. Comparison teachers' perceptions of the effectiveness of instructional materials in improving student learning in fluency and comprehension (n = 14).

Program Perceptions

Teacher-Related Perceptions

Comparison teachers were asked about the pacing of their reading and writing programs. Some teachers (36%) responded that the pace of their reading program allowed them adequate time to address the needs of all students. Similarly some teachers (29%) responded that the pace of their writing program allowed them adequate time to address the needs of all students.

Common Core

The teacher log included several questions to address how comparison teachers included the Common Core State Standards in their instruction. In the logs, 38% of teachers indicated that their reading and language materials explicitly addressed the Common Core State Standards and 36% of teachers indicated that their writing materials explicitly addressed the Common Core State Standards. In general, 72% of teachers indicated that they addressed the Common Core State Standards in their literacy instruction during that year. Some ways that comparison teachers addressed the Common Core State Standards included:

1. Bringing in supplemental materials *(n = 2)*
2. Focusing on the development and assessment of reading skills rather than reading comprehension *(n = 1)*
3. Making students reason and answer reading questions *(n = 1)*
4. Conducting research on the Internet *(n = 1)*
5. Collaborating with colleagues *(n = 1)*
6. Creating more student interaction *(n = 1)*

One teacher commented on the Common Core State Standards:

> *"The materials used this year are not aligned to CCSS. This created a challenge for teachers using it with the demands to implement CCSS in our daily instruction."* [Comparison teacher log]

Student-Related Perceptions

Teachers provided observations of student engagement during reading instruction. Teachers classified students as illustrating high engagement, average engagement, or low engagement with program materials (Table 11). The majority of teachers reported students exhibited average engagement (53%), followed by high engagement (34%), and low engagement (20%).

Table 11. Comparison teacher perspectives of student engagement in reading and writing (n = 11–13)

Engagement level	Mean percentages across logs
High Engagement	34%
Average Engagement	53%
Low Engagement	20%

Note: For each log, the percentages added up to 100%; however, the above data represent data across multiple teachers and logs, and as a result, might not add to 100%.

Reach for Reading Program Comparisons

The weekly logs reveal some similarities and differences between Reach for Reading and comparison teachers' perceptions of instructional materials and practices.

Reading and Writing Instructional Practices and Perceptions

Treatment and comparison teachers had different views with regard to student engagement in reading and writing. Treatment teachers indicated 74.9% of students were highly engaged, while comparison students indicated that 34% of students were highly engaged. Treatment teachers also utilized more digital instruction than comparison teachers. Almost all treatment teachers indicated utilizing digital instruction, while slightly more than half of comparison teachers utilized digital instruction. For treatment teachers, the top three most-used digital resources were the teacher eEdition (42%), student eEdition (36%), and the vocabulary games (36%). For comparison teachers, the top three most used digital resources were the teacher eEdition (21%), student eEdition (36%), and videos to build background (36%).

When asked about their perception of the effectiveness of their respective reading programs in improving student learning, treatment and comparison teachers' responses differed across various categories (Table 12). Overall, treatment teachers rated Reach for Reading components as being more effective than comparison teachers rated their reading instruction materials. Specifically, the greatest differences existed between teachers' ratings of program effectiveness in the areas of writing components (45 percentage-point difference) and language and vocabulary components (31 percentage-point difference).

Table 12. Percentage of responses indicating effectiveness of program components in improving student learning

Program instructional components	Percentage of responses indicating program was effective and very effective in improving student learning	
	Reach for Reading	Comparison
Speaking and listening	73%	43%
Language and vocabulary	74%	43%
Reading	78%	50%
Writing	66%	21%

Also according to the teacher logs, more treatment teachers reported that Reach for Reading assessments were effective in supporting instructional practices than did comparison teachers with regard to their programs' assessment tools. The greatest differences in teacher reports of their programs' assessment effectiveness were in areas of assessing student learning of new content and strategies (54 percentage-point difference), using data for instruction (46 percentage-point difference), and measuring application of strategies (36 percentage-point difference). (Table 13).

Table 13. Percentage of responses indicating effectiveness of program assessments in supporting instruction

Instructional supports	Percentage of responses indicating assessments were effective and very effective in supporting instructional practices	
	Reach for Reading	Comparison
Identifying learner differences	71%	43%
Using data for instruction	71%	25%
Assessing learning of new content	68%	14%
Assessing Common Core State Standards	70%	36%
Measuring application of strategies	72%	36%

Small group instruction

Treatment and comparison teachers both reported utilizing small group meetings in their lessons. Both groups of teachers reported similar average numbers of days per week spent with each of the small groups (Table 14). One notable exception was the amount of days spent with below-level students. Treatment teachers reported spending an average of 2.6 days per week with students reading one grade level below level, while comparison teachers reported spending an average of 3.6 days per week with this group of students. Similarly, treatment teachers spent an average of 2.6 days per week with students reading two grades below level, while comparison teachers spent an average of 4.3 days per week with this group of students.

