A Final Report for the Evaluation of the Achieve3000 Programs

October 13, 2015

magnolia consulting

cultivating learning and positive change
www.magnoliaconsulting.org

Findings at a Glance

- Overall, the learning gains of treatment students who used Achieve3000 were statistically significant and substantively important—based on the What Works Clearinghouse (WWC) threshold of 0.25—for all areas assessed: the GMRT-4 Vocabulary, Reading Comprehension, and Total Reading tests, as well as the LevelSet Lexile reading assessment.
- Across grades, Achieve3000 had a statistically significant positive impact on GMRT-4 Reading Comprehension and Total Reading when compared to study schools' standard English Language Arts (ELA) curricula. However, the effect sizes did not meet the WWC threshold of 0.25 for being substantively important.
- Within-grades, Achieve3000 had positive impacts on sixth-grade reading that were not statistically significant but approached the WWC threshold of 0.25 for being substantively important. Furthermore, the program had substantively important positive impacts on ninth-grade GMRT-4 Vocabulary, Reading Comprehension, and Total Reading.
- Teachers generally reported that Achieve3000 was effective, and most teachers noted that they would use the program again next year.

Executive Summary

Achieve3000, the publisher of differentiated online literacy programs, understands the importance of demonstrating the efficacy of its products through independent evaluation. Therefore, it contracted with Magnolia Consulting, LLC, an independent evaluation consulting firm to conduct this randomized control trial study of its Achieve3000 programs. Magnolia Consulting conducted this study with third-, sixth-, and ninth-grade teachers and students during the 2014/15 school year.

This study sample came from four districts located in three different regions of the United States: the West South region, the East North Central region and the Pacific region. Two districts were classified as "Suburb: Large" and two districts were classified as "City: Large" (U.S. Department of Education, National Center for Education Statistics, 2015). The number of schools in each district ranged from 21 to 23, and the student population in each district ranged from 19,257 to 130,271. The student-teacher ratios also varied and ranged from 15.57 to 25.14. Two districts had a higher percentage of Hispanics or Latinos, and race varied across the four districts. The Percentage of English Language Learner (ELL) students ranged from 2.49% to 11.52% and percentage of students with Individualized Education Programs (IEPs) ranged from 10.87% to 14.38%.

Study Design & Methods

Magnolia Consulting evaluators used a randomized control trial to conduct this mixed-methods evaluation study. Within each grade, evaluators randomly assigned half of participating teachers to the treatment condition and half to the comparison condition. Treatment teachers implemented the Achieve3000 program with their students, and comparison teachers implemented their usual ELA materials but not the Achieve3000 program. This design enabled evaluators to estimate an unbiased impact of Achieve3000 on student learning in reading.

This efficacy study used multiple student and teacher measures. Student measures included the Gates MacGinite Reading Test, fourth edition (GMRT-4) and Achieve3000's LevelSet. The GMRT-4 is a group-

administered, norm-referenced assessment that yields scores for Vocabulary, Reading Comprehension, and Total Reading. Teachers in both study conditions administered this assessment to their students at the beginning of the school year as a pretest and at the end of the school year as a posttest. The LevelSet is an online assessment that uses the Lexile Framework® to assess students' Lexile reading level. Teachers in the treatment condition administered it to their students at the beginning and end of the school year. In addition to student measures, the study used multiple teacher measures: weekly treatment teacher online implementation logs, a spring comparison-teacher survey, and spring classroom observations of treatment and comparison teachers.

Achieve3000 Program Implementation

KEY FINDING:

On average, treatment teacher implementation scores from the weekly logs, observation, and usage data were 71.05%. Thus, treatment teachers implemented the Achieve3000 program with moderate fidelity.

To measure program implementation and calculate an implementation fidelity score for each teacher, Magnolia Consulting and Achieve3000 jointly developed minimum implementation requirements for this study. These minimum implementation requirements asked teachers to implement the Achieve3000 program for at least 90 minutes per week. Teachers fulfilling the weekly implementation requirements had an implementation score of 100%.

The analyses of implementation data from weekly implementation logs, observations and student usage data, revealed an average implementation score of 71.05% across treatment teachers. This score reflects real-world implementation variation due to competing district and state requirements, assessments, holidays, weather delays, technology issues, and other issues.

Treatment students' online use of the Achieve3000 program was tracked by the LevelSet data. On average, students in this study logged into the Achieve3000 program 101 times during the year and logged 30.53 program hours. Treatment students averaged 50.58 valid activities during the year and 1.49 activities each week. Overall, they completed an average of 30.01 passing activities during the year. An activity was considered passing when a student answered 75% or more of the questions in the activity correctly. Achieve3000 uses this threshold as a measure for determining whether students are applying themselves to the activity and working within their instructional zone.

Comparison of Program Implementation across Study Conditions

Comparison teachers reported using various core literacy programs for more days per week than treatment teachers reported using Achieve3000. Comparison teachers reported planning and preparing for a longer period of time than treatment teachers and reported using more supplemental materials during the school year.

Treatment Group Student Learning Results

KEY FINDING:

As a group, students who used Achieve3000 during the 2014/15 school year demonstrated statistically significant and substantively important gains on the GMRT-4 Vocabulary, Reading Comprehension, and Total tests. They also demonstrated statistically significant and substantively important gains in LevelSet Lexile points.

Evaluators examined learning gains using multilevel modeling analyses. Findings regarding GMRT-4 gains showed that on average, treatment students who used Achieve3000 demonstrated statistically significant and substantively important learning gains on the GMRT-4 Vocabulary, Reading Comprehension, and Total Reading tests (effect sizes of 0.43, 0.47, and 0.48). Additionally, treatment students had statistically significant and substantively important LevelSet Lexile point gains (effect size of 0.33).

Figure 1 shows that as a group, third-grade treatment students exceeded the MetaMetrics expected gain of 100 Lexile points and sixth-grade treatment students exceeded the expected gain of 70 points. On average, ninth-grade students did not meet the expected gain of 50 Lexile points.

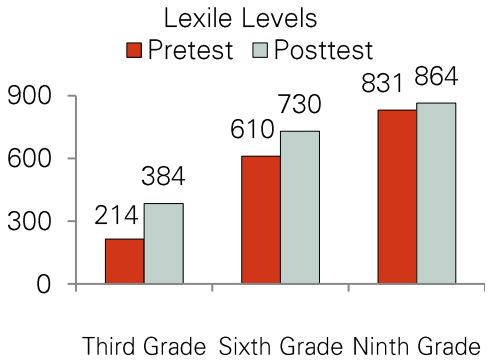


Figure 1. Pretest and posttest LevelSet Lexile levels for treatment students.

Exploratory analyses of the LevelSet data showed that as a group, more than half of the Achieve3000 users met or exceeded their expected growth in Lexile points. Additionally, whereas only 10.94% of treatment students met or exceeded the LevelSet college and career readiness benchmark at pretest, 23.44% met or exceeded this benchmark by posttest, and this difference was statistically significant. Student completion of Achieve3000 activities was positively—and statistically significantly—associated with Lexile point gains but not statistically significantly associated with GMRT-4 Vocabulary, Reading Comprehension, or Total Reading gains. However, performing well (i.e., passing 75% or more of the activities, on average) was statistically significantly related to greater GMRT-4 Reading Comprehension and Total Reading gains, as well as Lexile point gains. There was not a statistically significant relationship between afterschool use of Achieve3000 and learning gains.

Relationships between Treatment Teachers' Implementation Fidelity of Achieve3000 and Student Learning Gains

The degree to which treatment teachers implemented Achieve3000 with fidelity varied. Within the range of implementation for this study, there were positive relationships between implementation fidelity and learning gains, but they were not statistically significant for the GMRT-4 Vocabulary, Reading Comprehension, or Total Reading Tests. However, the relationship between implementation fidelity and Lexile point gains was statistically significant, with implementation fidelity increases of 10% associated with average gains of 31.01 Lexile points.

Student Learning Results Comparing Treatment and Comparison Groups

KEY FINDING:

Achieve3000 had a statistically significant positive impact on GMRT-4 Reading Comprehension and Total Reading performance. It did not have a statistically significant impact on GMRT-4 Vocabulary performance.

Evaluators used multilevel modeling analyses to determine if Achieve3000 had a

statistically significant impact on reading when compared to participating schools' usual language arts programs. Findings showed that overall, Achieve3000 did not have a statistically significant impact on students' GMRT-4 Vocabulary test scores (effect size of 0.12). However, the program had a statistically significant positive impact on students' GMRT-4 Reading Comprehension and Total Reading test scores (effect sizes of 0.22 and 0.20). The effect sizes for Reading Comprehension and Total Reading approached the WWC standards for substantively important effects.

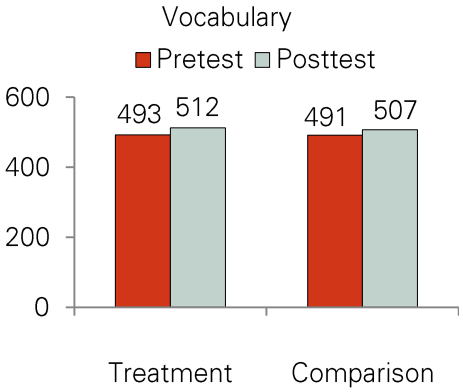


Figure 2. GMRT-4 Vocabulary unadjusted scale score means by study condition and time.

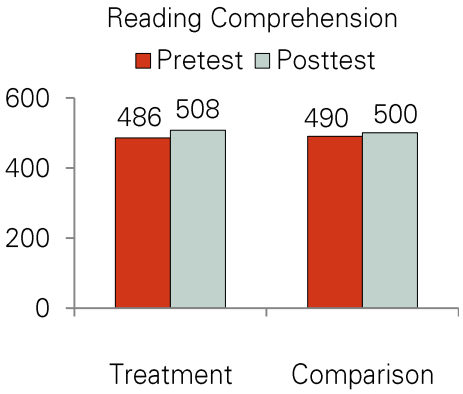


Figure 3. GMRT-4 Reading comprehension unadjusted scale score means by study condition and time.

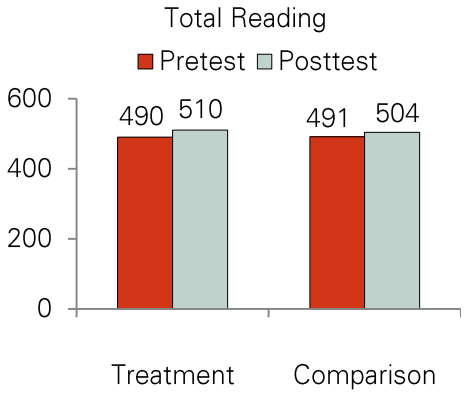


Figure 4. GMRT-4 Total Reading unadjusted scale score means by study condition and time.

Exploratory Analyses Comparing Achieve3000 Impacts by Grade

KEY FINDING:

Findings suggest that the impact of Achieve3000 on reading performance varied by grade, with the greatest impacts on ninth-grade reading.

After examining the impact of Achieve3000 across grades, evaluators used multilevel modeling to explore program impacts within each grade. Findings revealed that on average, third-grade students who participated in Achieve3000 performed similarly to comparison-group students who used their schools' typical literacy programs (with effect sizes of -0.02, 0.02, and 0.06 for Vocabulary, Reading Comprehension, and Total Reading). For sixth-grade students, there were no statistically significant differences in posttest GMRT-4 scores by study condition, but the effect sizes of 0.21, 0.22, and 0.22 for Vocabulary, Reading Comprehension, and Total Reading favored Achieve3000 users and approached the WWC standards for substantively important effects. Finally, although there were also no statistically significant differences in average ninth-grade treatment and

comparison-group posttest reading performance, the effect sizes of 0.28, 0.51, and 0.44 for Vocabulary, Reading Comprehension, and Total Reading favored Achieve3000 users and exceeded WWC standards for substantively important effects. Because these subgroup analyses had less statistical power than main analyses to detect effects, readers should use caution when interpreting the statistical significant of findings.

Exploratory Analyses Comparing Achieve3000 Impacts for ELL Students

KEY FINDING:

Findings suggest that ELL students who used Achieve3000 performed similarly on the GMRT-4 as ELL students who used their schools' typical literacy programs.

Evaluators also examined the impact of Achieve3000 on ELL students. The exploratory analyses showed no statistically significant differences or substantively important effect sizes by study condition for ELL students. The effect sizes for Reading Vocabulary, Reading Comprehension, and Total Reading (i.e., -0.07, -0.06, and -0.05) corresponded to WWC improvement indices of -3, -2, and -2 percentile points, respectively. These findings suggest that on average, ELL students who used Achieve3000 performed similarly to comparison-group ELL students who used their schools' typical literacy programs. Readers should interpret these findings with caution because of the small sample sizes for this ELL subgroup,

Teacher Perceptions of Achieve3000 and Comparison Programs

Overall, treatment teachers found the Achieve3000 program components useful and described many benefits to the program including differentiation, less time required for lesson preparation, and positive effects on student engagement and learning. However, some treatment teachers were frustrated with the monotony of the program routine, the amount of time the program took away from their core curriculum, the brevity of the training, program navigation, and technology issues. Many teachers offered suggestions for improvement such as improving teacher tools, adding visuals for vocabulary, improving various digital components, and navigation features. Most teachers said they would use the program again next year, but many teachers said they would implement it differently.

Comparison teachers relied heavily on supplemental materials and reported that their core materials were not engaging or dated. Comparison teachers struggled with finding interesting materials to meet all students' needs, but were generally happy with their students' achievement.

Comparisons of Teacher Program Perceptions across Study Conditions

KEY FINDING:

Treatment teachers described the Achieve3000 program as having higher student engagement, having the appropriate amount of materials to cover, and more applicable pacing than comparison programs.

When comparing treatment and comparison perceptions treatment teachers described the Achieve3000 as having higher student engagement, having the appropriate amount of materials to cover, and more applicable pacing than comparison programs. According to study teachers, Achieve3000 more *adequately* or *very*

adequately supported below-level, on-level and advanced-level readers, English Language Learners and special education students than comparison programs. For student skills, Achieve3000 more effectively supported building academic vocabulary, comprehending complex text, and critically evaluating informational text than comparison programs, while comparison programs more effectively supported reading fluency.

Limitations

This rigorous study had some limitations worth noting. First, findings only generalize to schools that met the participation requirements for this study, which included specific technology access and infrastructure as well as multiple ELA teachers at participating grades. Because two teachers who were uncomfortable with the program's technology requirements dropped out, it is unclear how teachers who are less skilled with technology might implement the program, and if that would have impacted study findings. It is possible that if teachers had received additional training earlier in the study, their implementation fidelity might have increased, which could have impacted the study findings. Finally, readers should use caution when interpreting within-grade subgroup analyses, as the relatively smaller sample sizes limited the statistical power of the analyses to detect effects.

Summary and Conclusions

Results from this 2014/15 evaluation of Achieve3000 showed that on average, treatment teachers implemented the program with moderate fidelity. Students who used Achieve3000 demonstrated statistically significant and substantively important gains on the GMRT-4 Vocabulary, Reading Comprehension, and Total Reading tests, as well as in their LevelSet Lexile levels. Comparisons of students who used Achieve3000 with students who used the schools' standard ELA programs showed that overall, Achieve3000 had a statistically significant but not substantively important impact on GMRT-4 Reading Comprehension and Total Reading. Within-grade analyses showed that Achieve3000 had impacts on sixth-grade reading that were not statistically significant but approached the WWC threshold for being substantively important, and it had substantively important impacts on ninth-grade GMRT-4 Vocabulary, Reading Comprehension, and Total Reading. Finally, ELL subgroup analyses revealed no statistically significant differences or substantively important effect sizes regarding the impact of Achieve3000 on ELL students' GMRT-4 Reading Vocabulary, Reading Comprehension, or Total Reading.

Treatment teachers generally reported that the Achieve3000 program components were useful and comprehensive, and they described many benefits to the program. However, some treatment teachers also expressed frustration with various aspects of the program. Comparisons of perceptions across conditions showed that treatment teachers often described Achieve3000 as having higher student engagement, having the appropriate amount of materials to cover, and more applicable pacing than comparison programs. According to study teachers, Achieve3000 more adequately or very adequately supported below-level, on-level and advanced-level readers, English Language Learners and special education students than comparison programs. For student skills, Achieve3000 more effectively supported building academic vocabulary, comprehending complex text, and critically evaluating informational text than comparison programs, while comparison programs more effectively supported reading fluency.

ACHIEVE3000
believe

An Efficacy Study of the Achieve3000 Programs

WHO WAS IN THE STUDY?

16 schools

46 teachers

1012 students

During the 2014/15 school year, 16 schools in four suburban and city districts across the U.S (the West South region, the East North Central region and the Pacific region) implemented Achieve3000. The study included 46 teachers and 1012 students.

HOW DID THEY IMPLEMENT THE PROGRAMS?

Treatment Teachers implemented the Achieve3000 program for 2 days per week, averaged 90 minutes on the program, and at least one lesson.
Comparison Teachers implemented their existing ELA programs.

WHAT DID THEY SAY?

"I enjoyed that each article was pushed out at their own level. I thought the articles were interesting and so did the students. I loved the growth reports and the Sunday report that was sent. **Overall one of the best programs I have worked with!** Teacher Quote

"I love that this allows me to level each of their articles. I can talk to my entire class about a topic and have them read at their own level. ***I have not found another literacy program that does this so thoroughly****" Teacher Quote*

Treatment teachers disliked the monotony of the program routine, the amount of time the program took away from their core curriculum, the brevity of the training, program navigation, and technology issues.

WHAT WERE THE IMPACTS ON STUDENT ACHIEVEMENT?

ACHIEVE 3000 STUDENTS

Statistically significant and substantively important learning gains on the GMRT-4 and LevelSet assessments

ACHIEVE 3000 VS. COMPARISON

Statistically significant positive impact on GMRT-4 Reading Comprehension and Total Reading when compared to study schools' standard ELA curricula.

ACHIEVE 3000 BY GRADE

Substantively important positive impacts on ninth-grade GMRT-4 Vocabulary, Reading Comprehension, and Total Reading.