Table 14. Percentage of responses indicating days per week spent on small group instruction

	Average number of days per week spent on small group instruction	
Small Group	Reach for Reading	Comparison
Above level	2.3 days	1.9 days
On level	2.6 days	2.5 days
Below level (1 grade below)	2.6 days	3.6 days
Below level (2 grades below)	2.6 days	4.3 days

Across all logs, treatment teachers reported that the Reach for Reading program was more adequate in meeting the needs of small groups. In fact, although comparison teachers spent more time with the below-level small groups, the teacher logs indicate that the greatest program differences in meeting the needs of small groups were evident with below-level students and in favor of the Reach for Reading program. Specifically, there was a 46 percentage-point difference in teacher reports of their respective program's adequacy meeting the needs of students reading one grade below level and a 43 percentage-point difference for students reading two grades below level (Table 15).

Table 15. Percentage of log responses indicating adequacy of small group instruction

	Percentage of responses indicating programs were adequate or very adequate in meeting the needs of small groups	
Small group	Reach for Reading	Comparison
Above level	92%	66%
On level	81%	58%
Below level (1 grade below)	67%	21%
Below level (2 grades below)	58%	15%

Student Performance Results

Evaluators measured the impact of Reach for Reading on student performance using the GMRT-4 and the DIBELS Oral Reading Fluency assessments. Treatment and comparison teachers administered each at the beginning and end of the study period. In addition, treatment students completed Reach for Reading Common Core benchmark assessments as part of program implementation and as measures of student learning relative to the Common Core State Standards. The following sections present findings from these measures. In order for evaluators to use all available data and maximize the study's power, evaluators used multiple imputation procedures to impute missing data. Multiple imputation procedures yielded five complete datasets, and estimates for imputed datasets were pooled using SPSS and HLM 7.0, as appropriate. The results in this report reflect the findings from the pooled estimates.

Treatment Student Performance on Common Core Benchmark Assessment

This section presents the results of treatment student performance on the Reach for Reading Common Core benchmark assessments. Test results are reported as the average percent correct for each strand of standards. Each strand encompasses multiple standards, and 1–3 items map onto each standard. The following list presents the number of standards per strand:

- Reading Literature: 9 standards
- Reading Informational Text: 10 standards
- Language for Reading: 6 standards
- Language for Writing: 12 standards
- Writing: 2 standards

It is important to note that treatment teachers had not provided instruction related to the CCSS prior to the study period, nor had students been assessed on the standards. Teachers implemented the assessment as part of the intervention being evaluated by this study.

Caution is warranted in interpreting these results because (a) the benchmark assessment lacks available validity and reliability data, (b) tests were scored by different independent subcontractors at pretest and posttest, who each were trained but not assessed for inter-rater agreement on the writing test, and (c) the analyses were of a descriptive nature and did not take into account the nested data structure or any variance that could influence scores based on student, teacher or school characteristics.

KEY QUESTION:

Did treatment students in Reach for Reading classrooms demonstrate significant learning gains in reading achievement scores after one year of implementation?

As shown in Figure 30, students' average scores on the benchmark Reading test increased from the beginning to the end of the year. Students showed the greatest gains, on average, on items related to the Reading Literature standards. Results for the Reading Informational Text standards suggest that this assessment was the most difficult for students with an average percent correct of 52% at the end of the year. On the Total Score for the benchmark Reading test, students' average performance increased from 54% correct to 65% correct. This reflects a mean difference of 6.58 points and a large effect size of 1.51. Although this difference is statistically significant (see Table 16), student performance on this assessment was still low by the end of the year. Teacher implementation feedback confirms that students found the assessments to be difficult.

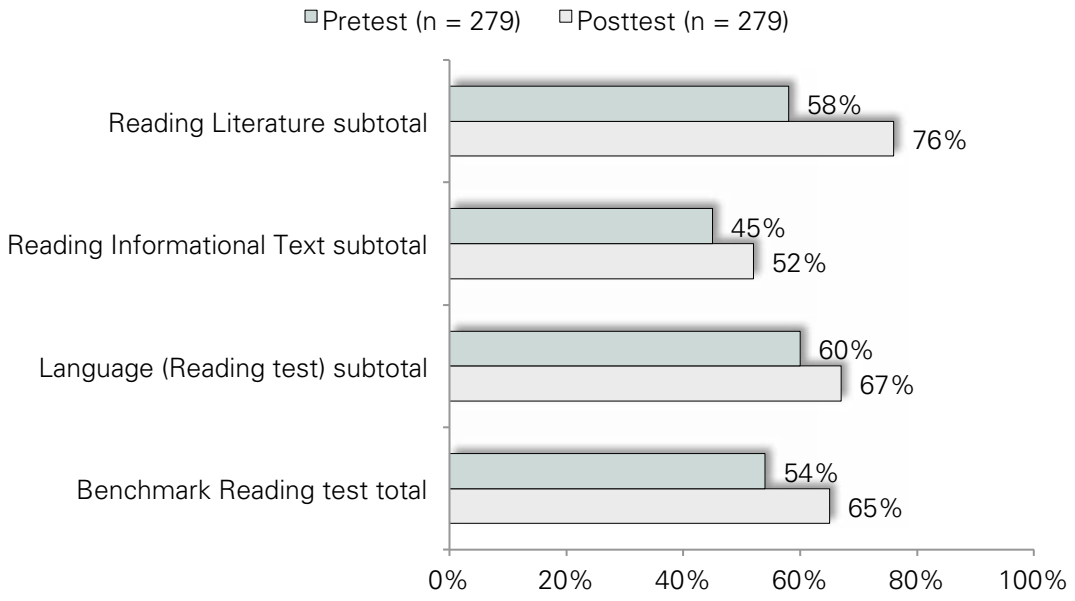

Figure 30. Common Core benchmark Reading test average pretest and posttest scores.