Acknowledgements

We sincerely value the collaborative efforts that contributed to the success of this study. We are especially grateful for the administrators, teachers, and students who participated. Additionally, we are thankful for the support that Achieve3000 provided throughout the project, and appreciate the value they place on independent evaluation of their products. Finally, we want to thank the entire Magnolia Consulting team for their support of the study.

The authors,

Lisa Shannon, Ph.D.
Billie-Jo Grant, Ph.D.

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

Table of Contents

Introduction

Literacy continues to be a priority in U.S. and global educational policy. The ability to read is essential for engaging in and contributing fully to society. Adults with low literacy skills tend to have lower incomes, struggle for employment, are less likely to vote, and are more likely to have legal trouble (National Institute for Literacy, 2008). Reading is a key factor in student achievement and progress through school (Rasinski et al., 2005; Mackenzie, Noella, and Hemmings, 2014). Thus, not only does reading ability have a significant impact on students' educational careers, it also has widespread implications for their economic livelihood and social and civic success (Lesnick, George, Smithgill, & Gwynn, 2010; NELPR, 2009).

Twenty-first century classrooms are filled with diverse learners who represent varied cultural and linguistic backgrounds and also embody a broad spectrum of cognitive abilities, knowledge bases, and learning styles. Differentiated instruction is a proven method for reaching students with different interests, preferences, learning strengths, and needs (Huebner, 2010; Reis, McCoach, Little, Muller, & Kaniscan, 2011). Differentiated instruction also makes it possible for teachers to enhance the success of students with disabilities, English language learners, students who are gifted, and students considered at risk for leaving school before completion (Alberta Education, 2010). However, though most teachers agree that differentiated instruction is important, they often find it easier to manage one lesson and one group of students than to plan different activities for multiple groups (Moody & Vaughan, 1997; Schumm, Moody, & Vaughn, 2000). New teachers, who are most likely to be in schools with a high diversity of learners, are the least likely to be prepared to deliver differentiated instruction (Parsons, Malloy, Vaughn, & La Croix, 2014; Santangelo & Tomlinson, 2012).

Digital technology is recognized as a valuable tool for supporting differentiated and personalized instruction (Watson & Watson, 2012). Technology-based differentiated instruction can help teachers tailor instruction to individual reading levels and provide students with high-quality learning experiences.

Achieve3000 is a differentiated online literacy program that provides standards-based content for students in grades 2 through 12. Achieve3000 strongly believes in providing the highest quality materials for use in the classroom. As such, it contracted with Magnolia Consulting, LLC, an external, independent consulting firm specializing in research and evaluation, to conduct an efficacy study of the Achieve3000 program.

Research Design

The primary purpose of this evaluation study was to examine the efficacy of Achieve3000 in increasing students' reading achievement. The study also examined teachers' implementation and perceptions of Achieve3000. Evaluators used a cluster randomized control trial (RCT) design to answer the following specific evaluation questions:

1. Did teachers implement Achieve3000 with high levels of fidelity based on the study's implementation guidelines?
2. Did students who used Achieve3000 during the 2014/15 school year demonstrate statistically significant gains in reading achievement? If so, what was the magnitude of these gains?
3. Did most students who used Achieve 3000 during the 2014/15 school year meet or exceed expected Lexile Level growth?
4. Were students who used Achieve3000 during the 2014/15 school year more likely to meet college and career readiness benchmarks by the end of the study than they were at the beginning of the study?
5. How did treatment student performance on Achieve3000 multiple choice activities relate to their learning gains during the study?
6. Was afterschool use of Achieve3000 related to student learning gains?
7. Was there a statistically significant relationship between the degree to which teachers implemented Achieve3000 with fidelity and student learning gains?
8. Did Achieve3000 have a statistically significant impact on reading achievement when compared to standard ELA programs? If so, what was the magnitude of the impact?
9. What were teachers' perceptions regarding the quality and utility of Achieve3000?

Methodological Approach

For this randomized control trial, Magnolia Consulting evaluators randomly assigned participating third-, sixth-, and ninth-grade teachers to study conditions. Teachers who participated in the treatment condition implemented Achieve3000 during the 2014/15 school year. Teachers who participated in the comparison condition implemented their current literacy programs but not Achieve3000. Therefore, half of the teachers in the study used Achieve3000, and the other half used their school's regular literacy programs. These types of clustered RCT designs are useful for studying the efficacy of educational programs in schools where students are grouped in classrooms and teachers deliver curricula at the classroom level (Raudenbush & Bryk, 2002). This study design enabled evaluators to estimate differences in treatment and comparison students' end-of-study reading skills and to determine if any differences were statistically significant.

Analytic Approach

Prior to analyzing study findings, evaluators followed precise data preparation and cleaning procedures (see Appendix A). Evaluators then employed several types of analyses to address the evaluation questions of interest. These included descriptive analyses to explore

student assessment data and multilevel modeling analyses[1] to determine (a) if students who participated in Achieve3000 demonstrated statistically significant gains in reading achievement, and (b) to estimate the impact of Achieve3000 on student reading achievement when compared to a typical ELA program.[2] Evaluators also conducted exploratory analyses using a variety of methods such as *t*-tests and McNemar's test.

When appropriate, evaluators calculated effect sizes to determine the magnitude of program effects. Standardized effect sizes reflect the strength of a relationship between two variables, or the magnitude of the effect of a treatment (Borenstein, Hedges, Higgins, & Rothstein, 2009). For example, a treatment effect size of 1.0 indicates that the treatment group's mean outcome was 1.0 standard deviation greater than the comparison group's mean outcome. For analyses of implementation and perception data, evaluators conducted a variety of analyses as appropriate, including various descriptive analyses, parametric, and nonparametric tests.

Evaluators considered findings statistically significant using an alpha level of 0.05. When interpreting effect sizes, evaluators followed the What Works Clearinghouse (WWC) guidelines that consider effect sizes substantively important when they are greater than or equal to 0.25 (What Works Clearinghouse, 2014). Therefore, for this study, evaluators considered effect sizes as notable when they met or exceeded the threshold of the absolute value of 0.25.

Measures

The study included multiple measures. Student measures were used to determine the impact of Achieve3000 on learning, and teacher measures were used to gauge teachers' use and perceptions of the Achieve3000 and comparison programs.

Student Measures

For this evaluation, participating treatment and comparison teachers administered the Gates MacGinite Reading Test, fourth edition (GMRT-4) to their students as a pretest and posttest. Additionally, treatment teachers administered the LevelSet to their students as a pretest and posttest.

GMRT-4 Standardized Assessment

This study's main student reading assessment for this study was the GMRT-4, a group-administered, norm-referenced reading assessment with broad visibility and acceptance in the field. The GMRT-4 is appropriate for the grade levels of students who participated in this study and yields scores for Vocabulary, Reading Comprehension, and Total Reading. The GMRT-4

[1] It was important to use hierarchical linear modeling (HLM) for the study's main analyses because of the clustering of students in teachers' classrooms. This clustering created a hierarchical, interdependent data structure because students who had the same classroom teacher also shared other teacher and classroom experiences, which might have affected the ways in which they responded to the educational programs they were exposed to during the study period (Borman et al., 2005). Traditional regression approaches assume independent observations, making them inappropriate for analyzing data collected from this type of cluster randomized trial.

[2] In this report, the term "impact" refers to the difference in outcomes between the treatment and comparison groups.

offers two forms—the S and T—which are appropriate for fall pretesting and spring posttesting, respectively. Riverside Publishing scoring services provide GMRT-4 raw scores, grade equivalent scores, and vertically-scaled extended scale scores. For this study, evaluators used pretest and posttest extended scale scores in the main analyses.

LevelSet

The LevelSet, developed by Achieve3000 and MetaMetrics, Inc., served as an additional assessment of reading skills for this study's treatment group. The LevelSet, an online assessment that uses the Lexile® Framework, measures both the difficulty of the text and students' reading abilities to assess students' Lexile reading scores. The assessment's Lexile measures are based on national norms, but the LevelSet is considered a criterion-reference test that provides teachers with a way to match students to informational texts by yielding individual Lexile placement scores for nonfiction texts. For this study, treatment students completed the LevelSet at the onset of the study, and the Achieve3000 program matched students to appropriately-leveled reading materials based on their resulting Lexile scores. Evaluators used pretest and posttest Lexile levels in the main treatment group analyses for this study. Evaluators also used additional data—such as Expected Reading Growth, pretest and posttest Career Readiness Levels, data regarding student participation performance on Achieve3000 multiple choice activities, and afterschool use—in exploratory treatment group analyses.

Teacher Measures

This evaluation included online weekly treatment-teacher implementation logs and a one-time online comparison-teacher survey to assess teachers' use and perceptions of their literacy programs. Additionally, evaluators conducted a spring site visit at each participating school to observe treatment teachers' use of Achieve3000 and comparison teachers' use of their regular literacy programs.

Treatment Teacher Implementation Logs

Throughout the study period, treatment teachers completed weekly online implementation logs designed to assess the breadth and depth of their implementation of Achieve3000 in the classroom. Each week, teachers spent approximately 15 minutes completing these logs, reporting the extent to which they adhered to Achieve3000 implementation guidelines, the dosage received by treatment students, student responsiveness, program differentiation or modifications, and teacher perceptions of the materials. These areas of implementation correspond with key areas of implementation fidelity as suggested by Carroll et al. (2007). Additionally, teachers used the logs to report interruptions in their implementation of the program (e.g., fire drill, field trips, etc.) and student attrition. The final implementation log included additional open-ended questions that encouraged summative reflection regarding the following:

- the classroom learning environment, including important school culture and student characteristics that influenced the learning experience,
- perceptions of program strengths and challenges,
- changes to instructional practices,
- observations of student impacts, and
- future program use.

In addition to serving as a measure of implementation fidelity and variety, the weekly logs permitted evaluators to report any local events that occurred during the study. At the conclusion of the study, evaluators aggregated log data and calculated individual ratings of each teacher's level of implementation.

Comparison Teacher Surveys

In the spring, evaluators administered a one-time online survey to comparison teachers. This survey enabled comparison teachers to report on their implementation and perceptions of their existing literacy curricula and supplemental use. To facilitate comparisons across study conditions, the survey items aligned, to the extent possible, with items on the treatment-teacher implementation log.

Classroom Observations

Also in the spring, evaluators conducted site visits to observe treatment and comparison classrooms. To facilitate these observations, evaluators created observation protocols. The protocols used the same format, but they differed by condition because the treatment protocol included items specific to Achieve3000, and the comparison protocol did not. Specifically, treatment protocols addressed the following five constructs: (a) teacher-student interactions, (b) equipment and technology, (c) procedures associated with the use of Achieve3000, (d) Achieve3000 program components, and (e) student engagement. Comparison protocols addressed the following four constructs: (a) teacher-student interactions, (b) equipment and technology, (c) instructional strategies and procedures, (d) the lesson, and (e) student engagement. Each larger construct was subdivided into smaller sections with checklists and notes regarding particular classroom observations. Evaluators measured implementation using a 25-item treatment teacher rubric and a 23-item comparison teacher rubric. Each item was scored on a scale of 0 to 3, (0=*Not at all, does not meet this indicator*; 1=*Partially, indicator is apparent but inconsistent*; 2=*Mostly, indicator is apparent most of the time*; 3=*Fully, consistently meets indicator*).

Study Procedures

Magnolia Consulting evaluators worked closely with participating sites and with Achieve3000 throughout the study to ensure that all study procedures were carried out as planned.

Site Selection and Group Assignment

Magnolia Consulting worked collaboratively with Achieve3000 to recruit the sites for this randomized control trial study. For this study, schools were eligible to participate if they had at least two teachers who were not current users of Achieve3000 at grades 3, 6, and 9, and if they would be able to implement the program for at least 90 minutes per week. To reach potential sites, Magnolia Consulting contacted schools referred to them by Achieve3000 and drawn from its database of district contacts in curriculum and instruction. After contacting potential sites, Magnolia Consulting screened them using a district-level informational survey that aligned to the criteria and included relevant demographic and contact information, as well as a telephone interview to follow up with any questions about participation requirements. Once the participating school was selected, a school and district representative signed a memorandum of understanding that outlined all study procedures and responsibilities.

Magnolia Consulting evaluators used random assignment procedures in SPSS statistical software to randomly assign participating teachers to study conditions in each grade (3, 6, and 9). Evaluators took several steps to avoid contamination across study conditions. First, the design itself reduced contamination compared to a design using student-level random assignment because all students within a classroom used the same program (i.e., either Achieve3000 or the school's regular ELA program). Comparison group teachers did not have any access to the Achieve3000 program until the end of the study. Additionally, evaluators asked treatment teachers not to discuss any aspects of the Achieve3000 program with comparison group teachers, and a site study coordinator was assigned to ensure contamination did not occur.

Study Timeframe

This study occurred during the 2014/15 school year. Evaluators conducted site recruitment in the spring and summer of 2014. Site selection and random assignment to study conditions occurred in August 2014. Immediately following random assignment to study conditions, Magnolia Consulting evaluators conducted an on-site study orientation with participating teachers and the site coordinator. After the study orientation, Achieve3000 conducted an initial Achieve3000 training seminar with treatment teachers. Treatment and comparison teachers administered the GMRT-4 in the fall and spring, and treatment teachers administered the LevelSet in the fall and spring[3]. Magnolia Consulting completed a site visit to observe treatment and comparison teacher classrooms in the spring and administered a comparison teacher survey in the spring. Table 1 displays an overview of the study's time frame.

Table 1. Timeline of Study Activities

STUDY ACTIVITY	Spring & Sum. 2014	Sept. 2014	Oct. 2014	Nov. 2014	Dec. 2014	Jan. 2015	Feb. 2015	Mar. 2015	Apr. 2015	May 2015	Jun. 2015
Site recruitment and random assignment to study condition	→										
Training, study orientation, and program implementation begins	→										
Administration of GMRT-4 for treatment and comparison classrooms		◆								◆	
Administration of LevelSet for treatment students		◆								◆	
Implementation of Achieve3000 in treatment classrooms		→	→	→	→	→	→	→	→	→	
Administration of weekly treatment teachers' implementation logs		→	→	→	→	→	→	→	→	→	

[3] In one district, two sixth-grade teachers and three ninth-grade teachers administered the GMRT-4 pretest about six weeks later than other teachers because they misunderstood administration specifications. Based on the field's recommendations for handling late pretests (i.e., Schochet, 2008), evaluators conducted sensitivity analyses to determine if the late pretesting for these classrooms had any impacts on study findings. The sensitivity analyses revealed that these classrooms' relatively later pretest administrations did not impact student GMRT-4 performance.

STUDY ACTIVITY	Spring & Sum. 2014	Sept. 2014	Oct. 2014	Nov. 2014	Dec. 2014	Jan. 2015	Feb. 2015	Mar. 2015	Apr. 2015	May 2015	Jun. 2015
Site visit (treatment and comparison teacher classroom observations)								◆			
Administration of comparison teacher survey									◆		
End of study											◆

Implementation Fidelity

Evaluators calculated teachers' fidelity to implementation of their programs using the minimum program requirements as a denominator for each usage variable. For this study, teachers' implementation fidelity scores were calculated using data from three sources: teachers' weekly log reports, teacher observations, and student usage reports. Each of these variables was equally weighted (33%) to calculate an overall implementation fidelity score for each teacher, and these teacher fidelity scores were averaged to calculate an overall implementation fidelity score for the study.

For the weekly log data, evaluators used the following variables and criteria to calculate an implementation score:

- Number of days teachers used Achieve3000 each week (minimum two days),
- Number of minutes teachers used Achieve3000 each week (minimum 90 minutes),
- Number of lessons teachers covered each week (minimum of two lessons), and
- If the teacher completed the multiple choice activity question(s) each week.

For the teacher observation data, evaluators used 25 items from the following five constructs to calculate an implementation fidelity score:

- Teacher-student interactions,
- Equipment and technology,
- Procedures associated with the use of Achieve3000,
- Achieve3000 program components, and
- Student engagement.

For the usage reports, evaluators used the following variables and criteria to calculate an implementation fidelity score:

- Total valid activities (minimum two activities per week), and
- Passing activities[4] (minimum two activities per week).

Settings

This study sample came from four districts located in three different regions of the United States: the West South region, the East North Central region and the Pacific region. Two districts were classified as "Suburb: Large" and two districts were classified as "City: Large"

[4] An activity was considered passing when a student answered 75% or more of the questions in the activity correctly. Achieve3000 uses this threshold as a measure for determining whether students are applying themselves to the activity and working within their instructional zone.

(U.S. Department of Education, National Center for Education Statistics, 2015). As shown in Table 2, the number of schools in each district ranged from 21 to 23, and the student population in each district ranged from 19,257 to 130,271. The student-teacher ratios also varied and ranged from 15.57 to 25.14. Districts A and B had a higher percentage of Hispanics or Latinos than Districts C and D. Race also varied across the four districts, with greater diversity in Districts A, B, and C than District D, which was predominately White. Over 20% of District A and B students were classified as English Language Learner (ELL) students compared to District C at 11.52% and District D at 2.49%. Districts were moderately similar in the percentage of students with Individualized Education Programs (IEPs), which ranged from 10.87% to 14.38%.

Table 2. Site Characteristics

	District A	District B	District C	District D
Geographic location* and city description	Pacific City: Large	Pacific Suburb: Large	East North Central City: Large	West South Central Suburb: Large
Number of Schools	231	39	27	21
Total student enrollment	130,271	34,922	19,257	20,209
Student–Teacher ratio	19.83	25.14	15.57	20.22
Ethnicity				
Hispanic or Latino	38.28%	43.96%	8.55%	2.01%
Non-Hispanic or Latino	61.72%	56.04%	91.45%	97.99%
Race				
White alone	44.79%	42.51%	68.19%	92.32%
Black or African American alone	12.30%	21.10%	15.21%	2.76%
American Indian or Alaska Native alone	0.73%	0.86%	3.43%	0.10%
Asian alone	13.43%	4.31%	3.58%	2.53%
Hawaiian or other Pacific Islander alone	0.64%	0.53%	0.09%	0.01%
Some other race alone	19.17%	22.49%	3.35%	0.56%
Two or more races	8.94%	8.20%	6.16%	1.72%
ELL Students	22.66%	24.35%	11.52%	2.49%
Students with IEPs	11.12%	12.00%	14.38%	10.87%

Sources: The National Center for Education Statistics at http://nces.ed.gov/ccd/districtsearch/
*Geographic location from U.S. Census Bureau

Participants

This section presents information about the teacher and student participants included in this study. It describes participant attrition, the analysis sample, teacher and student demographic characteristics, and student group equivalence.