The Reach for Reading program offers a mid-year administration of the benchmark assessments, which was optional for the study. Of the treatment students, 107 students completed the benchmark Reading test with an average percent correct of 70% for the Reading Literature subtotal, 57% correct for the Reading Informational Text subtotal, 55% correct for the Language subtotal, and 62% correct for the benchmark Reading test total.

Figure 31 presents the results of the benchmark Writing test, which revealed an increase of average treatment student performance on the Language portion and a decrease on the Writing portion. The average percent correct decreased from 56% at the beginning of the year to 52% at the end of the year on the benchmark Writing test total. The mean difference in points from pretest to posttest was statistically significant and reflects an effect size of -0.47 (see Table 16).

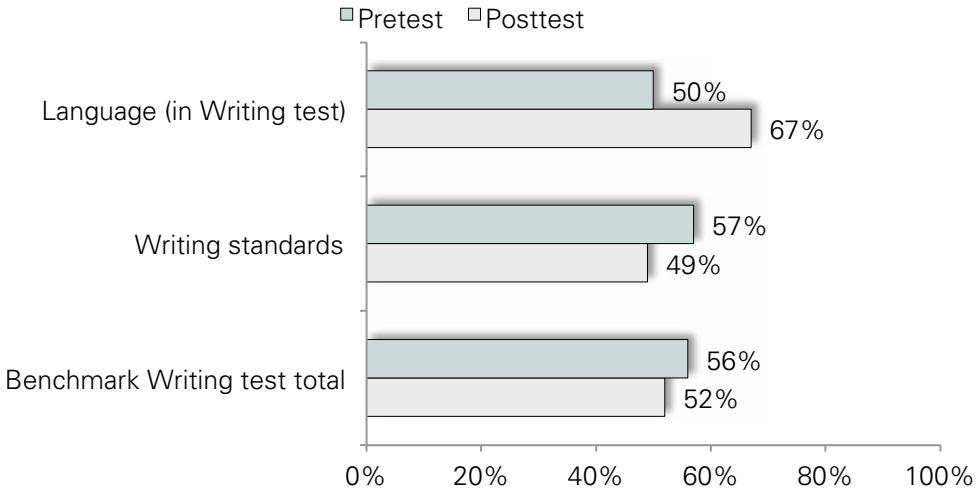

Figure 31. Common Core benchmark Writing test average pretest and posttest scores.

Table 16. Common Core benchmark Reading and Writing test outcomes

Outcome variable	Mean difference	Standard deviation	*t* value	Approx. df	*p* value	Effect size
Benchmark Reading Test Total	6.58 points	8.15	13.66	278	<0.001***	1.51
Benchmark Writing Test Total	-3.78 points	13.55	4.91	278	<0.001***	-0.47

*** Significant at the 0.001 level.

Comparisons of Student Learning Gains by Treatment and Comparison Group

Evaluators examined differences between Reach for Reading students and comparison students to determine if Reach for Reading improved student performance over and above what would be expected had the students not participated in Reach for Reading. Researchers conducted multilevel modeling to estimate the impact of Reach for Reading on students' reading achievement and to account for two sources of variance in student performance outcomes: (1) teacher-level and (2) student-level. This acknowledges that students' learning experiences within classrooms are not independent or unrelated to each other and thus, should not be analyzed as such. The notion is that two students in the same classroom with the same teacher and exposed to the same classroom-level influences are more likely to respond similarly than two students randomized from different classrooms (Borman et al., 2005).

The analytical models for the GMRT-4 and DIBELS ORF outcomes included a pretest covariate at Level 1 and study condition (treatment or comparison) at Level 2. To control for any variance associated with students being in different schools, the models included dummy-

coded school variables as Level 2 covariates. Evaluators calculated effect sizes to determine the magnitude of the difference between treatment and comparison students.[1]

The main analyses presented in this section consisted of tests for statistically significant findings. A statistically significant finding is one that is unlikely to have occurred by chance. In the case of the analyses presented in this report, individual findings were considered statistically significant if the probability of the finding occurring by chance was less than 5%. Conducting many individual tests (e.g., GMRT-4 Vocabulary and GMRT-4 Comprehension), as done for this report, increases the probability that some test results may be statistically significant by chance. For example, with a threshold of 5%, one would expect one test out of 20 to be statistically significant by chance alone. This means that some correction must be made when conducting many tests of statistical significance.

Correcting for the increased chances of statistically significant results is known as correcting for multiple comparisons (Schochet, 2008). There are many different methods for correcting for multiple comparisons. For this report, evaluators used the Benjimini-Hochberg (BH) correction method. Evidence suggests that the BH method may be the best solution to the multiple comparisons problem in many practical situations (Williams, Jones, & Tukey, 1999). The BH method was conducted for the GMRT-4 domains of Vocabulary, Comprehension, and Total score. All tests that were statistically significant prior to the BH correction were still statistically significant after the BH correction.

This section presents student performance comparisons on the DIBELS ORF and the GMRT-4 assessments. Evaluators also examined differences between subgroups of students including those qualifying for free- or reduced-price lunch (FRL) and students with limited English proficiency (LEP).