Attrition

To determine attrition, evaluators compared the numbers of teacher and student participants at the start and end of the study. At the beginning of the school year, two of the 25 original treatment teachers left the study because they did not feel comfortable implementing

the program and did not receive follow-up training. Thus, the treatment teacher attrition rate was 8.00%. There was no attrition in the comparison group, where all 23 comparison teachers who began the study remained in the study throughout its duration.

In addition to examining teacher attrition, evaluators also determined student attrition by assessing the number of students who were enrolled at the start of the study but were no longer enrolled by the end of the study. In this study, there were 1,270 student participants at the start of the year (683 treatment and 587 comparison students), and 1,039 of these students remained enrolled in a study classroom for the duration of the study period. This difference of 231 students (156 treatment students and 75 comparison students) counted towards student sample attrition, yielding an overall attrition rate of 18.19%. Evaluators then calculated the differential attrition rate to compare attrition by study condition. The treatment group attrition rate was 22.84%, and the comparison group attrition rate was 12.78%, yielding a differential attrition rate of 10.06%. A chi-square test revealed that this difference was statistically significant, χ^2 (1, n = 1270) = 20.81, p = >.001. More specifically, a larger number of treatment students than comparison student left the study between its start and conclusion. It is important to note that a large wave of treatment student attrition (35.90%) occurred when the two treatment teachers left the study, resulting in the removal of their 56 students from the study.

Analysis Sample

To be included in the teacher analysis sample, teachers had to participate throughout the entire length of the study year. Two treatment teachers left the study, leaving a final analysis sample of 46 teachers (23 treatment and 23 comparison teachers). To be included in the student analysis sample, students had to be enrolled in a treatment or comparison classroom throughout the study period. Thus, the 231 students who left the study prior to the end (156 treatment and 75 comparison students) were removed from the analysis sample. Additionally, 27 students were removed from the analysis sample because they did not have adequate program dosage (13 treatment and 3 comparison students), they had been exposed to the Achieve3000 program (7 comparison students) or because they changed condition partway through the study (2 treatment and 2 comparison students). Based on these criteria, the final analysis sample included 512 treatment and 500 comparison students, for a total of 1,012 students. The following section describes the demographic characteristics of the students and teachers included in the analysis sample.

Teacher Participants

The final analysis sample included 46 teachers (23 treatment and 23 comparison) from 16 schools in four districts who participated throughout the entire length of the study year.

Demographics

As part of the study, all treatment and comparison teachers provided information about their level of education and years of teaching experience. In the treatment group, 12 teachers reported holding a bachelor's degree and 11 teachers reported holding a master's degree. The comparison teachers had the reverse, with 11 teachers reporting holding a bachelor's degree and 12 teachers reporting holding a master's degree. Treatment teachers reported teaching for an average of 11.59 years with an average of 5.87 years at their current school and 6.09 years

at their current grade. Comparison teachers reported teaching for an average of 12.35 years with an average of 4.78 years at their current school and 5.39 years at their current grade. Lastly, the number of students per treatment teacher ranged from 20 to 34, for an average of 27.22 students. Comparison teachers had a range of 15 to 36 students, for an average of 25.04 students per teacher.

Evaluators then determined if there were differences in these characteristics by study condition by conducting independent samples *t*-tests. As seen in Table 43, these analyses showed no statistically significant differences between treatment and comparison groups in regard to teaching experiences and the number of students per teacher.

Table 3. Teacher Demographics by Group

	Comparison Teachers (*N* = 23)			Treatment Teachers (*N* = 23)			Independent *t*-test		
Characteristics	*n*	Mean	*SD*	*n*	Mean	*SD*	*df*	*t*	Sig. (alpha =.05)
Experience									
Total years teaching	23	12.35	9.58	23	11.59	8.30	44	0.29	.77
Years at current school	23	4.78	4.40	23	5.87	4.49	44	-0.83	.41
Years at current grade	23	5.39	4.46	23	6.09	4.86	44	-0.51	.62
Number of Students	23	25.04	5.17	23	27.22	4.53	44	-1.52	.14

Student Participants

The final analysis sample included 512 treatment and 500 comparison students. In this section, evaluators present these students' demographic information, as well as results from the group equivalence analyses.

Demographics

Nearly half of the students in the sample were in the sixth grade (48.81%), 26.68% were in the third grade and 24.51% were in the ninth grade. As shown in Table 4, slightly more students in the study were male (52.77%) than female (47.23%). A little over a third of the analysis sample was of Hispanic or Latino ethnicity (37.06%). Sixty-seven percent of students were classified as White, 20.55% as Black or African American, 4.45% as Asian and 7.91% as two or more races or other. Across both study conditions, 12.55% of students were categorized as English Language Learners (ELL). Three districts provided student-level data for the remaining demographic variables. Of these students 61.98% qualified for free or reduced-priced lunch (FRL), 12.07% were classified as special education students (SPED), and 2.31% were classified as Section 504 students.

Table 4. Student Demographics by Group

	Comparison Students (*N* = 500)		Treatment Students (*N* = 512)		Total Students (*N* = 1012)		Chi-square Results	
Characteristics	Percent	*n*	Percent	*n*	Percent	*n*	X^2 Value	Sig. (alpha =.05)
Grade								
Third	28.60%	143	24.80%	127	26.68%	270		
Sixth	46.20%	231	51.37%	263	48.81%	494	2.94	.23
Ninth	25.20%	126	23.83%	122	24.51%	248		

Characteristics	Comparison Students (*N* = 500)		Treatment Students (*N* = 512)		Total Students (*N* = 1012)		Chi-square Results	
	Percent	*n*	Percent	*n*	Percent	*n*	X^2 Value	Sig. (alpha =.05)
Gender								
Male	48.80%	244	56.64%	290	52.77%	534	5.93	.01
Female	51.20%	256	43.36%	222	47.23%	478		
Ethnicity								
Hispanic/Latino	39.40%	197	34.77%	178	37.06%	375	2.13	.14
Not Hispanic/Latino	60.60%	303	65.23%	334	62.94%	637		
Race								
White	70.00%	350	64.26%	329	67.09%	679	4.42	.22
Black/African America	18.40%	92	22.66%	116	20.55%	208		
Asian	3.80%	19	5.08%	26	4.45%	45		
Two or More Races/Other	7.80%	39	8.01%	41	7.91%	80		
Free/Reduced Lunch								
FRL	62.14%	192	61.82%	183	61.98%	375	.00	1.00
Non-FRL	37.86%	117	38.18%	113	38.02%	230		
English Proficiency								
ELL	13.60%	68	11.52%	59	12.55%	127	.81	.37
Non-ELL	86.40%	432	88.48%	453	87.45%	885		
Special Education								
Special Ed.	10.36%	32	13.85%	41	12.07%	73	1.43	.23
Non-Special Ed.	89.64%	277	86.15%	255	87.93%	532		
Section 504								
Section 504	2.91%	9	1.69%	5	2.31%	14	.53	.47
Non-Sect. 504	97.09%	300	98.31%	291	97.69%	591		

Note a. Free or Reduced Priced Lunch, Section 504 and Special Education analyses do not include students from District C. This district provided classroom level data per district requirements.

Group Equivalence

Evaluators examined pretest equivalence between the treatment and comparison group by conducting chi-square analyses on demographic characteristics (Table 4) and running multilevel modeling analyses on pretest reading skills (Table 5). Analyses show that treatment and comparison groups were similar in all characteristics examined except for gender. Specifically, there was a higher percentage of males in the treatment group and a higher percentage of females in the comparison group. Likewise, multilevel modeling analyses revealed no statistically significant differences between groups on mean pretest student GMRT-4 Vocabulary, Reading Comprehension, or Total Reading scores.

Table 5. Group Equivalence on Pretest Reading Skills

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. *df*	*p*-value	Effect Size
Pretest Vocabulary	3.08	10.43	0.30	965	.77	0.07
Pretest Reading Comprehension	-2.47	9.42	-0.26	965	.79	-0.05
Pretest Total Reading	0.50	10.01	0.05	965	.96	0.01

Program Description

Achieve3000 is a differentiated online literacy program that provides standards-based content for students in grades 2 through 12. Achieve3000 can be used for whole-group instruction, providing lessons and assignments to the entire class, and it can be customized to meet individual student needs. The program uses a five-step literacy routine to support: (1) learning from informational text, (2) acquisition of content knowledge, (3) use of higher-order thinking skills, (4) routine use of reading strategies, and (5) awareness that information contributes to opinions, which should be confirmed using evidence. Additionally, the program provides ongoing assessments designed to help teachers target instruction and it continually adjusts to meet individual student needs. Achieve3000 also has integrated reporting systems to provide timely diagnostic and achievement data for use by teachers and school administrators.

Achieve3000 and Comparison Program Implementation

As a condition of study participation, evaluators expected teachers to complete all of the study's data collection activities. Before the 2014/15 school year, Magnolia Consulting and Achieve3000 facilitated a study orientation for all teachers, as well as an Achieve3000 training for treatment teachers. Throughout the study period, treatment teachers were required to implement the Achieve3000 program with their students for at least 90 minutes per week. Ideally, Achieve3000 program implementation occurred in a computer lab or classroom with a 1:1 computer-to-student ratio. For this study, comparison teachers implemented their typical literacy program, but not Achieve3000. It is important to note that treatment teacher program implementation data included weekly online logs administered over 32 weeks, as well as student usage data, while comparison teacher implementation data consisted of a one-time online teacher survey. Evaluators observed treatment and comparison teachers' classrooms during a spring site visit at each school.

Achieve3000 Response Rates and Implementation Measures

Implementation measures for the treatment group included teacher-reported online weekly implementation logs, observation data collected by evaluators, and student usage data compiled by the Achieve3000 program.

Treatment teachers completed online weekly implementation logs comprised of questions about classroom use, perceptions, and student engagement in the program. These logs were based partly on the study's implementation guidelines, which evaluators and Achieve3000 staff developed collaboratively. Within these logs, teachers provided feedback on their experiences with the Achieve3000 program. As a group, the 23 participating treatment teachers completed a total of 740 weekly logs for an average of 32 weekly logs per teacher and an overall response rate of 100%.

Evaluators observed treatment teachers in the spring for 30–60 minutes using an observation checklist. The observation checklist, developed by Magnolia Consulting in collaboration with Achieve3000, included ratings for the following domains: teacher-student interactions, equipment and technology, procedures associated with the use of Achieve3000, Achieve3000 program components, and student engagement. Evaluators scored each indicator with a 4-point rating scale (0 = *Not at all—does not meet this indicator*, 1 = *Partially—apparent but on somewhat inconsistent basis*, 2 = *mostly—apparent but not fully consistent*, 3 = *Fully—fully meets indicator*). Two treatment teachers were absent during the scheduled observations, thus observations were available for 21 of the 23 treatment teachers (91.30%).

In addition to collecting implementation data from teachers, the Achieve3000 program tracked treatment students' use of the program with LevelSet data generated by the program. Student usage data included information about the number of logins, program hours, reading connections, writing assignment thought questions, and activities.

Achieve3000 Implementation Fidelity

The study's implementation guidelines defined minimum usage requirements for teachers, which were used to determine the denominator for calculating each teacher's implementation fidelity score. As discussed in the Measures section of this report, evaluators calculated implementation fidelity scores using data from three sources: teacher's weekly log reports, observations, and student usage reports. Each of these variables was equally weighted (33.33%) to calculate an overall implementation fidelity score.

Based on the weekly log data, teachers' implementation scores ranged from 36.19% to 96.10%, for an average score of 73.33%. Based on the observation scores, teachers' implementation scores ranged from 63.89% to 100.00%, for an average of 86.24%. Based on student usage data, teachers' implementation scores ranged from 23.00% to 93.00%, for an average of 57.67%. As shown in Figure 5, implementation levels varied by type of measure. This supports the importance of triangulating the findings from teacher self-reported logs, observations completed by evaluators, and average student usage data.

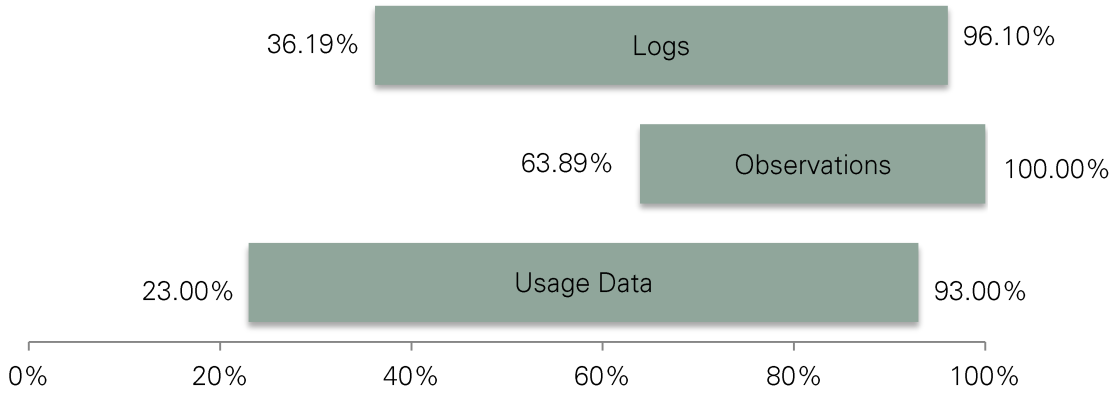
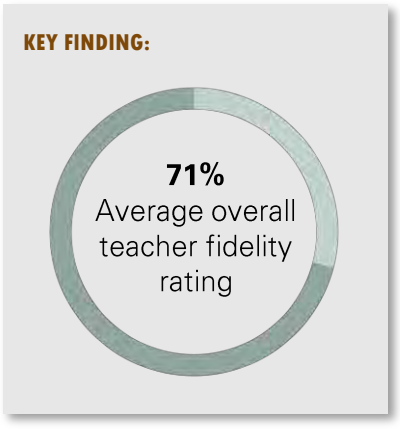


Figure 5. Implementation score ranges for weekly log data, observations and usage data.

Combining data from the teacher logs, observation, data and student usage reports, evaluators determined that the overall fidelity scores for the treatment teachers ranged from 34.99% to 96.25% for an average fidelity score of 71.05%. Given that a 100% fidelity rating represents prefect implementation, this overall fidelity score indicates that overall, teachers implemented the program 29% less than the minimum implementation guidelines prescribed by program developers. It is important to note that perfect implementation fidelity is very difficult for teachers to achieve in the real world due to competing district and state requirements, assessments, holidays, weather delays, technology issues, and other issues.



Achieve3000 Program Use, Planning, and Supplementation

KEY FINDING:

On average, treatment teachers reported that they used the Achieve3000 program for 1.86 days each week for a total of 88.43 minutes.

On the weekly logs, treatment teachers reported how often they used the Achieve3000 program and its components, as well as how much time they spent planning each week's lessons. On average, treatment teachers reported that they used the Achieve3000 program for 1.86 days each week for a total of 88.43 minutes and an average of 1.94 lessons each week. They indicated that they covered a featured lesson on 66.67% of their weekly logs and a bonus lesson on 7.63% of the logs. Additionally, on average, treatment teachers reported that they spent 25.64 minutes preparing and planning for their lessons each week.

The logs also provided an opportunity for treatment teachers to indicate whether or not they had supplemented the Achieve3000 program each week. On average, treatment teachers reported supplementing the Achieve3000 curriculum with additional materials on 12.58% of the weekly logs. Teachers described the supplemental materials as:

- reading books,
- self-made power points,
- videos (YouTube videos, DVDs, documentaries)
- Google images
- 3D strategies,
- short example paper,
- constructed responses,
- summarization strategies,
- using capitals and commas and textual evidence to support arguments,
- graphic organizers,
- literacy templates,
- teacher created curriculum,
- required district curriculum,
- nonfiction texts and articles (History, social studies, *Narrative of the Life of Frederick Douglass*, *USA Today*, *National Geographic*)
- online dictionary,
- fiction texts (*To Kill a Mockingbird, Animal Farm, Romeo and Juliet*),
- incentive programs,
- small groups,
- Prezi's,
- SOAPS strategy,
- argument writing,
- notebooks, and
- Rigby.

Treatment teachers were also asked to indicate if they needed to supplant any of their core curricula in order to implement the Achieve3000 program. On 8.84% of the weekly logs treatment teachers reported supplanting their core curriculum to implement the Achieve3000 program. Teachers reported supplanting the following core materials:

- vocabulary lesson,
- grammar lessons,
- ancient history curriculum,
- history curriculum,
- social studies curriculum,
- reading assignments
- nonfiction texts, articles and reading materials,
- fiction novels,
- literature (*To Kill a Mockingbird, Of Mice & Men, Narrative of the Life of Frederick Douglass*)
- argumentative writing,
- whole group reading,
- small group work,
- literacy centers,
- figurative language,
- reading strategies, and
- plot diagrams.

Implementation of Achieve3000 Components

In addition to asking teachers about their overall program use, the weekly logs also asked treatment teachers to report the degree to which their students used various Achieve3000 components for their independent work. On average, teachers most frequently reported that students used the "complete the activity questions" (92.39%), "read the article" (91.07%), and the "respond to the before the reading poll" (89.30%) (see Figure 6).