KEY QUESTION:

Did the Reach for Reading program significantly impact treatment students' reading achievement compared to comparison students' achievement after one year of implementation?

Program Impacts on Oral Reading Fluency

Teachers administered the DIBELS Next ORF test at the beginning and end of the study with each administration yielding a score for total words correct per minute (wcpm). The test established different benchmarks for fall and spring administration periods. As part of progress monitoring, benchmarks included bands for students performing at or above benchmark, below benchmark, and well below benchmark. The fall benchmarks were lower than the spring benchmarks to account for expected student growth during the school year. For example, for a student to score at or above benchmark, a student needed to read 70 or more words correct per minute in the fall and 100 or more words correct per minute in spring. Figures 32 and 33

[1] An effect size is a unit of measurement that expresses the difference in outcome for the average treatment participant from the average comparison student. It also is used to indicate the strength of the increase or decrease in achievement

present the percentages of treatment and comparison students scoring within each benchmark band for fall 2012 and spring 2013. At both time points the majority of treatment and comparison students scored at or above the benchmarks. Student distributions across these benchmark levels remained relatively stable during the study period. By the end of the year, 69% of treatment students and 60% of comparison students scored at or above the benchmark of 100 wcpm. For the Reach for Reading classrooms, there was a slight upward shift in the percentage of students progressing from well below benchmark to below benchmark. For comparison classrooms, there was a downward shift in the percentage of students scoring at or above benchmark in the fall to scoring below benchmark in the spring.

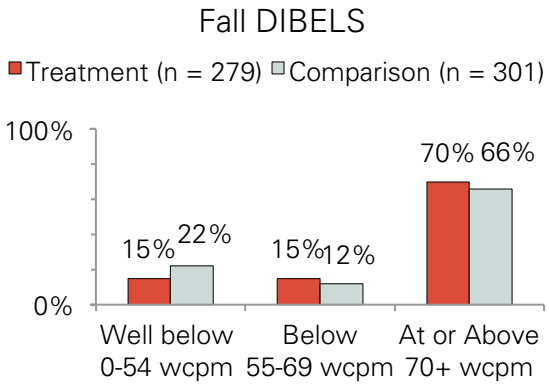

Figure 33. Percentages of treatment and comparison students meeting DIBELS ORF benchmarks in fall 2012.

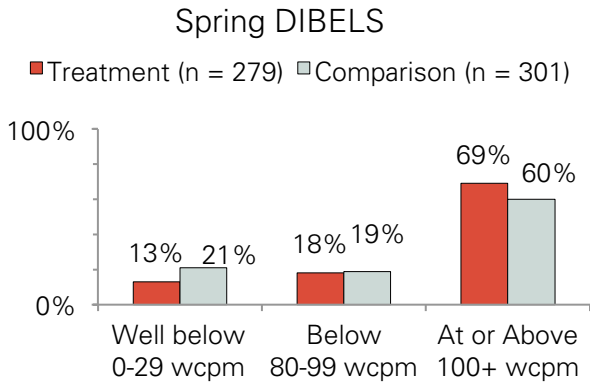

Figure 32. Percentages of treatment and comparison students meeting DIBELS ORF benchmarks in spring 2013.

To assess if statistically significant differences existed between treatment and comparison students' oral reading fluency by the end of the school year, evaluators used a effects multilevel model with the spring ORF score as the outcome variable. The model included covariates to control for students' fall ORF performance and the school students attended. The difference between treatment and comparison students' performance on the ORF was not statistically significant. The positive coeffient for ORF (i.e., 0.34) indicates that treatment students had a higher odds of scoring at or above benchmark in the spring than comparison students (see Table 17). More specifically, the odds of scoring at or above benchmark in spring were 1.41 times greater for Reach for Reading students than for comparison students.

Evaluators conducted the same analyses to examine subgroup differences between treatment and comparison students qualifying for free- or reduced-price lunch (FRL) and students with limited English proficiency (LEP) status. There were no statistically significant differences within each subgroup. However, the odds of scoring at or above benchmark in spring were 1.42 times greater for FRL treatment students than for FRL comparison students, and 2.12 times greater for LEP treatment students than LEP comparison students.

Table 17. Estimation of effects for DIBELS Oral Reading Fluency outcomes

Outcome variable	Coefficient	Standard error	*t* value	Approx. df	*p* value	Odds ration effect size[a]
All Students ORF	0.34	0.30	1.13	20	0.27	1.41
FRL - ORF	0.35	0.30	1.17	18	0.26	1.42
LEP - ORF	0.75	0.46	1.63	16	0.12	2.12

* Significant at the .05 level.

[a] The effect size is expressed as an odds ratio, which is appropriate for characterizing the magnitude of a program effect for binary outcomes (e.g., performing/not performing at or above benchmark). The odds ratio represents the impact of Reach for Reading in terms of how much greater (or smaller) the odds of a positive fluency outcome are for a treatment student than for a comparison student. Evaluators used the exponent function to calculate odds ratios.

Program Impacts on Vocabulary and Comprehension

Evaluators based students' reading achievement on GMRT-4 extended scale scores[2] (ESS) for the domains of Comprehension, Vocabulary, and Total Reading. Figures 34-36 provide descriptive presentations of the average pre-test and posttest scores for Reach for Reading and comparison students after accounting for student- and teacher-level variance. In all three domains, treatment students' average scores were higher than comparison students' average scores at pretest and posttest. These figures also show the effect size for each comparison, which will be discussed next.