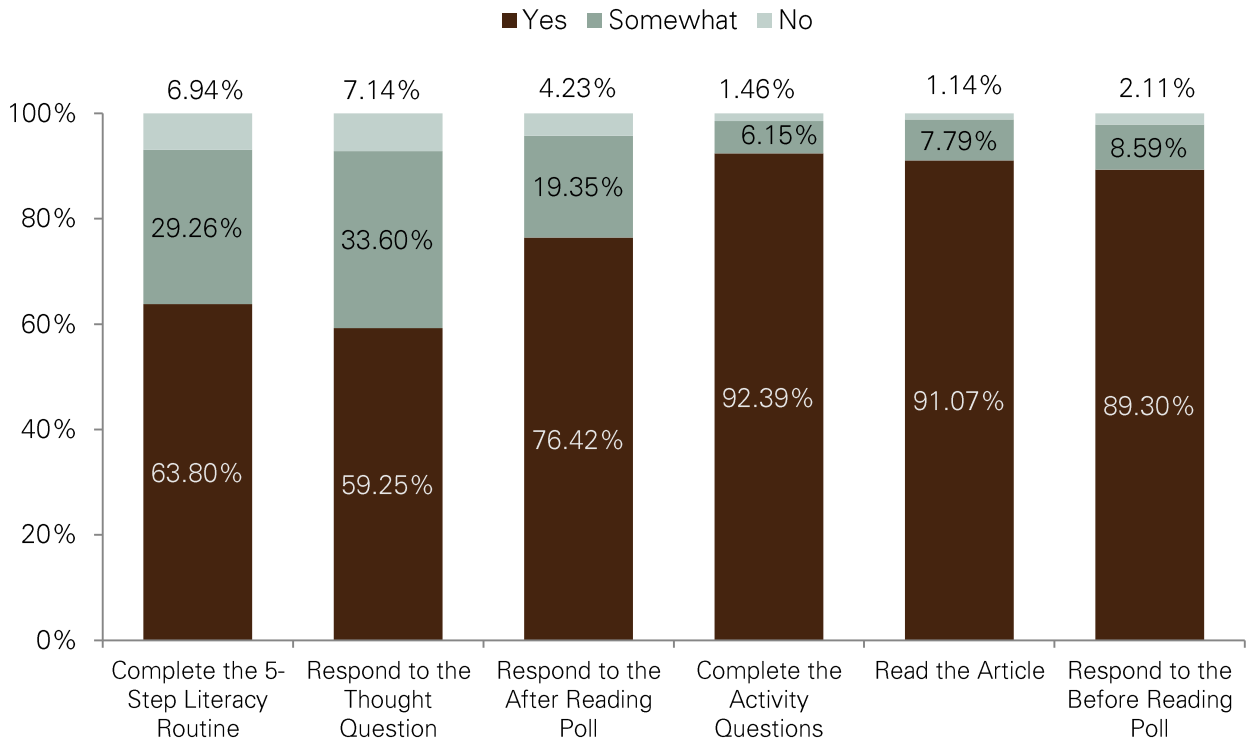


Figure 6. Average treatment teacher ratings of student use of Achieve3000 components for independent work.

Treatment teachers also described the extent to which their students participated in any additional Achieve3000 activities each week. Teachers most frequently reported that their students used or sometimes used the "poll results." On the majority of treatment teacher logs, teachers indicated that students did not use the "math," "stretch article," and "stretch activity" (see Figure 7).

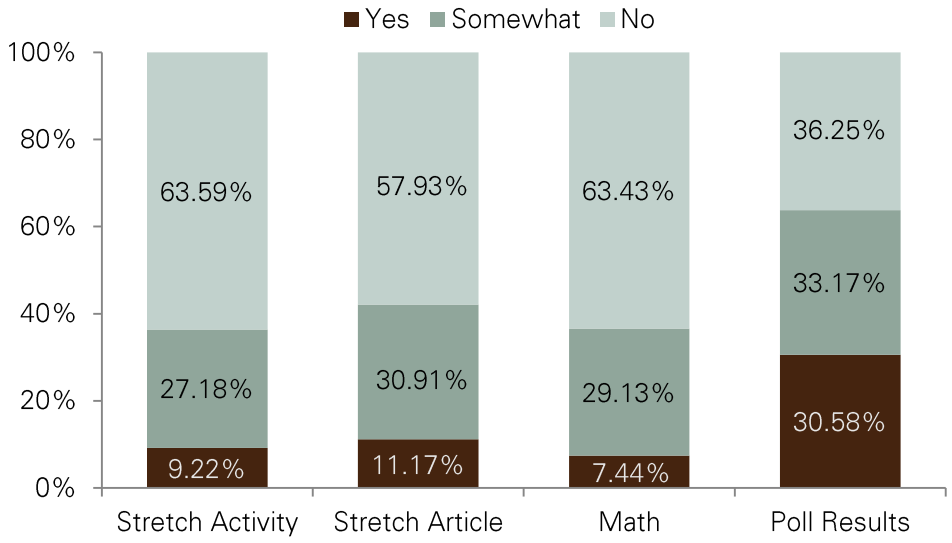


Figure 7. Average teacher ratings of student participation in additional Achieve3000 actitvities.

Use of Achieve3000 Administrative Components

On the weekly logs, treatment teachers also reported on how often they used various administrative components. For this question, teachers were asked to select one of four answers (*never, rarely, often,* or *always*). For "student work," on average most teachers reported that they *always* (27.51%) or *often* (36.08%) used the component. For "usage reports," and performance reports, most teachers reported that they *often* used the components. And finally, for "assessment tools" and "home communication" most teachers reported that they *never* used these components (see Table 6).

Table 6. Average Teacher Use of Achieve3000 Adminsitrative Components

Student Needs	Never	Rarely	Often	Always
Student Work	20.55%	15.86%	36.08%	27.51%
Usage Reports	22.98%	13.27%	42.88%	20.87%
Performance Reports	23.21%	12.99%	40.42%	23.38%
Assessment Tools	45.15%	22.49%	24.43%	7.93%
Home Communication	82.85%	12.94%	3.39%	0.81%

Use of Achieve3000 Teacher Materials

In addition to reporting on their use of administrative components, treatment teachers also rated how often they used Achieve3000 teacher materials using the same Likert scale (*never, rarely, often,* or *always*). On average, the majority of teachers reported *never* using "answer keys," "ELL & struggling readers supports," and "gifted and talented supports" on the weekly logs. For other components— "teacher recommendations," "discuss/review lesson vocabulary," "curriculum key," "graphic organizers," "standards," and "strategy lesson"— usage varied, but teachers most often indicated that they *never* used the materials. See Table 7 for average teacher use of Achieve3000 materials.

Table 7. Average Teacher Use of Achieve3000 Teacher Materials

Student Needs	Never	Rarely	Often	Always
Teacher Recommendations	48.06%	18.45%	21.52%	11.97%
Discuss/review Lesson Vocabulary	29.77%	11.97%	28.64%	29.61%
Answer Keys	51.38%	20.75%	21.07%	6.81%
Curriculum Key	42.46%	13.29%	29.01%	15.24%
Graphic Organizer	48.71%	21.36%	23.46%	6.47%
Standards	35.28%	18.44%	26.38%	19.90%
Strategy Lesson	42.39%	17.80%	29.61%	10.19%
ELL & Struggling Readers Supports	57.12%	17.64%	12.94%	12.29%
Gifted & Talented Supports	66.99%	15.86%	9.06%	8.09%

Classroom Instruction with Achieve3000

Treatment teachers also rated how often they completed various classroom instruction practices using the same scale (*never, rarely, often,* or *always*). Most frequently, teachers

reported that they *always* "actively supervised students as they worked independently at their own Lexile level" (48.78%), "responded to students' questions" (49.84%), and "helped students navigate the program" (35.58%). Additionally, teachers most often reported that they *often* (44.70%) or *always* (31.32%) "discussed what students did when working at their Lexile Level." Teachers' answers varied across the four scales for the following strategies to meet student needs: "modeled strategies of more rigorous text," "guided students as they worked on more rigorous text," and "discussed what worked as students worked with more rigorous text." Finally, teachers most often reported that they *never* "worked on other work" (55.90%) while students were using Achieve3000.

Table 8. Average Teacher Reports Of Classroom Instruction Practices While Students Were Working Through The Achieve3000 Program

Strategies to Meet Student Needs	Never	Rarely	Often	Always
Actively supervised students as they worked independently at their own Lexile level	2.60%	6.67%	41.95%	48.78%
Discussed what students did when working at their own Lexile Level	7.18%	16.80%	44.70%	31.32%
Modeled strategies on more rigorous text	18.46%	21.73%	34.48%	25.33%
Guided students as they worked on more rigorous text	16.37%	20.79%	35.52%	27.33%
Discussed what worked as students worked with more rigorous text	16.94%	24.51%	32.40%	26.15%
Responded to student questions	4.78%	10.86%	34.54%	49.84%
Helped students navigate the program	20.43%	21.91%	22.08%	35.58%
Worked on other work (such as lesson planning, grading papers, etc.)	55.90%	34.26%	6.07%	3.77%

Challenges with Achieve3000 Implementation

Treatment teachers were asked if they experienced any challenges or difficulties implementing Achieve3000 each week. On average, treatment teachers reported experiencing difficulties implementing Achieve3000 on 23.50% of the weekly logs.

CHALLENGES WITH ACHIEVE3000 IMPLEMENTATION

- Software and hardware issues with iPads and computers.
- Wifi access issues.
- Competing assessment administration schedules (LevelSet, GMRT, district and state assessments).
- Not having enough time to implement the program.
- Subject matter/Lexile levels too hard for students.
- Poor student engagement.
- Program glitches (issues with thought questions, logging students out, not being able to log in, screen freezing when reading out loud, tabs missing for before and after reading polls).
- School activities (parent-teacher conferences, field trips, professional development days).
- Sick days.
- Holidays and weather delays.
- Insufficient training.

Student Achieve3000 Online Program Usage

KEY FINDING:

On average, students in this study logged into the Achieve300 program 101 times during the year and logged 30.53 program hours.

Treatment students' online use of the Achieve3000 program was tracked by the program. On average, students in this study logged into the Achieve3000 program 101 times during the year and logged 30.53 program hours. Treatment students averaged 50.58 valid activities during the year and 1.49 activities each week. Students averaged 30.01 passing activities during the year. Passing activities were activities in which a student answered 75% or more of the questions in the activity correctly. Achieve3000 uses this threshold as a measure for determining whether students are applying themselves to the activity and working within their instructional zone. See Table 9 for complete student usage data results.

Table 9. Treatment Student Online Program Usage Descriptives ($N = 512$)

	N	Mean	*SD*	Min	Max	Median
Total Logins	512	101.07	64.62	20	421	89
Program Hours	512	30.53	16.62	7.0	149.8	29.6
Reading Connections: Summarization	512	8.78	12.88	0	102	4
Reading Connections: Generate Questions	512	3.86	7.21	0	57	1
Reading Connections: Setting the Purpose	512	7.82	18.34	0	140	1
Writing Assignment Thought Questions	512	39.12	28.19	2	239	32
Activities	512	53.54	29.84	4	233	50
Invalid Activities	512	2.96	7.76	0	93	1
Total Valid Activities	512	50.58	26.97	4	193	48
Average Weekly Activities	512	1.49	0.80	0.1	6.0	1.40
Passing Activities	512	30.01	20.07	0	122	26

Achieve3000 Observations

As reported earlier, observation checklists were completed by evaluators in the spring of 2015 for 21 of the 23 treatment teachers. For this study, the Achieve3000 program was administered with grades 3, 6 and 9 using multiple forms of available technology, class-times, and settings, resulting in a variety of implementation models. Teachers were given a total observation score, which ranged from 63.89% to 100.00%, for an average of 86.24%.

On average, treatment teachers *mostly* or *fully* met each of the observation indicators. However, on average, teachers did not fully meet some of the indicators for implementation including "supporting struggling or gifted and talented readers," "discussing or reviewing lesson vocabulary," and "whole group instruction." Average scores for each of these indicators

ranged between *partially* or *mostly* meeting the indicators, with some teachers not implementing these indicators during the observations.

Program implementation varied across grades with some teachers using all program components and a very hands-on approach to the lesson, while other teachers asked students to use the program independently and used very little whole group instructional time. To illustrate the types of program implementation evaluators observed during this study, three vignettes have been developed showing examples of (1) independent implementation, (2) whole group implementation, and (3) small group implementation. It is important to note that these vignettes represent examples and do not reflect the implementation of every teacher observed.

Vignette #1: Sample Independent Implementation

Mr. Swanson* teaches a high school English language class. The students are in Grade 9 and the class period is 57 minutes long. Students enter class quietly and are ready to follow the routines established for daily instruction. As they enter, Mr. Swanson instructs them to gather a Chrome Book from the computer cart to implement an Achieve3000 lesson. The students quickly get settled at their desks and start up their computers.

Students log onto the Achieve3000 program and can choose which article they would like to complete. Mr. Swanson provides very little instruction for students to log on and select an article. Students complete various lessons and follow the 5-step routine. Mr. Swanson walks around the room multiple times during the class period and makes sure students are on task. The room is silent for the majority of the lesson. During the lesson Mr. Swanson helps individual students as needed and hands back graded homework (non-Achieve related). He does not engage students in a vocabulary lesson or whole group instruction. As students complete their Achieve3000 lesson they return their Chrome Books to the computer cart and silently complete other work at their desks.

In a conversation with Mr. Swanson he comments that he has the students use the program independently and they do not use the whole group instruction because students choose their own articles.

*Teachers' names have been changed to protect confidentiality

Vignette #2: Sample Whole Group Implementation

Ms. Burton* teaches a sixth-grade English language arts class. Students enter class and are ready to follow the routines established for daily instruction. As they enter, each student grabs an iPad and logs onto the Achieve3000 program. Ms. Burton tells the students which lesson they will be completing and walks them through logging onto the program, the thought question, and the vocabulary. All of the students eagerly put their hands up to answer the questions. Students complete the poll responses and discuss their responses as a class. The teacher provides some examples of "scams" when talking through the topic of the article and reminds students about strategies for comprehending text and citing textual evidence.

Students launch into the 5-step literacy routine. Ms. Burton walks around the room to help and probe individual students with guiding questions as necessary. The students are interested and engaged for the majority of the lesson. Ms. Burton helped students figure out the meaning of vocabulary without giving them the answers. During the lesson Ms. Burton hands back completed worksheets from a previous lesson and gives students a piece of candy if the students complete the daily Achieve activity.

Ms. Burton leads a whole group discussion with students and asks students to share their responses to the questions. She asks questions and calls on a variety of students. Ms. Burton is very encouraging and provides a lot of positive reinforcement throughout the lesson.
*Teachers' names have been changed to protect confidentiality

Vignette #3: Sample Small Group Implementation

Ms. Little* teaches a third-grade class. Students are seated in groups while Ms. Little discusses each of the 4 stations they will be completing in small groups (independent reading, Achieve3000, vocabulary lessons with Ms. Little, and worksheets). She explains each station and then asks students to go to their first station. Some students forget where they are going or are off task and Ms. Little has to remind them about where they should be and what they should be doing.

At the Achieve3000 station, students all complete the same Achieve3000 lesson. The students complete the 5-step literacy routine independently. Some students have issues with logging on to their computers or with their computer screens freezing, and they have to ask Ms. Little for help. The teacher has to get up from the small group she is leading to help them. Some students are off task and are sharing the questions and answers with their neighbor, clicking through the questions without taking time to consider the answers, or getting up out of their seats and walking around the room. Some students are on task and complete the 5-step routine.

In a conversation with Ms. Little, she said she struggles with student engagement in her class in general and has a few students who act out on a daily basis. She said they typically take two days to rotate through small groups, and then they have a whole group discussion.
*Teachers' names have been changed to protect confidentiality

Comparison Teachers' Implementation of Their Literacy Programs

Comparison teachers were asked a series of implementation questions on the spring comparison teacher online survey (single administration) about their comparison programs. This

section describes comparison teachers' implementation of the various literacy programs used in their classrooms.

Comparison Teachers' Implementation

On average, comparison teachers reported using their core materials 4.48 days per week for 54.65 minutes each day. Comparison teachers reported spending 129.78 minutes on average each week planning and preparing to teach their core literacy lessons. Comparison teachers reported using a wide variety of formal core literacy programs including: Rigby, Houghton Mifflin Harcourt: Literacy By Design, Reading A-Z, www.readworks.org, McDougal Littell Language of Literature, Scholastic Story Works and district-created ELA units. Comparison teachers also reported using the following informal materials to teach their core literacy instruction: fiction and nonfiction reading (i.e. novels, media articles, textbook reading), grammar practices, handouts, technology (i.e. iPad programs), films, art and pictures, and vocabulary lessons.

The majority of comparison teachers (91.30%) reported supplementing their core literacy materials with additional materials an average of 2.62 days per week. Comparison teachers' formal supplemental materials included: Houghton Mifflin Common Writing Book, EPIC, Reading A-Z, Teachers Pay Teachers, Moby Max, Time for Kids, *National Geographic* articles, Reading Minute, Making Connections, comprehension toolkit by Harvey and Goudvis, *Scholastic Scope* magazine, and Leveled Literacy Intervention (LLI). Comparison teachers also reported using various nonformal and teacher-created materials such as: printed materials or handouts from professional learning communities or other teachers, various texts (i.e. nonfiction articles, trade books for differentiated reading, district unit books, fiction and nonfiction passages, novels, plays, short stories, and essays), reading comprehension skills and strategies, computer based activities (i.e.,iPad lessons), graphic organizers, notes, science and social studies materials, skill-based activities, and educational videos or tutorials (YouTube).

Comparison teachers were also asked how often they assessed students formally or informally in literacy. The frequency of teacher assessments varied with most teachers reporting assessing students *a few times a week* (30.43%) or *once a week* (26.09%) followed by *daily* (26.09%). Comparison teachers reported moving students to higher or lower literacy groups, and working with students on target skills based on student assessment results. Comparison teachers reported using the following formal literacy assessments: Aimsweb, Accelerated Reader Assessments, Direct Reading Assessment (DRA) fluency, Galileo STAR reading, Response to Literature (RTL) responses, McGraw Hill Early Reading Intervention (ERI), Northwest Evaluation Association (NWEA), Fountas and Pinnell: Running Record, www.readtheory.org, Mastery Connect, Scholastic Reading Inventory (SRI) tests, benchmark assessments, district assessments, and unit assessments. In addition to these formal assessments, comparison teachers also reported using the following informal assessments: vocabulary quizzes, group projects or presentations, grade level comprehension and vocabulary theme skills tests, conferencing, one-on-one assessments, informal classroom assessments (pulling sticks, tickets at the door), world maps, reading comprehension packets, iPad games/quizzes, teacher created assessments, short responses, journal assignments, blog responses, written, verbal and observation.

In the spring of 2015, evaluators completed observation checklists for all 23 comparison teachers. A few comparison teachers had issues with student behavior and engagement (students were off task and appeared bored) or the teacher did not have materials prepared and ready for the lesson. Overall, comparison teachers used a variety of instructional strategies and techniques to implement their lessons, and students were mostly on task and engaged in the lessons. The majority of comparison teachers blended together formal core literacy materials and teacher created curriculum and activities. Most comparison teachers were very positive and provided a lot of scaffolding, modeling, and opportunities for student discussion.

Summary

KEY FINDING:

Comparison teachers reported using various core literacy programs for more days per week than treatment teachers reported using Achieve3000. Comparison teachers reported planning and preparing for a longer period of time than treatment teachers and reported using more supplemental materials.

Overall, treatment teachers implemented the Achieve3000 program with a combined average fidelity score of 71%. On the majority of the weekly logs teachers reported logging onto the program twice a week for the appropriate amount of time. Some treatment teachers reported supplanting their core curriculum in order to implement Achieve3000 and reported removing various literacy and social studies lessons, readings, small group work, and activities.

Treatment teachers also reported their use of various Achieve3000 program components and described any challenges with program implementation. On the majority of the weekly logs, teachers reported completing or "somewhat completing" all program components, but did not consistently use the additional activities, the administration components, or teacher materials. Program implementation varied across grades and classrooms, and evaluators observed multiple implementation models including whole group instruction, independent activities, and small groups. Overall, teachers reported having some challenges with implementation such as: not having enough time to implement the program, low student engagement, various class delays, competing standardized testing schedules, issues with computer and iPad software/hardware, and various program glitches. These challenges may have impacted teachers' implementation of the program.

Comparison teachers were asked on a one-time survey to describe their program implementation. Comparison teachers reported using various core literacy programs for more days per week than treatment teachers reported using Achieve3000. Comparison teachers reported planning and preparing for a longer period of time than treatment teachers and reported using more supplemental materials. Observations of comparison teachers revealed that most comparison teachers used a blend of formal core literacy program materials and teacher-created materials and activities. In most comparison classrooms students were engaged in the lessons and on-task.

Findings Regarding Student Learning

This portion of the report addresses the evaluation results regarding student learning in reading. It begins with a description of GMRT-4 and LevelSet outcomes among students in the treatment group (i.e., students who participated in Achieve3000). Then it describes findings comparing GMRT-4 outcomes among treatment-group students and comparison-group students who participated in their schools' usual literacy programs.

Reading Achievement among Treatment-Group Students (Achieve3000 Users)

KEY FINDING:

Students in the treatment condition who used Achieve3000 during the 2014/15 school year demonstrated substantively important and statistically significant gains on the GMRT-4 and LevelSet assessments.

As noted earlier in this report, evaluators conducted descriptive and multilevel modeling analyses to determine if students who used Achieve3000 during the 2014/15 school year demonstrated gains in reading achievement over the course of the study. Evaluators also ran exploratory analyses to examine Achieve3000's LevelSet assessment data. This section provides a description of findings from these analyses.

Descriptive Findings for Treatment-Group Students

Before running multilevel modeling analyses to measure reading gains among treatment students, evaluators examined descriptive statistics for the GMRT-4, which teachers administered as a pretest at the beginning of the school year and as a posttest at the end of the school year. Figures 8–10 display the grade equivalent scores corresponding to each grade's GMRT-4 mean scale score for the Vocabulary, Reading Comprehension, and Total Reading tests. Examining these scores visually shows that within each grade, students in the treatment group began the school year scoring below grade level. Throughout the study period, grade level equivalent scores increased. Furthermore, within each grade and test, the average increase corresponded to greater than or equal to the amount that a typical student would be expected to grow during a nine-month school period. Thus, by the end of the school year, students who had used Achieve3000 were generally closer to scoring on grade level than they had been at the beginning of the school year, before they had used the program.

LEARNING GAINS

By the end of the school year, students who had used Achieve3000 were generally closer to scoring on grade level than they had been at the beginning of the school year, before they had used the program.

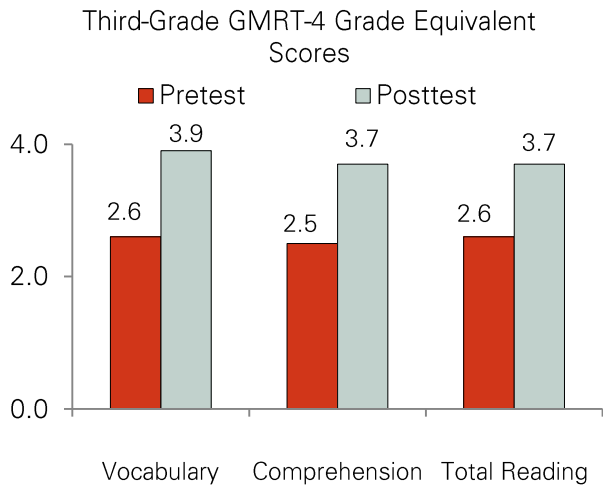


Figure 8. Pretest and posttest GMRT-4 grade equivalent scores for third-grade treatment students.

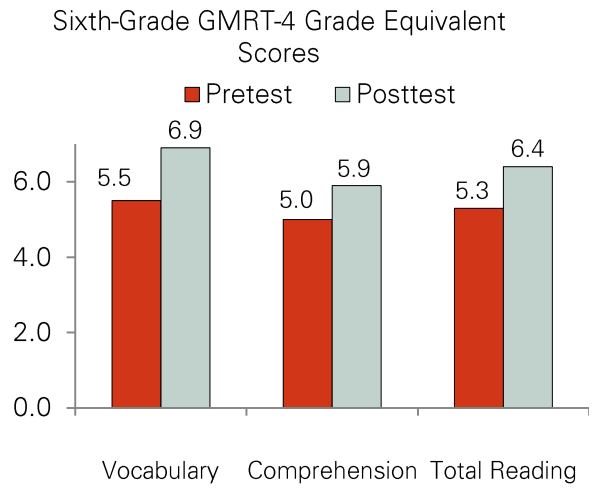


Figure 9. Pretest and posttest GMRT-4 grade equivalent scores for sixth-grade treatment students.

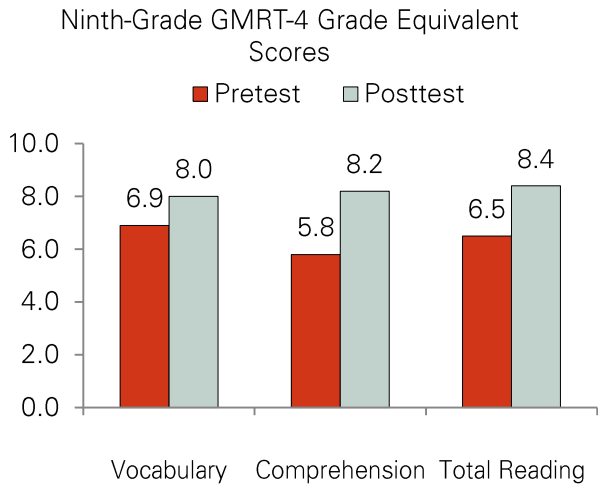


Figure 10. Pretest and posttest GMRT-4 grade equivalent scores for ninth-grade treatment students.

Additionally, evaluators examined descriptive statistics for treatment students' pretest and posttest LevelSet Lexile levels (see Figure 11). Visual examination of the raw means suggests that third-grade, sixth-grade and ninth-grade treatment students who used Achieve3000 demonstrated increases of 170, 120, and 33 in their Lexile levels, respectively. The increases for third- and sixth-grade students exceeded the expected Lexile gains for average students established by MetaMetrics, which correspond to 100 for an average third-grade

ACHIEVE3000 STUDENTS' LEXILE GAINS

- The average third-grade Lexile gain was 70% greater than the expected gain for an average Grade 3 student.
- The average sixth-grade Lexile gain was 71% greater than the expected gain for an average Grade 6 student.
- The average ninth-grade Lexile gain was 34% smaller than the expected gain for an average Grade 9 student.

student and 70 for an average sixth-grade student[5]. The increase for ninth-grade did not exceed the expected Lexile gain for an average ninth-grade student established by MetaMetrics, which corresponds to 50 for an average ninth-grade student. Overall, treatment students made gains, on average, in their Lexile levels from pretest to posttest, and they became more likely as a group to be classified as *on* or *above* the LevelSet assessment college and career readiness benchmark by the end of the study. However, their average end-of-year Lexile levels did not correspond to levels associated with being on track for college and career readiness based on the Achieve3000 LevelSet assessment benchmarks for college and career readiness (Achieve3000, 2011).

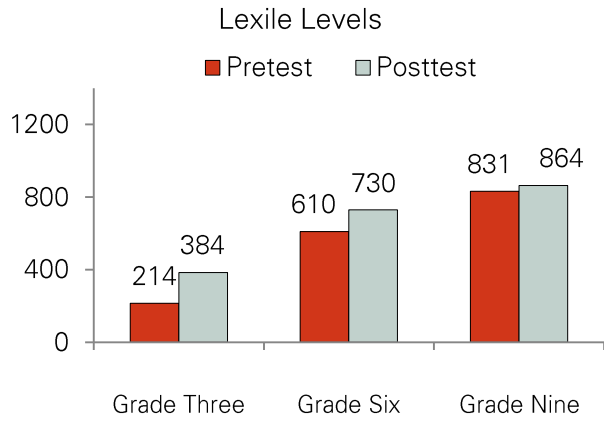


Figure 11. Pretest and posttest LevelSet Lexile levels for third, sixth, and ninth-grade treatment students.

Multilevel Modeling Analyses Examining Treatment Students' Pretest to Posttest Reading Gains

After examining descriptive statistics regarding treatment students' reading gains from pretest and posttest, evaluators used multilevel modeling analyses to determine if the learning gains were statistically significant. In each analysis, the outcome variable was the gain score corresponding to the GMRT-4 subtest of interest (i.e., Vocabulary, Reading Comprehension, or Total Reading) or the LevelSet outcome of interest (i.e., Lexile level), and each model accounted for the clustering of students in teachers' classrooms. Additionally, evaluators calculated standardized effect sizes by dividing each pretest-to-posttest gain by the corresponding pretest standard deviation.

[5] In this context, an average student is a student scoring at the 50th percentile.

KEY FINDING

Achieve3000 users' Vocabulary, Reading Comprehension, and Total Reading gains corresponded to effect sizes of 0.43, 0.47, and 0.48.

Table 10 displays the findings from the multilevel modeling analyses conducted on the GMRT-4. On average, treatment students who participated in Achieve3000 during the 2014/15 school year demonstrated statistically significant and substantively important learning gains based on the What Works Clearinghouse standards (WWC, 2014) threshold of 0.25. More specifically, average Vocabulary, Reading Comprehension, and Total Reading gains corresponded to effect sizes of 0.43, 0.47, and 0.48, respectively.

Table 10. Treatment Students' GMRT-4 Pretest to Posttest GMRT-4 Gains

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. *df*	*p*-value	Effect Size
Vocabulary	19.82	3.32	5.97	22	<.001*	0.43**
Reading Comprehension	21.24	4.10	5.18	22	<.001*	0.47**
Total Reading	20.24	3.11	6.51	22	<.001*	0.48**

* Statistically significant at the 0.05 level.
**Substantively important based on the WWC Standards.

Table 11 displays findings from multilevel modeling analyses conducted on the LevelSet Lexile levels. Analyses conducted across grades revealed that the average gain was statistically significant, and the positive effect size of 0.33 was substantively important. Thus, on average, treatment students who participated in Achieve3000 demonstrated statistically significant, substantively important gains in their Lexile levels over the study period (WWC, 2013).

KEY FINDING

Achieve3000 users' Lexile gains corresponded to an effect size of 0.33

Table 11. Treatment Students' Pretest to Posttest LevelSet Lexile Gains

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. df	*p*-value	Effect Size
Lexile level	109.62	15.96	6.87	22	<.001*	0.33**

* Statistically significant at the 0.05 level.
**Substantively important based on the WWC Standards.

EXPECTED GROWTH

On average, over half of the students who used Achieve3000 during the 2014/15 school year met or exceeded expected Lexile level growth.

Exploratory Analyses of LevelSet Data

In addition to yielding data regarding student reading levels and student Lexile levels, the LevelSet provided data regarding: (a) the degree to which treatment students' reading skills grew as expected over the course of the study; (b) whether treatment students' classifications regarding college and career readiness changed over the course of the

study; (c) the degree to which treatment students participated in and passed Achieve3000 multiple choice activities; and (d) afterschool program use. Evaluators conducted exploratory analyses in each of these four areas.

Expected Growth

Evaluators examined treatment students' reading growth over the study year. The Achieve3000 program generated an expected Lexile growth based on students' pretest Lexile levels.[6] Evaluators compared students' expected Lexile growth and actual Lexile growth to determine if growth exceeded expectations. By the end of the study year, over half of the treatment students (57.87%) met or exceeded their expected growth levels, and on average, these students' actual Lexile growth was 34.78 points higher than their expected growth. A paired-samples *t*-test (see Table 12) showed that the difference between expected and actual growth was statistically significant.

Table 12. Descriptive Statistics Regarding Treatment Students' Expected and Actual Lexile Growth

	N[7]	Mean	*SD*	Min	Max	*t* value	*p* value
Expected Lexile Growth	508	642.29	286.00	-299	1387	6.404	<.01*
Actual Lexile Growth	508	677.07	300.51	-113	1581		

* Statistically significant at the 0.05 level.

COLLEGE AND CAREER READINESS

Achieve3000 users were more likely to be classified as *on* or *above* the LevelSet college and career readiness benchmark by the end of the study (23.44%) than they were at the beginning of the study (10.94%).

College and Career Readiness

Next, evaluators examined the change in treatment students' college and career readiness from pretest to posttest. On the LevelSet, the Lexile levels were classified as *far below, below, on* or *above* in regards to students' college and career readiness. The percentage of treatment students in each classification at each timepoint is shown in Table 13. Using a McNemar's Test, evaluators determined that over the course of the year, there was a statistically significant change in the proportion of students meeting the benchmarks. More specifically, the percentage of participants *on* or *above* benchmark after participating in Achieve3000 (23.44%) was statistically significantly higher than the proportion of students at the start of the program (10.94%). In other words, treatment students were more likely to be classified as *on* or *above* the LevelSet college and career readiness benchmarks at the end of the study than at the beginning.

[6] The Achieve3000 program calculated expected Lexile growth using each student's initial reading level and the number of days the student used the Achieve3000 program.

[7] Expected Lexile scores were determined based on pretest Lexile scores. Four students had imputed pretest values and thus, their expected Lexile score data was no longer valid and they could not be included in analyses.

Table 13. Treatment Students' College and Career Readiness Levels by Time Point

College and Career Readiness	Pretest (*N* = 512) *n*	Pretest (*N* = 512) Percent	Posttest (*N* = 512) *n*	Posttest (*N* = 512) Percent	McNemar's Test Results[8] Exact. Sig. (2-sided)
On or Above	56	10.94%	120	23.44%	<.01*
Above	12	2.34%	32	6.25%	
On	44	8.59%	88	17.19%	
Below or Far Below	456	89.06%	392	76.56%	
Below	228	44.53%	249	48.63%	
Far Below	228	44.53%	143	27.93%	

* Statistically significant at the 0.05 level.

Treatment Students' Completion of Achieve3000 Activities and Reading Gains

To address whether or not there was a relationship between completion of Achieve3000 activities and reading gains, evaluators used multilevel modeling to examine the relationship between the valid number of Achieve3000 activities that students had completed and their GMRT-4 Reading Vocabulary, Reading Comprehension, Total Reading, and LevelSet Lexile Gains. These analyses included a variable indicating whether students had completed a low number (i.e., 1-39), moderate number (40-79), or high number (i.e., 80 or more) of Achieve3000 activities throughout the study. As shown in Table 14, students who completed a moderate or high number of activities had average GMRT-4 Vocabulary, Reading Comprehension, and Total Reading gains that were approximately 3-4 points higher than those of students who completed a low number of activities. However, these differences were not statistically significant. For the Levelset Lexile gains, students who completed a moderate or high number of activities had Lexile gains that were statistically significantly greater than students who completed a low level of activities. Additionally, students who completed a high number of activities had Lexile gains that were statistically significantly higher than students who completed a moderate number of activities. Thus, on average, students who completed greater numbers of Achieve3000 activities during the study tended to have higher Lexile gains than students who completed relatively a lower number of Achieve3000 activities.

COMPLETION OF ACHIEVE3000 ACTIVITIES AND LEARNING GAINS

There was a statistically significant relationship between completion of Achieve3000 activities and Lexile level gains but not between activity completion and GMRT-4 Vocabulary, Reading Comprehension, or Total Reading gains.

[8] Due to the nature of McNemar's test, the four benchmark categories were combined at both time points to create two categories, one for students *on* or *above* benchmark and one category for students *below* or *far below* benchmark.

Table 14. Relationship Between Completion of Achieve3000 Activities and Reading Gains

Completion of a Moderate Number (40-79) versus a Low Number (1-39) of Activities					
Outcome	Coefficient	Standard Error	*t*-value	Approx. *df*	*p*-value
GMRT-4 Vocabulary	3.85	2.89	1.33	487	.18
GMRT-4 Reading Comprehension	3.94	4.40	0.90	487	.37
GMRT-4 Total Reading	3.66	2.68	1.36	487	.17
LevelSet Lexile Gain	42.21	14.97	2.82	487	.01*
Completion of a High Number (80+) versus a Low Number (1-39) of Activities					
Outcome	Coefficient	Standard Error	*t*-value	Approx. *df*	*p*-value
GMRT-4 Vocabulary	3.03	4.25	0.71	487	.48
GMRT-4 Reading Comprehension	3.96	6.47	0.61	487	.54
GMRT-4 Total Reading	3.11	3.95	0.79	487	.43
LevelSet Lexile Gain	83.16	22.02	3.78	487	<0.001*

* Statistically significant at the 0.05 level.

Performance on Achieve3000 Activities and Reading Gains

PERFORMANCE ON ACHIEVE3000 ACTIVITIES AND LEARNING GAINS

Overall, students with better performance on Achieve3000 activities had greater GMRT-4 Reading Comprehension, Total Reading, and Lexile gains, but there were no differences in GMRT-4 Vocabulary gains based on Achieve3000 performance.