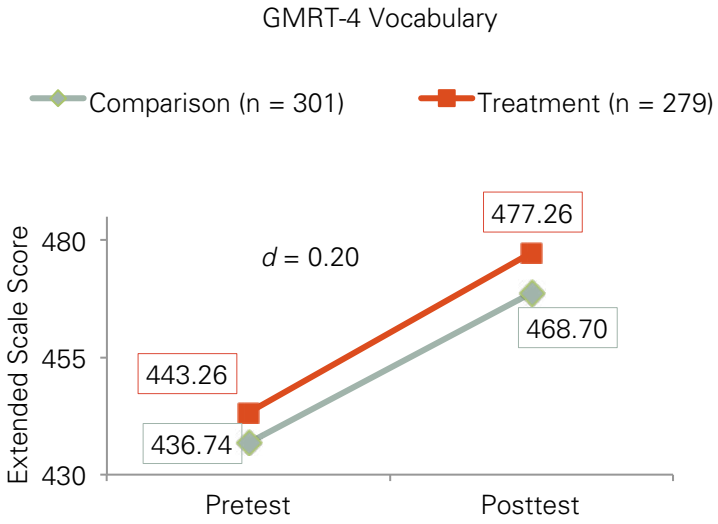

Figure 34. Pretest and posttest adjusted Vocabulary means by condition.

[2] An extended scale score is an equal-unit scale from the lowest achievement in kindergarten to the highest achievement in Grade 12. It is used to track student progress over a period of time and is used primarily for statistical analyses.

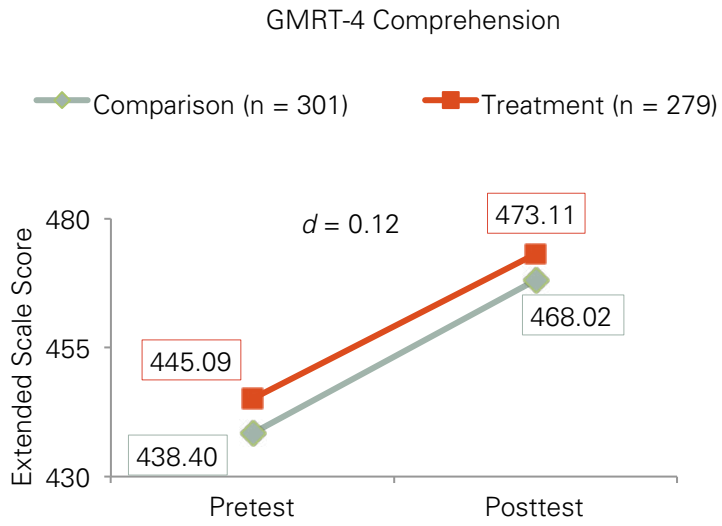

Figure 35. Pretest and posttest adjusted Comprehension means by condition.

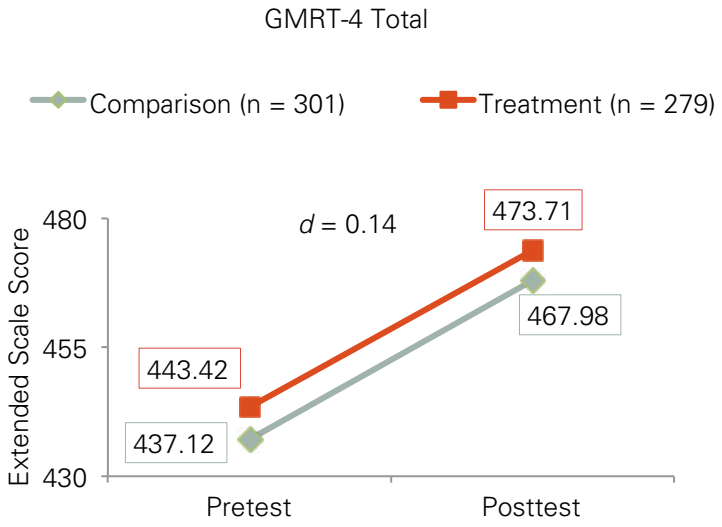

Figure 36. Pretest and posttest adjusted Total reading scores by condition.

As indicated previously, treatment and comparison students' pretest scores did not differ significantly, indicating that they were statistically equivalent at baseline in reading. At posttest, there was a statistically significant difference between treatment and comparison students for Vocabulary and Total Reading (see Table 18). To interpret the practical significance of these findings, the magnitude of these differences is reflected in an effect size of 0.20 for Vocabulary. This translates into an improvement index of eight percentile points, which means that had comparison students participated in Reach for Reading, the average student would

have achieved an 8% increase in percentile rank in Vocabulary. This also indicates that 58% of Reach for Reading students scored above the comparison group mean in Vocabulary. The effect size for Total Reading is 0.14 and the improvement index is six percentile points, which indicates that 56% of treatment students scored above the comparison group mean in Total Reading.

The difference between treatment and comparison students' performance on the Comprehension test was not statistically significant and reflects an effect size of 0.12. This difference translates to a difference of five percentile points between the average treatment student and the average comparison student.