To determine whether or not there was a relationship between performance on Achieve3000 activities and learning gains, evaluators used multilevel modeling analyses to determine if reading gains differed, on average, between two groups of treatment students: (1) those who gave correct answers to 75% or more of the Achieve3000 multiple choice questions they completed during the course of the study, and (2) those who did not answer correctly to 75% or more of the multiple choice questions they completed. Thirty percent (155) of the 512 treatment-group students were in the first group, with an average score of 75% or more correct. As shown in Table 15, multilevel modeling analyses revealed that on average, there were no statistically significant differences between the two groups in GMRT-4 Vocabulary gains. However, treatment students who averaged 75% or more correct answers had statistically significantly higher pretest-to-posttest GMRT-4 Reading Comprehension, Total Reading, and LevelSet Lexile gains than students who did not average 75% or more correct.

Table 15. Relationship between Treatment Students' Performance on Achieve3000 Multiple Choice Activities and Pretest-to-Posttest Reading Gains

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. *df*	*p*-value
GMRT-4 Vocabulary	3.14	2.58	1.22	488	.22
GMRT-4 Reading Comprehension	8.64	3.96	2.18	488	.03*
GMRT-4 Total Reading	6.37	2.38	2.67	488	.008*
LevelSet Reading Lexile Gains	117.86	12.85	9.17	488	<.001*

* Statistically significant at the 0.05 level.

Afterschool Users versus Non-Afterschool Users

Evaluators also examined treatment-group students' afterschool use of Achieve3000. About three-fourths of the treatment students (382) used the Achieve3000 program outside of school at least once during the study, and these students were considered afterschool users. On average, these 382 afterschool treatment students logged into the program afterschool 11.01 times. The other 130 treatment students never accessed the program outside of school.

AFTERSCHOOL USE AND LEARNING GAINS

On average, there was no statistically significant relationship between afterschool use of Achieve3000 and learning gains.

Using multilevel modeling, evaluators examined the relationship between the number of times students logged in after school and their pretest-to-posttest reading gains. Table 16 shows that none of these relationships was statistically significant. Thus, findings suggest no relationship between the number of times Achieve3000 users logged into the program after school and reading gains during the study period.

Table 16. Relationship between Treatment Students' Afterschool Use of Achieve3000 and Pretest-to-Posttest Reading Gains

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. df	*p*-value
GMRT-4 Vocabulary	0.01	0.07	0.10	488	.93
GMRT-4 Reading Comprehension	-0.09	0.11	-0.75	488	.45
GMRT-4 Total Reading	-0.05	0.07	-0.70	488	.49
LevelSet Reading Lexile Gains	0.72	0.39	1.83	488	.07

* Statistically significant at the 0.05 level.

Relationship between Teacher Implementation Fidelity of Achieve3000 and Student Learning Gains

KEY FINDINGS:

Relationships between implementation fidelity and learning gains were positive but not statistically significant for GMRT-4 Vocabulary, Reading Comprehension, or Total Reading gains. The positive relationship between implementation fidelity and Lexile gains was statistically significant.

As noted in the Implementation section of this report, teachers varied in the degree to which they implemented Achieve3000 with fidelity to study guidelines. To determine if there was a statistically significant relationship between implementation fidelity and student learning gains, evaluators ran a series of multilevel modeling analyses. In each of these analyses, the gain score of interest served as the outcome variable. Each model also accounted for the clustering of students in teachers' classrooms. Table 17 displays the results and shows no statistically significant relationships between implementation fidelity and GMRT-4 Vocabulary, Reading Comprehension, or Total Reading gains. On average, within the range of

implementation fidelity represented by this sample (i.e., 35.99% to 96.25% out of a possible 100%), an implementation fidelity increase of 10% corresponded to an increase of 0.90, 4.01, and 2.42 scale score points on the GMRT-4 Vocabulary, Reading Comprehension, and Total Reading tests. The positive relationship between implementation fidelity and LevelSet Reading Lexile gains was statistically significant, with an implementation fidelity increase of 10% corresponding to average gains of 31.01 Lexile levels.

Table 17. Relationship between Treatment Teachers' Program Implementation Fidelity and Student Learning Gains

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. *df*	*p*-value
GMRT-4 Vocabulary	9.43	20.99	0.45	21	.66
GMRT-4 Reading Comprehension	40.05	24.53	1.63	21	.12
GMRT-4 Total Reading	24.23	19.01	1.28	21	.22
LevelSet Reading Lexile Gains	310.65	76.61	4.06	21	<.001*

* Statistically significant at the 0.05 level.

Summary of Findings for Students Who Used Achieve3000

Together, treatment-group findings indicate that as a group, students who used Achieve3000 demonstrated substantively important and statistically significant gains on the GMRT-4 and LevelSet assessments (see Table 18).

Table 18. Summary of Main Findings for Acieve3000 Users

Assessment	Effect Size	Statistically Significant Impact	Substantively Important Effect Size*
GMRT-4 Vocabulary Gain	0.43	◆	◆
GMRT-4 Reading Comprehension Gain	0.47	◆	◆
GMRT-4 Total Reading Gain	0.48	◆	◆
LevelSet Lexile Gain	0.33	◆	◆

*Substantively important based on the WWC Standards.

Additionally, exploratory analyses showed that over half of the Achieve3000 users met or exceeded their expected Lexile level growth. During the course of the school year, students who used Achieve3000 were more likely to be classified as *on* or *above* the LevelSet college and career readiness benchmark compared to their beginning-of-year classifications. Although treatment students' performance on Achieve3000 activities was positively related to their learning gains on the GMRT-4 Reading Comprehension, GMRT-4 Total Reading and LevelSet Lexile levels, their afterschool use of the program was not statistically significantly associated with learning gains.

There was variability in the extent to which treatment teachers implemented Achieve3000 with fidelity. Analyses examining the relationship between implementation and learning gains showed that the positive relationships were not statistically significant for

implementation fidelity and GMRT-4 Vocabulary, Reading Comprehension, or Total Reading gains. However, there was a statistically significant positive relationship between implementation fidelity and student Lexile level gains. Thus, on average, teachers who implemented Achieve3000 with higher fidelity tended to have students who made greater Lexile level gains compared to teachers who implemented the program with relatively lower fidelity.

Analyses of Students' Reading Achievement by Treatment and Comparison Groups

KEY FINDINGS:

Achieve3000 had a statistically significant impact on posttest GMRT-4 Reading Comprehension and Total Reading scores when compared to typical ELA programs. Exploratory analyses suggest that program impacts varied by grade, with greatest impacts (substantively important impacts) on Vocabulary, Reading Comprehension, and Total Reading evident among ninth-grade study participants.

Evaluators conducted descriptive analyses and multilevel modeling analyses to compare reading achievement among treatment group students who used Achieve3000 and comparison group students who used their school's typical ELA program. Evaluators also ran descriptive multilevel modeling analyses to examine the impact of Achieve3000 on reading achievement separately within third, sixth, and ninth grades. When appropriate, evaluators calculated effect sizes and WWC improvement indices (WWC 2014) to help readers interpret the magnitude of program impacts. This portion of the report describes the results of these analyses.

Descriptive Findings Comparing Reading Achievement by Study Condition

Before running multilevel modeling analyses, evaluators calculated means corresponding to pretest and posttest GMRT-4 scores by study condition. Examining the means visually (see Figures 12-14) revealed that on average, students in the treatment group gained an average of 3, 12, and 7 points more than students in the comparison group in the Vocabulary, Reading Comprehension, and Total Reading tests, respectively. Readers should note that evaluators calculated these means for descriptive purposes rather than to determine if differences in reading performance by study condition were statistically significant.

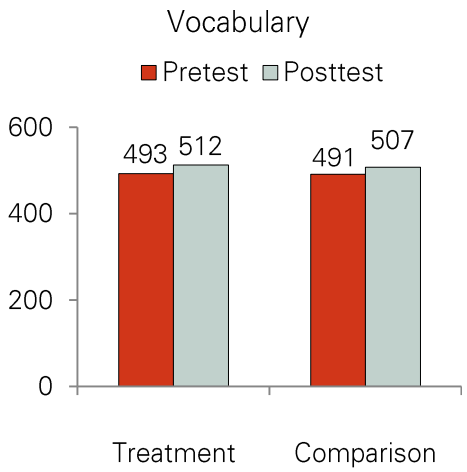


Figure 12. GMRT-4 unadjusted Vocabulary scale score means across grades by study condition and time.

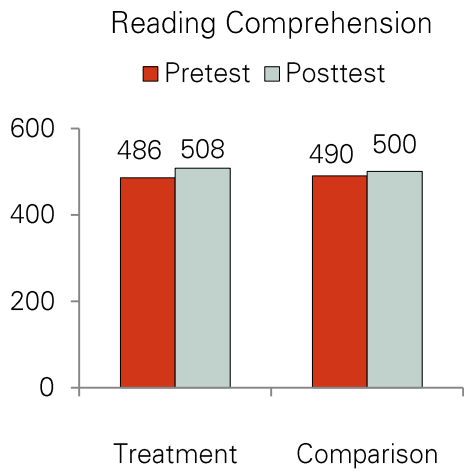


Figure 13. GMRT-4 unadjusted Reading Comprehension scale score means across grades by study condition and time.

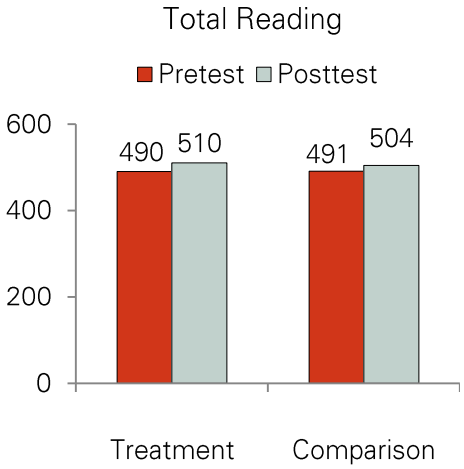


Figure 14. GMRT-4 unadjusted Total Reading scale score means across grades by study condition and time.

Multilevel Modeling Analyses Comparing Reading Achievement by Study Condition

After visually examining descriptive statistics, evaluators used multilevel modeling analyses to establish whether or not the Achieve3000 program had a statistically significant impact on reading achievement when compared to participating schools' typical publisher-developed language arts curricula and supplemental activities. For each of these analyses, the posttest score corresponding to the GMRT-4 subtest of interest (i.e., Vocabulary, Comprehension, or Total Reading) served as the outcome variable. The models accounted for the clustering of students in teachers' classrooms and included a study condition variable at the teacher-level of the model to indicate random assignment to the treatment or comparison group. Each model also included student-level pretest achievement scores as a covariate to increase the precision of the impact estimate and account for potential preexisting reading

achievement differences between the treatment and comparison groups (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007). After running each model, evaluators calculated standardized effect sizes by dividing the adjusted difference between treatment and comparison groups by the standard deviation of the comparison group. This enabled evaluators to examine the magnitude of the program impacts. In addition to calculating effect sizes, evaluators calculated WWC improvement indices when appropriate (WWC, 2014). Each improvement index reflects the change in an average comparison group student's percentile rank that would be expected if that student had participated in the Achieve3000 program instead of their school's typical curriculum.[9]

Findings from the multilevel modeling analyses calculated across grades are displayed in Table 19. On average, the Achieve3000 program did not have a statistically significant impact on students' Vocabulary scale scores. However, Achieve3000 had a statistically significant positive impact on students' Reading Comprehension and Total Reading scale scores. More specifically, at the end of the study period, treatment students who used Achieve3000 scored an average of 5.04 points, 9.49, and 7.76 points higher on the Vocabulary, Reading Comprehension, and Total Reading tests, respectively. The effect sizes for these impacts were 0.12, 0.22, and 0.20, corresponding to WWC improvement indices of 5, 9, and 8 percentile points, respectively. Although none of these was considered substantively important based on the WWC standards, the effect sizes for Reading Comprehension and Total Reading approached the WWC 0.25 threshold.

Table 19. Impact of Achieve3000 on Student GMRT-4 Performance

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. df	*p*-value	Effect Size	Improvement Index
Vocabulary	5.04	3.57	1.41	31	.17	0.12	0.05
Reading Comprehension	9.49	4.55	2.09	31	.045*	0.22	0.09
Total Reading	7.76	3.44	2.26	31	.03*	0.20	0.08

* Statistically significant at the 0.05 level.

Exploratory Analyses Comparing Reading Achievement within Grades by Study Condition

In addition to running the main analyses across grades to examine the impact of Achieve3000 on reading achievement, evaluators conducted exploratory analyses to examine the program's impact within each participating grade level. First, within each grade level, evaluators calculated descriptive statistics (means are displayed in the figures below, and other descriptive statistics are displayed in Appendix F). Next, evaluators conducted multilevel modeling analyses within each grade level. Like the main analyses, each of the subgroup analyses conducted by grade accounted for the clustering of students in teachers' classrooms and included a study condition variable at the teacher-level of the model to indicate random assignment to the treatment or comparison group. Each model also included student-level pretest achievement scores as a covariate to increase the precision of the impact estimate and

[9] According to the WWC (2014), an improvement index of 10 percentile points (corresponding to an effect size of 0.25), would suggest that an intervention would likely yield a 10% increase in percentile rank if a typical comparison-group student were to participate in the program. Additionally, it would suggest that 60% of the treatment-group students scored higher than the mean for the comparison-group students.

account for potential preexisting reading achievement differences between the treatment and comparison groups (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007). To determine the magnitude of impacts within grade levels, evaluators calculated standardized effect sizes and WWC improvement indices. It is important to note that these subgroup analyses divide the study's sample into smaller groups, which yields analyses with less statistical power to determine effects compared to the main analyses. Therefore, readers should interpret findings with caution.

Exploratory Analyses for Third Grade

KEY FINDINGS FOR THIRD GRADE:

Overall, third-grade students who used Achieve3000 during the study period performed similarly to comparison-group students who used their schools' typical literacy programs.

First, evaluators calculated means corresponding to pretest and posttest GMRT-4 scores for third-grade students by study condition. Visual examination of these means revealed that on average, students in the treatment and comparison groups gained an equal amount on the Vocabulary test, and on average, treatment-group students gained an average of 10 and 5 points more than students in the comparison group on the Reading Comprehension and Total Reading tests, respectively. Readers should note that evaluators calculated these means for descriptive purposes rather than to determine if differences in reading performance by study condition were statistically significant.

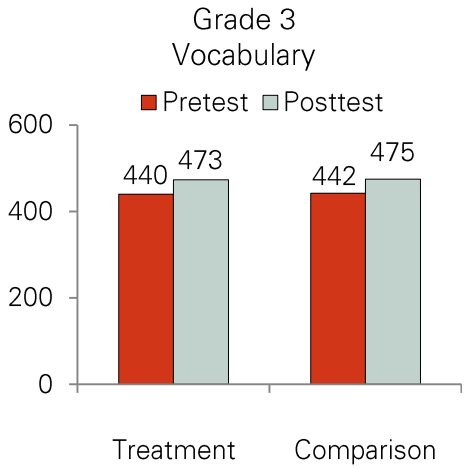


Figure 15. Third-grade unadjusted GMRT-4 Vocabulary scale score means by study condition and time.

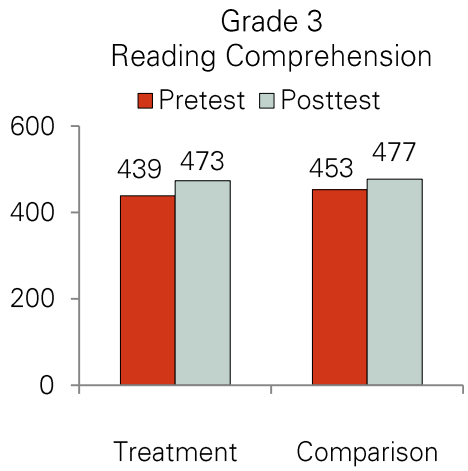


Figure 16. Third-grade unadjusted GMRT-4 Reading Comprehension scale score means by study condition and time.

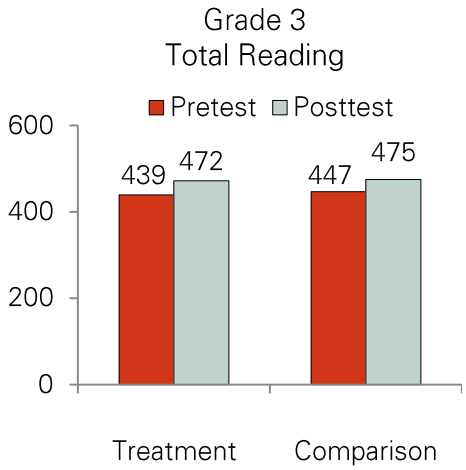


Figure 17. Third-grade unadjusted GMRT-4 Total Reading scale score means by study condition and time.

Table 20 displays the findings regarding the impact of Achieve3000 on reading achievement among participating third-grade students. Findings from these exploratory analyses revealed no statistically significant differences or substantively important effect sizes for Reading Vocabulary, Reading Comprehension, or Total Reading. The effect sizes for Reading Vocabulary, Reading Comprehension, and Total Reading (i.e., -0.02, 0.02, and 0.06) corresponded to WWC improvement indices of -1, 1, and 2 percentile points, respectively. Thus, on average, third-grade students who used Achieve3000 during the study period performed similarly to comparison-group students who used their schools' typical literacy programs.

Table 20. Impact of Achieve3000 on GMRT-4 Performance Third-Grade Students

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. df	*p*-value	Effect Size	Improvement Index
Reading Vocabulary	-0.72	7.43	-0.10	11	.93	-0.02	-0.01
Reading Comprehension	0.88	8.15	0.11	11	.92	0.02	0.01
Total Reading	2.51	6.99	0.36	11	.73	0.06	0.02

Exploratory Analyses for Sixth Grade

KEY FINDINGS FOR SIXTH GRADE:

There were no statistically significant differences in average sixth-grade treatment and comparison-group posttest reading scores, but the effect sizes favored Achieve3000 users and approached the WWC threshold of 0.25.

Before running multilevel models, evaluators calculated pretest and posttest GMRT-4 means for sixth-grade students study condition. Examining these means visually showed that

on average, students in the treatment group gained an average of 6, 7, and 8 points more than students in the comparison group on the Vocabulary, Reading Comprehension, and Total Reading tests, respectively. It is important to note that evaluators calculated these means for descriptive purposes rather than to determine if differences in reading performance by study condition were statistically significant.

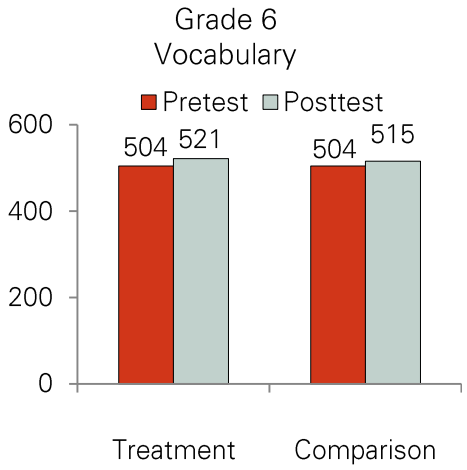


Figure 18. Sixth-grade unadjusted GMRT-4 Vocabulary scale score means by study condition and time.

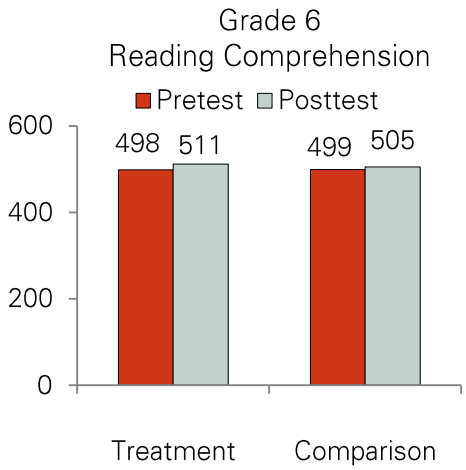


Figure 19. Sixth-grade unadjusted GMRT-4 Reading Comprehension scale score means by study condition and time.

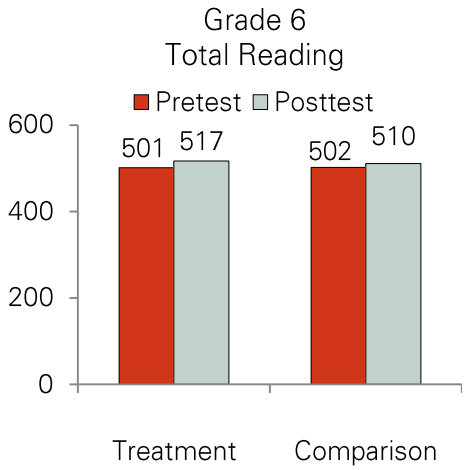


Figure 20. Sixth-grade unadjusted GMRT-4 Total Reading scale score means by study condition and time.

Table 21displays the findings regarding the impact of Achieve3000 on reading achievement among participating sixth-grade students. Although findings from these exploratory analyses revealed no statistically significant differences by study condition or substantively important effect sizes, the effect sizes (i.e., 0.21, 0.22, and 0.22) approached the WWC threshold of 0.25. Additionally, the WWC improvement indices were 8, 9, and 9

percentile points for the Reading Vocabulary, Reading Comprehension, and Total Reading, respectively.

Table 21. Impact of Achieve3000 on GMRT-4 Performance among Sixth-Grade Students

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. df	*p*-value	Effect Size	Improvement Index
Reading Vocabulary	7.97	11.38	0.70	19	.49	0.21	0.08
Reading Comprehension	8.10	6.66	1.22	19	.24	0.22	0.09
Total Reading	7.16	5.54	1.29	19	.21	0.22	0.09

Exploratory Analyses for Ninth Grade

KEY FINDINGS FOR NINTH GRADE:

There were no statistically significant differences in average ninth-grade treatment and comparison-group posttest reading scores, but the effect sizes favored Achieve3000 users and were substantively important based on the WWC threshold of 0.25.

Visual examination of ninth-grade pretest and posttest GMRT-4 means by study condition revealed that on average, treatment-group students gained an average of 5, 23, and 14 points more than comparison-group students on the Vocabulary, Reading Comprehension, and Total Reading tests, respectively. It is important to note that these means were calculated for descriptive purposes rather than to determine if differences in reading performance by study condition were statistically significant.

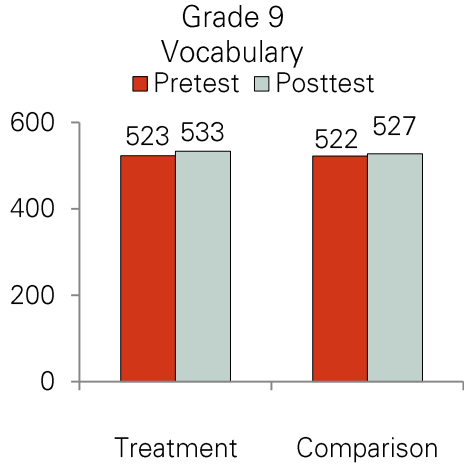


Figure 21. Ninth-grade unadjusted GMRT-4 Vocabulary scale score means by study condition and time.

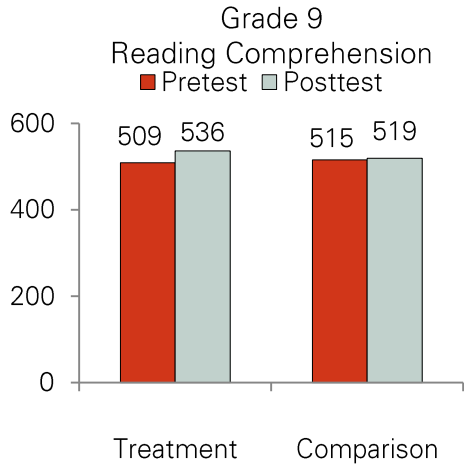


Figure 22. Ninth-grade unadjusted GMRT-4 Reading Comprehension scale score means by study condition and time.

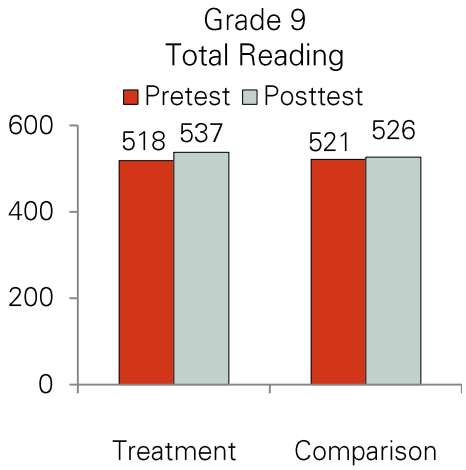


Figure 23. Ninth-grade unadjusted GMRT-4 Total Reading scale score means by study condition and time.

Table 22 displays the findings regarding the impact of Achieve3000 on reading achievement among participating ninth-grade students. Although findings from these exploratory analyses revealed no statistically significant differences by study condition, all of the associated effect sizes (i.e., 0.28, 0.51, and 0.44) exceeded the WWC threshold of 0.25 for determining whether or not effect sizes were substantively important. The WWC improvement indices corresponded to 11, 19, and 17 percentile points for Vocabulary, Reading Comprehension, and Total Reading, respectively. As indicated previously, these subgroup analyses had less statistical power to detect effects than the main analyses, so readers should use caution when interpreting the *p*-values for these findings, as the substantively important effect sizes suggest that ninth-grade students who used Achieve3000 during the study period out-performed comparison students who used their schools' typical literacy programs.

Table 22. Impact of Achieve3000 on GMRT-4 Performance among Ninth-Grade Students

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. df	*p*-value	Effect Size	Improvement Index
Reading Vocabulary	9.86	8.71	1.13	10	.28	0.28**	0.11
Reading Comprehension	19.26	9.01	2.14	10	.06	0.51**	0.19
Total Reading	14.51	7.24	2.00	10	.07	0.44**	0.17

**Substantively important based on the What Works Clearinghouse Standards.

Exploratory Analyses Comparing ELL Reading Achievement by Study Condition

KEY FINDINGS FOR ELL STUDENTS:

Findings suggest that ELL students who used Achieve3000 performed similarly on the GMRT-4 as ELL students who used their schools' typical literacy programs.

After running within-grade analyses, evaluators ran additional exploratory analyses to examine Achieve3000's impact on students classified as ELL. Before running multilevel modeling analyses, evaluators calculated means for pretest and posttest (visually displayed in the figures below). Visual examination of these means revealed that on average, ELL students in the treatment group gained 3 more points on the Vocabulary test and 2 more points on the Total Reading test than those in the comparison group. The average Reading Comprehension gains of ELL students were similar by study condition. Readers should note that evaluators calculated these means for descriptive purposes rather than to determine if differences in reading performance by study condition were statistically significant.

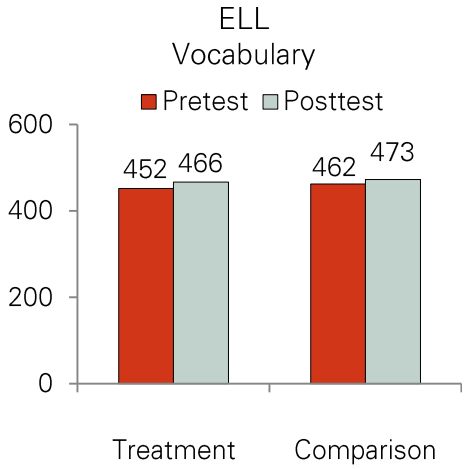


Figure 24. ELL students' unadjusted GMRT-4 Vocabulary scale score means by study condition and time.

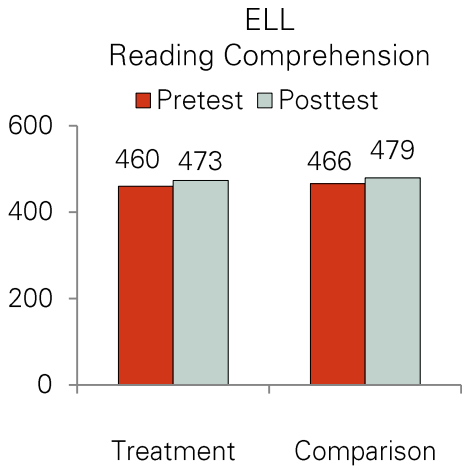


Figure 25. ELL students' unadjusted GMRT-4 Reading Comprehension scale score means by study condition and time.

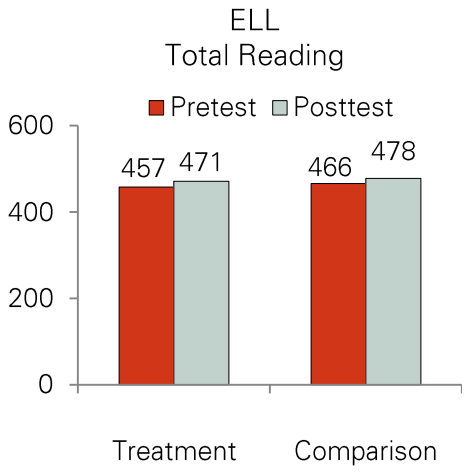


Figure 26. ELL students' unadjusted GMRT-4 Total Reading scale score means by study condition and time.

After calculating means, evaluators conducted multilevel modeling analyses that accounted for the clustering of ELL students in teachers' classrooms. Evaluators also included a study condition variable at the teacher-level of the model to indicate random assignment to the treatment or comparison group. Additionally, the models included student-level pretest achievement scores as a covariate to increase the precision of the impact estimate and account for potential preexisting reading achievement differences between the treatment and comparison groups (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007). Evaluators then calculated standardized effect sizes and WWC improvement indices. Readers should note that because these ELL subgroup analyses divide the study's sample into smaller groups, the analyses reduced statistical power to determine if there were statistically significant effects compared to the main analyses. Thus, findings should be interpreted with caution.

Table 23 displays the findings regarding the impact of Achieve3000 on reading achievement among ELL students. Findings showed no statistically significant differences or substantively important effect sizes for Reading Vocabulary, Reading Comprehension, or Total Reading. The effect sizes for Reading Vocabulary, Reading Comprehension, and Total Reading (i.e., -0.07, -0.06, and -0.05) corresponded to WWC improvement indices of -3, -2, and -2 percentile points, respectively. Thus, on average, ELL students who used Achieve3000 during the study period performed similarly to comparison-group ELL students who used their schools' typical literacy programs.

Table 23. Impact of Achieve3000 on GMRT-4 Performance for ELL Students

Outcome Variable	Coefficient	Standard Error	*t*-value	Approx. df	*p*-value	Effect Size	Improvement Index
Reading Vocabulary	-2.65	5.32	-0.50	20	0.62	-0.07	-0.03
Reading Comprehension	-2.14	6.65	-0.32	20	0.75	-0.06	-0.02
Total Reading	-1.98	4.85	-0.41	20	0.69	-0.05	-0.02

Summary of Findings Comparing Reading Achievement by Study Condition

Overall, findings from analyses comparing the performance of treatment and comparison group students indicated that the Achieve3000 program had a statistically significant impact on posttest GMRT-4 Reading Comprehension and Total Reading scores when compared to typical ELA programs. Exploratory analyses suggested that impacts varied by grade, and although none of the within-grade findings were statistically significant, readers should note that these subgroup analyses had less statistical power to detect program effects than the main analyses. The effect sizes for third grade did not approach the WWC threshold of 0.25, but the effect sizes for sixth-grade approached it, and the effect sizes for ninth-grade exceeded the WWC threshold. Thus, across grades, Achieve3000 had a statistically significant impact on GMRT-4 Reading Comprehension and Total Reading, and it had a substantively important impact on ninth-grade GMRT-4 Vocabulary, Reading Comprehension, and Total Reading. Table 24 summarizes these findings. Finally, there were no statistically significant differences by study condition for ELL students, suggesting that the Achieve3000 and comparison programs performed similarly for this subgroup of students.

Table 24. Summary of Achieve3000 Program Impacts

All Grades Combined			
Assessment	Effect Size	Statistically Significant Impact	Substantively Important Effect Size*
GMRT-4 Vocabulary	0.12		
GMRT-4 Reading Comprehension	0.22	◆	
GMRT-4 Total Reading	0.20	◆	
Third Grade			
Assessment	Effect Size	Statistically Significant Positive Impact	Substantively Important Effect Size*
GMRT-4 Vocabulary	-0.02		
GMRT-4 Reading Comprehension	0.02		
GMRT-4 Total Reading	0.06		
Sixth Grade			
Assessment	Effect Size	Statistically Significant Positive Impact	Substantively Important Effect Size*
GMRT-4 Vocabulary	0.21		
GMRT-4 Reading Comprehension	0.22		
GMRT-4 Total Reading	0.22		
Ninth Grade			
Assessment	Effect Size	Statistically Significant Positive Impact	Substantively Important Effect Size*
GMRT-4 Vocabulary	0.28		◆
GMRT-4 Reading Comprehension	0.51		◆
GMRT-4 Total Reading	0.44		◆
ELL Students			
Assessment	Effect Size	Statistically Significant Positive Impact	Substantively Important Effect Size*
GMRT-4 Vocabulary	-0.07		
GMRT-4 Reading Comprehension	-0.06		
GMRT-4 Total Reading	-0.05		

*Substantively important based on the What Works Clearinghouse Standards.

Teacher Perceptions

KEY FINDING:

Treatment teachers found the Achieve3000 program components useful and described many benefits to the program including differentiation, less time for lesson preparation, and positive effects on student engagement and student growth. However, some treatment teachers expressed some frustration with program navigation and technology issues and offered suggestions for improvement.

This section presents the results of teacher perceptions of the Achieve3000 program and comparison programs. Evaluators asked treatment teachers to respond to a series of questions about their perceptions of the Achieve3000 program on the online weekly treatment teacher logs. On the final weekly log, evaluators asked treatment teachers to respond to additional retrospective questions about their experiences throughout the entire study. Evaluators asked comparison teachers to respond to perception questions on a one-time spring comparison teacher survey. Therefore, it is important to note that treatment teacher perceptions were collected over 32 online weekly logs and comparison teacher perceptions were collected on a one-time comparison teacher online survey.

Treatment Teacher Perceptions of Achieve3000

On the online weekly logs, treatment teachers responded to questions about the effectiveness of Achieve3000 program activities, administrative components, teacher program components, and program impacts. Treatment teachers also compared the Achieve3000 program to other literacy programs they had used previously and indicated whether or not they would want to continue using Achieve3000. Finally, treatment teachers shared what they liked and disliked about the program and had an opportunity to provide suggestions for program improvements.

Perceptions of Achieve3000 Activities

On the weekly logs, treatment teachers rated the effectiveness of Achieve3000 activities at engaging students using a 6-point Likert scale (*not applicable*, *very ineffective*, *ineffective*, *neither ineffective nor effective*, *effective, very effective).* Teachers most often said the activities of "poll results," "stretch article," and "stretch activity" were *effective* at engaging students. Teachers most often said the math activities were *neither ineffective nor effective*.

Table 25. Average Teacher Ratings of Effectiveness of Achieve3000 Activities at Engaging Students

Student Needs	Very Ineffective	Ineffective	Neither Ineffective nor Effective	Effective	Very Effective	Not Applicable
Poll Results	0.51%	4.60%	39.13%	44.75%	11.00%	0.00%
Math	0.13%	2.24%	52.02%	41.26%	3.14%	0.00%

Student Needs	Very Ineffective	Ineffective	Neither Ineffective nor Effective	Effective	Very Effective	Not Applicable
Stretch Article	1.16%	1.94%	27.13%	41.09%	28.68%	0.00%
Stretch Activity	1.35%	1.79%	25.56%	40.81%	30.49%	0.00%

Perceptions of Achieve3000 Administrative Components

In addition to responding to perception questions about program activities, treatment teachers also rated the usefulness of the Achieve3000 administrative components on a 6-point Likert scale (*not applicable, not useful at all, not very useful, moderately useful, useful, very useful*). Teachers rated the components of "student work," "usage reports," "performance reports," and "assessment tools" as *useful* or *very useful*. Teachers most often rated the component of "home communication" as *moderately useful* or *useful*.