Table 18. Program impacts on Reading Vocabulary and Comprehension

Outcome variable	Coefficient	Standard error	*t* value	Approx. df	*p* value	Effect size	WWC improvement index
Vocabulary	8.56	2.88	2.97	20	0.008**^	0.20	8 percentile points
Comprehension	5.09	3.00	1.70	20	0.10	0.12	5 percentile points
Total Reading	5.73	2.55	2.25	20	0.03*^	0.14	6 percentile points

* Significant at the .05 level.
** Significant at the 0.01 level
^ Significant after application of the Benjamini Hochberg correction for multiple comparisons.

Program Impacts on Learning Outcomes for Subgroups of Students

KEY QUESTION:

Were there differential effects between treatment and comparison student subgroups?

To explore if Reach for Reading had a significant impact on students who qualify for free- and reduced-price lunch (FRL), evaluators analyzed posttest GMRT-4 scores for this subgroup of treatment (n = 171) and comparison (n = 222) students. As shown in Figures 37-39, the average performance of treatment students was higher than the average performance of comparison students in all three domains: Vocabulary, Comprehension, and Total Reading. These differences were only statistically significant for Vocabulary with an effect size of 0.28 (see Table 19). This translates to an improvement index of 11 percentile points for treatment students and indicates that 61% of FRL students in Reach for Reading scored above the mean of FRL students in comparison classrooms. Effect sizes for Comprehension and Total Reading were 0.15 and 0.20, respectively, and were not statistically significant.

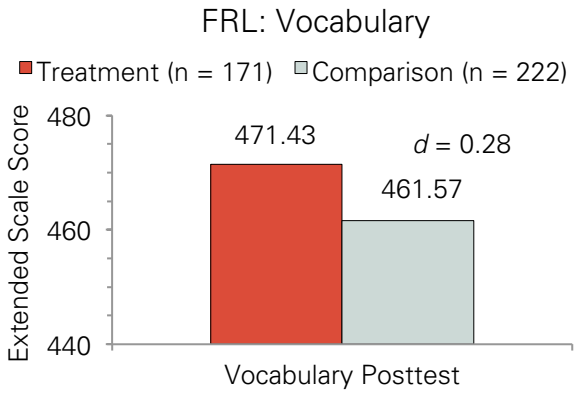

Figure 37. Posttest Vocabulary adjusted means for FRL students by condition.

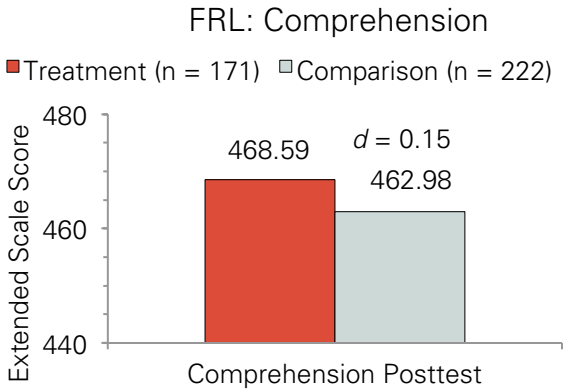

Figure 38. Posttest Comprehension adjusted means for FRL students by condition.

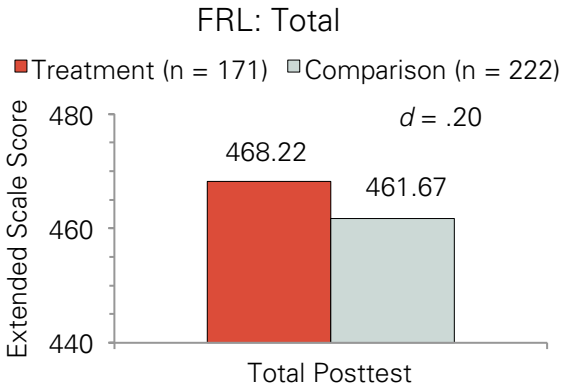

Figure 39. Posttest Total adjusted total scores for FRL students by condition.

Table 19. Program impacts for FRL student subgroup

Outcome variable	Coefficient	Standard error	*t* value	Approx. df	*p* value	Effect size	WWC improvement index
FRL: Vocabulary	10.43	3.38	3.11	18	0.007*^	0.28	11 percentile points
FRL: Comprehension	5.61	4.01	1.40	18	0.18	0.15	6 percentile points
FRL: Total	6.55	3.24	2.02	18	0.06	0.20	8 percentile points

* Significant at the .05 level.
** Significant at the 0.01 level
^ Significant after application of the Benjamini Hochberg correction for multiple comparisons.

Evaluators also explored whether differences existed between performances of treatment and comparison students categorized as limited English proficient (LEP), as measured by the GMRT-4. As shown in Figures 40-42, the average posttest performance for treatment students (*n* = 100) was higher than the average posttest performance of comparison students (*n* = 101) in Vocabulary, Comprehension, and Total Reading. These differences were statistically significant for Vocabulary and Total Reading with moderate effect sizes of 0.57 and 0.40, respectively (see Table 20). Based on the improvement index, this indicates that 72% of LEP treatment students scored above the mean of LEP comparison students in Vocabulary, and 66% scored above the comparison mean in Total Reading. Although the difference in Comprehension performance between LEP treatment and comparison students was not statistically significant, the effect size of 0.23 indicates that 59% of treatment students scored above the mean for comparison students.

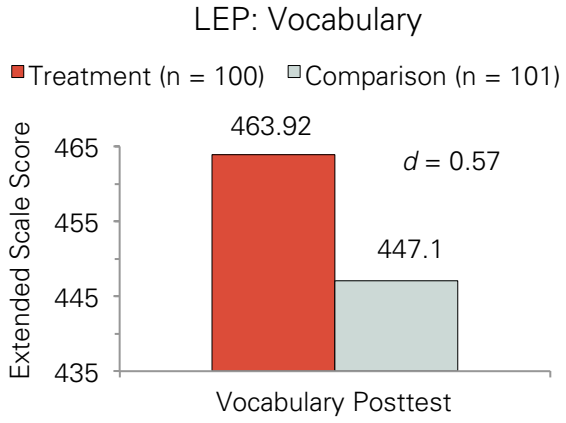

Figure 40. Posttest Vocabulary adjusted means for LEP students by condition.

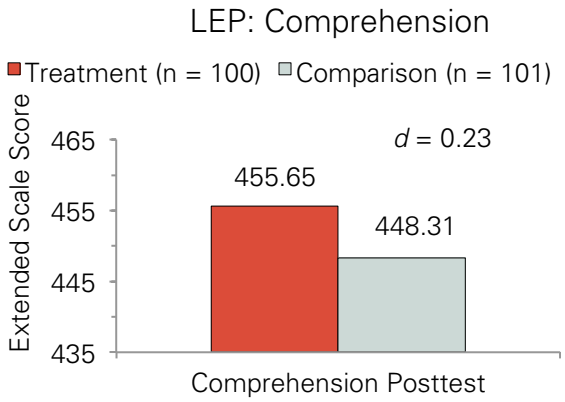

Figure 41. Posttest Comprehension adjusted means for LEP students by condition.

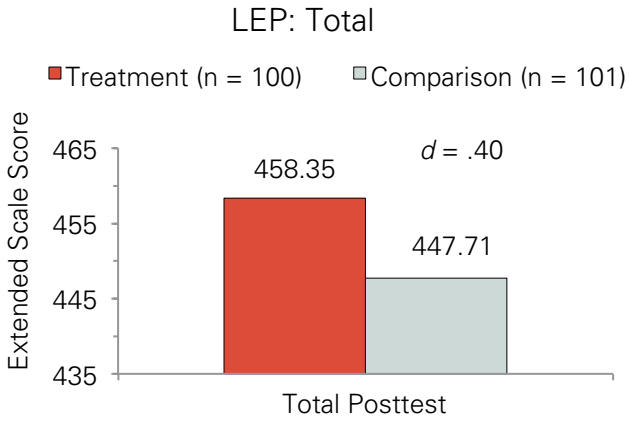

Figure 42. Posttest Total adjusted reading scores for LEP students by condition.

Table 20. Program impacts for LEP student subgroup

Outcome variable	Coefficient	Standard error	*t* value	Approx. df	*p* value	Effect size	WWC improvement index
LEP: Vocabulary	16.81	3.50	4.80	16	< 0.001***^	0.57	22 percentile points
LEP: Comprehension	7.34	3.69	1.99	16	0.06	0.23	9 percentile points
LEP: Total	10.64	2.76	3.85	16	0.002**^	0.40	16 percentile points

** Significant at the .01 level.
*** Significant at the .001 level.
^ Significant after application of the Benjamini Hochberg correction for multiple comparisons.

Limitations of the Study

The findings of this study are generalizable only to third-grade students in similar school settings. The lack of a valid and reliable norm-referenced assessment that aligns with the Common Core State Standards precluded evaluators from making causal claims about the impact of Reach for Reading on student learning related to the standards. Findings presented in this report are for descriptive purposes. Based on budget constraints, the measurement of writing outcomes was relegated to the Reach for Reading Common Core benchmark assessment. Because different scorers were used at pretest and posttest, it was not possible to establish inter-rater agreement on the writing portion of the assessment. Subsequently, writing scores could have been influenced by rater bias and should be interpreted with caution.

Summary and Discussion

The purpose of this study was to measure the efficacy of Reach for Reading in improving third-grade students' reading and writing skills. This study also included an examination of teachers' implementation of Reach for Reading and comparison teachers' curricula. The final analytical sample for the study consisted of 28 teachers and 580 students across seven schools and four school districts. Student outcomes were measured through multiple assessments, and teacher fidelity of implementation was measured through online logs and classroom observations.

Implementation. Reach for Reading teachers implemented the program on average 90 minutes or more per day for 3.8 days per week. This was consistent with the amount of time comparison teachers implemented reading and writing each week. Reach for Reading teachers demonstrated a high level of fidelity in implementing the required program components, with an overall fidelity rating of 94%. As part of their implementation, teachers differentiated instruction with small groups, engaged students in academic talk, and used vocabulary and reading teaching routines with every lesson or most lessons, on average.

Treatment teachers rated the effectiveness of Reach for Reading in improving student learning higher than comparison teachers rated their reading and writing programs and materials in several areas. The areas with the largest and most noteworthy differences in teachers' effectiveness ratings were writing (45 percentage-point difference), and language and vocabulary (31 percentage-point difference). On average, treatment teachers also gave Reach for Reading higher effectiveness ratings than comparison teachers gave their programs with regard to assessing student learning of new content and strategies (54 percentage-point difference), using assessment data for instruction (46 percentage point difference), and measuring students' application of strategies (36 percentage-point difference).

There were also notable differences between treatment and comparison teachers' small group instruction practices and perceptions. Despite comparison teachers reporting that they spent more time with the below-level small groups than did treatment teachers, only 21% of comparison teachers rated their program as *adequate* or *very adequate* for meeting the needs of their below-level students, compared to 67% of treatment teachers. Similarly, 15% of comparison teachers rated their programs as *adequate* or *very adequate* for meeting the needs of students reading significantly below grade level, compared to 58% of treatment teachers.

Common Core Benchmarks. Overall, teachers thought Reach for Reading supported them in addressing the Common Core State Standards (CCSS) to *a great extent*, and particularly in the areas of emphasizing academic language vocabulary and using more informational text. Treatment teachers reported during interviews that students' scores on the Common Core benchmark assessments were lower than expected. In general, they attributed this to the high bar set by the CCSS for writing and reading informational text skills. Based on teacher interviews, the majority of students had not experienced the level of writing rigor and stamina reflected in the standards prior to Reach for Reading. Average treatment student scores on all portions of the benchmark assessments were 60% correct at pretest and 76% correct at posttest. Student gains from pretest to posttest on the benchmark Reading test were statistically significant, $t(278) = 6.58$, $p < .001$. A decrease in student scores on the benchmark Writing test also was statistically significant, $t(278) = -3.78$, $p < .001$. Caution is warranted when interpreting these results, given the nature of the assessment, its scoring, and use of an analytical significance test that did not account for student, teacher, or school variance.

Oral Reading Fluency. There were no statistically significant differences in the number of treatment and comparison students demonstrating an oral reading fluency level at or above benchmark at the end of the study. The majority of students in each group met or exceeded the spring benchmark of reading 100 or more words correct per minute with 69% of treatment students and 60% of comparison students demonstrating fluency at this level. The odds of scoring at or above benchmark in spring were 1.41 times greater for Reach for Reading students than for comparison students. There were no statistically significant differences for subgroups of students qualifying for free- or reduced-price lunch or LEP students.

Reading Vocabulary and Comprehension. There was a statistically significant difference in treatment and comparison students' scores on the GMRT-4 Vocabulary and Total Reading tests. The difference in scores on the Comprehension test was not statistically significant. Effect sizes were 0.20 for Vocabulary, 0.12 for Comprehension, and 0.14 for Total Reading, which translates to the average treatment student scoring eight, five, and six percentile points higher than the average comparison student, respectively.

Reach for Reading also resulted in positive impacts for FRL and LEP students. The program had a statistically significant and positive impact on FRL students' performance on the GMRT-4 Vocabulary test with an effect size of 0.28 and a percentile difference of 11 points between the average treatment student and the average comparison student. Although not statistically significant, the effect size for FRL students on the Comprehension test was 0.15 and 0.20 on the Total Reading test, which are considered substantively important positive effects by the U.S. Department of Education's What Works Clearinghouse (What Works Clearinghouse, 2008). For the LEP student subgroup, there were statistically significant

differences in student performance on the GMRT-4 Vocabulary and Total Reading tests. Effect sizes for both of these tests were moderate with 0.57 for Vocabulary and 0.40 for Total Reading. This translates to the average treatment student scoring 22 percentile points higher than the average comparison student on the Vocabulary test and 16 percentile points on the Total Reading test. The effect size of 0.23 on the Comprehension test is considered substantively important, although not statistically significant, and translates to a nine percentile point difference between the average treatment and comparison student.

Overall. Through a rigorous, well implemented RCT, this study found that Reach for Reading has a statistically significant positive effect on student reading outcomes. This positive effect also is evident on reading outcomes for students with limited English proficiency. The program also positively impacts vocabulary outcomes for low-income students. Treatment students' oral reading fluency gains were comparable to those of comparison students.

References

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educational Evaluation and Policy Analysis, *29*(1), 30–59.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N. A., & Chambers, B. (2005). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42(4), 673*–696.

Erickson, F. (1986). *Qualitative methods in research on teaching*. In M.C. Wittrock (Ed.), Handbook of research on teaching (3rd ed.) (pp. 119-159). New York: Macmillan Publishing Co.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87.

Liebling, C. & Meltzer J. (2011). Making a difference in student achievement using the Common Core State Standards for English language arts: What school and district leaders need to know. Portsmouth, NH: PGA Education.

McTighe, J. & Wiggins, G. (2012). From Common Core standards to curriculum: Five big ideas. Retrieved from http://educore.ascd.org/Resource/Content/d92b6a9f-c938-4fe6-94d1-6f6bc174fa89

National Governors Association, Council of Chief State School Officers, & Achieve. (2008). Benchmarking for success: Ensuring U.S. students receive a world-class education. Retrieved from http://www.achieve.org/files/BenchmarkingforSuccess.pdf

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2005). Optimal design for longitudinal and multilevel research, Version 1.55 [computer software]. Retrieved from http://sitemaker.umich.edu/group-based/optimal_design_software.

Schochet, P. Z. (2008). Technical methods report: Guidelines for multiple testing in impact evaluations (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2008). What Works Clearinghouse procedures and standards handbook (Version 2.0). Retrieved from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*(1), 42–69.

Appendix A CONSORT Flow Diagram