Table 26. Average Teacher Ratings of Usefulness of Achieve3000 Administrative Components

Student Needs	Not Useful at All	Not Very Useful	Moderately Useful	Useful	Very Useful	Not Applicable
Student Work	0.20%	1.02%	11.25%	41.31%	46.22%	0.00%
Usage Reports	0.00%	1.06%	8.46%	47.57%	42.92%	0.00%
Performance Reports	0.21%	0.64%	7.43%	47.13%	44.59%	0.00%
Assessment Tools	0.30%	5.06%	17.86%	49.11%	27.68%	0.00%
Home Communication	1.90%	13.33%	34.29%	33.33%	17.14%	0.00%

Perceptions of Achieve3000 Teacher Components

On the weekly logs, treatment teachers also rated the usefulness of various teacher program components using the same 6-point scale described in Table 27. Teachers most often rated all of the teacher program components as *useful* or *very useful*.

Table 27. Average Teacher Ratings of Achieve3000 Teacher Components

Student Needs	Not Useful at All	Not Very Useful	Moderately Useful	Useful	Very Useful	Not Applicable
Teacher Recommendations	1.56%	5.31%	18.44%	47.50%	27.19%	0.00%
Discuss/review Lesson Vocabulary	0.23%	0.23%	10.23%	47.44%	41.86%	0.00%
Answer Keys	0.67%	5.37%	23.83%	46.31%	23.83%	0.00%
Curriculum Key	0.57%	0.85%	16.43%	47.59%	34.56%	0.00%
Graphic Organizer	0.32%	1.29%	20.32%	48.06%	30.00%	0.00%
Standards	0.00%	1.02%	16.24%	43.65%	39.09%	0.00%
Strategy Lesson	0.00%	0.00%	20.98%	53.16%	25.86%	0.00%
ELL & Struggling Readers Supports	0.00%	9.96%	22.99%	42.91%	24.14%	0.00%
Gifted & Talented Supports	1.48%	15.27%	18.23%	45.81%	19.21%	0.00%

Treatment Teachers' Comparisons of Achieve3000 and Other Programs

"I LOVE that this program is differentiated for my students, it's ready to use, with lots of different lesson options from which I can choose, and many of the articles are interesting to the students. We basically have nothing else like this for NONFICTION material, and it's better than anything else I've seen. Achieve3000 is very effective in informational text that is current and interesting for students."
Quote from teacher logs

On the final implementation log, treatment teachers were asked to compare the Achieve3000 program to other literacy programs they had used. A few teachers had not used another digital literacy program before or were first-year teachers, so they did not provide a comparison, but the majority of teachers were very positive about the Achieve3000 program and its components in comparison to other literacy programs. Teachers commented, for example, that Achieve3000 was "by far the BEST," "the best I have seen in 20 years" and "I have never seen one [a program] come close to what Achieve offers."

When comparing Achieve3000 to other literacy programs, treatment teachers appreciated various program components and especially the differentiated Lexile levels, the constant feedback, and the reporting options. For example, one teacher said, "It is very effective for individual students because of the differentiated levels." Another teacher said, "I love that this allows me to level each of their articles. I can talk to my entire class about a topic and have them read at their own level. I have not found another literacy program that does this so thoroughly." Teachers also mentioned the Achieve3000 program was different from other programs because of the constant feedback and in-depth reading comprehension reports.

Treatment teachers were also very impressed by student engagement with the Achieve3000 program in comparison to other literacy programs. Teachers mentioned they liked the high interest articles and the variety of options for teachers and students to choose from. One teacher said, "[Achieve3000 is] THE BEST nonfiction program I've ever used. The rigor is amazing and kids get into it!" Another teacher said, "In comparison to worksheets or textbooks, I saw more engagement from my kids with Achieve 3000."

Teachers' Perceptions Regarding Achieve3000 Improvements

"The growth my kids made was amazing. They took pride in wanting to see how they grew!"
Quote from teacher logs

On each weekly log, treatment teachers were asked to describe if the Achieve3000 program improved students' reading skills. Teachers' comments about the effects of the Achieve3000 on student learning were mixed. Most teachers said the program helped with improving student literacy and comprehension skills, but others said that certain students struggled with the program, and it did not improve their literacy and comprehension skills as much as other programs.

Of the treatment teachers who said student reading and comprehension levels improved, some said

students made "significant gains." One said, "It exceeded my expectations," and another said, "It really helped with comprehending nonfiction texts." Similarly, another teacher reported, "I think that it's helped my students a lot. I haven't had access to anything this helpful in the past." In addition, some treatment teachers also mentioned they had seen more growth with Achieve3000 than with other literacy programs. For example, one teacher said, "Wonderful - I have never seen the overall growth in classes as I have had this year."

Treatment teachers also mentioned that Achieve3000 impacted student confidence in reading. For example, one teacher said, "Students are ready to tackle higher level thought questions more than they were at the beginning of the year." Another teacher said, "They [my students] are more confident and are careful readers!"

While the majority of teachers said the program had a positive impact on students, some teachers said that their students did not make as much progress in the program as they would have liked, and Achieve3000 may not have been effective for students with low reading levels. In particular, teachers pointed out that students who read on lower levels (including special education students) had a hard time using this program independently and may not have made as much growth as students with average or high reading levels. One teacher said, "I had mixed results mid-year and not as much gain as I anticipated. Scores were a little better but what I found was good readers, who could read independently, made more progress. My struggling readers did not make gains to "close the gap" as I had hoped (according to A3000 Lexile levels)." Another teacher said, "I am not sure that it helped. I see bigger gains and higher interest when using other materials. I feel a lot of students were resistant to the program because they did not like just sitting there reading silently. I do think it helped those students who were open to the program."

Teachers' Plans for Future Use of Achieve3000

Using the logs, evaluators also asked treatment teachers if they would use the Achieve3000 program next year and if they would make any changes to their implementation. Most treatment teachers (73.91%) said they would use the Achieve3000 program next year, but 52.82% said they would implement it differently. Some teachers noted that they would use the program less next year, and others said they would use it more. For example, one teacher said the model of implementing the program twice a week in regular English classrooms did not allow enough time to teach the district-mandated curriculums. Teachers offered various options for implementing the program in the classroom such as using it as a tool when students finish assignments. One teacher said, "I would like to utilize the program more frequently than I did

TEACHER SUGGESTIONS FOR IMPLEMENTATION CHANGES

- Do more pre-reading activities before the lesson.
- Select articles and thought questions that better align with the curriculum.
- Have students use a notebook or journal to track articles read and scores.
- Use more of the graphic organizers.
- Use more of the teacher recommendations.
- Use more of the various reports.
- Use more suggestions from support staff for implementing the program & tracking student progress.
- Use incentives earlier.
- Grade more writing components.
- Do more stretch article work.

this year. I would also like to use this program to administer assessments and support student grading."

Achieve30000 Components that Teachers Particularly Liked

KEY FINDING:

Overall, teachers were very positive about the Achieve3000 program and said that the materials were comprehensive, engaging for students, and increased student achievement.

The weekly log also provided treatment teachers an opportunity to report anything they particularly liked or disliked about using the Achieve3000 materials. Overall, teachers were very positive about the Achieve3000 program and found the materials comprehensive, engaging for students, and increased student achievement. However, treatment teachers expressed some frustration with program navigation and technology issues and reported that the program may not be meeting the needs of all students.

Teachers liked various Achieve3000 program components such as the "my lesson plan" option, the articles, the graphic organizer, writing activities, and grammar activities. One teacher said, "I enjoyed the "It's a Rap!" bonus article & lesson. We watched the video & the students liked that it was about rap music. The fact that the activity had two parts was a nice change for the students." Teachers also reported enjoying using the graphic organizer, the vocabulary lessons, classroom discussion, and the scoreboard feature. One teacher said, "I really liked the "scoreboard" feature. I have made it a competition and my students have really starting getting into it." Another teacher said, "I really liked the gifted/talented video and ideas to show with all kids."

Teachers appreciated the assessment component of the program because it encouraged student motivation and monitoring. For instance, one teacher said, "It gave me concrete reading scores when meeting with parents." Another teacher said, "This week, the students were able to retest for their placement. I thought it was great and it allowed the students to see their growth which reinforced their desire to continue Achieve."

"I like the fact I am providing high quality, high interest informational text for my students. I also love the fact it is on their reading level."
Teacher log quote

"The way it makes the kids think. I Love everything!"
Teacher log quote

Teachers also appreciated that the Achieve3000 lessons are well-planned and they needed less planning time to prepare for a lesson. For example one teacher said she liked "the idea that the lessons are well-planned and prepared and I didn't have to do it." Another teacher said it was easy for a sub to fill in and implement the first 4 steps of Achieve. Another teacher said, "It is not too challenging to plan for." One teacher said the program "has a systematic approach to teach skills. They are easy to use and implement."

Treatment teachers reported on their weekly logs how much their students were engaged with the Achieve3000 program. Overall, teachers liked the "engaging articles," the "pictures," and the general "engagement of students." Teachers said that students were engaged in the program because they get to use technology (i.e. iPads, chromebooks), they like to search for their own high-interest articles and read nonfiction texts, and they get to "self-manage" their learning experience. One teacher said, "On one of the implementation days, I let the students search for any article they thought looked interesting. They always enjoy doing that." In addition, most teachers said the program has been able to engage most students, even the low readers, and students generally "really like it" "enjoy it" or "love it." Teachers also appreciated the accountability system and the incentives to keep students engaged. One teacher said, "They [the students] are finally grasping the importance of close reading!"

Most treatment teachers also expressed satisfaction with the progress students made and the advantage of having assessment components to see and measure student growth. One teacher said, "I like seeing student progress from the beginning of the year until now. I like that it stretches the students' vocabulary." Another teacher said, "The results this time of year were amazing. I feel Achieve has been the key to the results I am seeing." Another said, "My students are making significant progress using textual evidence. They are also increasing their reading levels more than I thought possible." Another said, "I am enjoying seeing their progression and how they've gotten a higher reading score."

"I love the article topics. There are so many articles I cannot wait to teach! As a social studies teacher, I love the world history connections." *Teacher log quote*

Most teachers reported that they liked the differentiation that the Achieve3000 program offers. The Lexile leveled readings are "interesting" and "allows almost everyone to be actively engaged the entire time we use the program." For example, one teacher said, "I really like that the program is written on their reading level and has the vocabulary words for the students." Another teacher liked "having appropriate informational text."

Teachers said that they and their students liked having various options to choose from. For example, teachers mentioned that they loved the following content areas: dinosaur articles, science (space, planets), articles about lions and tigers, holidays, the Fenway article, "It's a Rap," and the flying car lesson. Teachers mentioned these articles were "really interesting to the kids," that they align to the curriculum and probe good discussion. Teachers also liked having the option to pick stories or topics relevant to the holiday or seasons (i.e. baseball season), or stories that align with existing classroom curriculum. For example, one teacher said, "The students are writing a persuasive essay. I was able to find articles at their level related to the topic they are writing about. It has been very helpful." Another teacher liked, "being able to piggy back on the science lessons students had with an Achieve3000 lesson." Treatment teachers also appreciated that the articles teach students about various cultures and events.

On the final log, teachers were asked to reflect on what they particularly liked about the program over the entire year. Teachers reported similar positive aspects of the program, and noted that they especially liked the variety of articles, high-interest and engaging content, the

use of technology, the instant feedback, independence, using the graphic organizer, the engagement of low-level readers, growth tracking, and the impact on student literacy (see Figure 27). In particular, one teacher said "the activities and the instant feedback were very effective for my students. It showed them right away whether they were right or wrong. I love that I was able to find high interest articles for my students. It was very helpful in keeping my students engaged."

"The high interest articles and the fact even lower level readers were reading the same thing in a non-babyish way. This whole program held the students' interest all year long.
Teacher log quote

"I loved seeing my students growth! Every month when the Lexile levels were adjusted, we celebrated success. As the months went on, more and more of my students were making gains!
Teacher log quote

"I enjoyed that each article was pushed out at their own level. I thought the articles were interesting and so did the students. I loved the growth reports and the Sunday report that was sent. Overall one of the best programs I have worked with!
Teacher log quote

"Everything! I love the program and my students have made big gains in their reading level and ability to use textual evidence."
Teacher log quote

Figure 27. Teacher Final Log Quotes

Achieve30000 Components that Teachers Particularly Disliked

KEY FINDING:

Some treatment teachers expressed frustration with program navigation and technology issues and reported that the program may not be meeting the needs of all students.

Treatment teachers were also asked on the weekly logs to report anything they disliked about using Achieve3000. Some teachers disliked components of the program such as its applicability for low and high-level readers, student engagement, and not having enough time to implement their curriculum. Others were dissatisfied with the training, and some experienced problems using the technology.

While some teachers appreciated the differentiated reading levels and the program's ability to meet the needs of all students, some teachers found the program too difficult for low-level readers and high-level readers. For low-level readers, one teacher reported that low-level students were frustrated because they weren't able to get 75% or better no matter how hard they tried. One teacher said, "My SPED and ELLs students are struggling." Another teacher said, "Students with very low reading levels are having difficulty achieving a 75% or higher. They become frustrated and give up easily. The wording of the questions is difficult because words they don't know are in the actual question not the answer choices." A few teachers said the articles are too complex for their higher-level students. One teacher said, "They [my higher level students] are having a hard time completing as many assignments due to the complexity of their articles." Another teacher said, "I had a girl, who is about a 1320 Lexile, really struggle with one of her articles so we did a second one together. I know 1320 is high, but wow, it was

difficult." Another teacher said, "Still don't like that kids with a 1400 or higher Lexile can't take the stretch article for practice at grade level. I have been surprised the number of kids whose Lexile is higher than the stretch article are having trouble with it."

Some teachers said they struggled with student engagement including interest in content, the lessons being too repetitive, and student gaming (students clicking through the system or logging onto different applications during the Achieve3000 lesson). Some teachers also noted that the subject matter of the articles is too hard and said it was necessary to build background knowledge in the content area to engage students. A few teachers wished there were more entertaining articles and said they were out of date and not engaging for students because the content is "not interesting." One teacher said, "My students are losing interest and it is becoming a battle to get them to use it." Other teachers said the program became monotonous throughout the year and students got "burnt out on informational texts." One teacher said, "Having the same five step process for each article can be repetitive for both the teacher and the student." Multiple teachers reported that students have figured out that in the activity section, they can just click on the answers twice and then it gives them the answer and they complete the activity too quickly. For example, one teacher said, "I don't like that after two guesses in the activity section that they are given the answer. It would be nice to have some sort of way to not allow the students to just guess to move on." Lastly, some teachers said they noticed students are logged onto different applications during the lesson and they have to police students to stay on task.

Multiple teachers said the program took too much time away from their literacy curriculum. These teachers qualified that this lack of time is not Achieve3000's fault, but they felt rushed to try and complete the district and state requirements in addition to Achieve3000. One teacher said "I don't like that it takes away from my main curriculum. I can't fully discuss the fictional literature with my class because I have to do this twice a week." Another said:

> "I continue to dislike that I have to take two days out of my week to implement the program. There is often a huge disconnect between what I am teaching and what they are reading using the program. I feel like it's reading just to read and they aren't gaining much from it. I also do not like that it is done independently when I am being taught Common Core is all about students working together. I don't like that it takes away from my curriculum and the literature I am supposed to be teaching in my class. I then feel stressed to balance this program as well as the rest of the freshmen curriculum I am supposed to be teaching."

KEY FINDINGS:

Many teachers said they were dissatisfied with the training and either needed more in depth training, more follow-up training or needed the training to take place earlier in the year.

Multiple teachers also reported disliking the technology problems they experienced with the application including issues with scoring, grading, and being logged off the program.

Many teachers said they were dissatisfied with the training and either needed (a) more in- depth training, (b) more follow-up training or (c) training sessions earlier in the year. One teacher said, "I struggle to use the articles and have not been trained to use the program effectively." Another teacher said, "We were able to get additional training from an

Achieve3000 representative this week. It was extremely helpful and will really make a difference in my students' progress. It would have been much better if this training would have taken place right after my students completed the LevelSet testing but better late than never. My students have been making huge gains in their Lexile scores even without the additional training but now I think it will be even better."

Multiple teachers also reported disliking the technology problems they experienced with the application including issues with scoring, grading, and being logged off the program. One teacher said, "The Achieve3000 application was very glitchy this week." Another said, "My most frustrating moments came from technology problems, like I mentioned earlier in the survey (problems with the iPads, but not really with the program itself)." Specifically, teachers had trouble with the score reports, which would not match what the students were doing. One teacher said the program "would count correct answers as incomplete and score them as wrong." Another teacher had an issue with the activities disappearing from the task bar and the congratulations pages saying students passed an activity, but not showing up on the homepage as completed. One teacher said students could not complete their thought questions on the iPads.

TEACHER SUGGESTIONS FOR TRAINING MODIFICATIONS

- More one-on-one training with the school teams to navigate through the program and questions teachers have.
- Spend more time planning and practicing.
- Have a quicker follow-up training at the beginning of the year so teachers can ask questions once they are using it.
- Have a mini training with teachers throughout the year to discuss problems/questions and in-depth activities.

Perceptions of Training

Treatment teachers were asked if they received effective Achieve3000 training. None of the treatment teachers said "no," 34.78% of teachers said they received a partially effective training and 65.22% said they received an effective training.

In particular, one teacher said, "I really appreciated the visits from [the Achieve3000 trainer] and would like more of those. I don't feel I used the teacher resources in the most effective ways. If in training, we could have seen a print-out of each resource and known how to access them and USE them, it would have really helped me. If I can't see a hard copy of things, I don't necessarily know they even exist to access them."

Treatment Teachers' Suggestions for Improving Achieve3000

On the logs, treatment teachers were also asked to provide recommendations for improving the program. While most teachers offered positive comments about the program and multiple areas they liked, teachers also reported areas that could be improved. Multiple teachers said they had issues with students rushing right to questions, clicking through, and not reading the article. One teacher said, "The

"I would like to see more articles in pop culture, current events, and sports. It would be nice if my students could read about things, people, and events that interest them and are relevant to their lives." *Log Teacher Quote*

only consistent issue that I had was that I had some students just click answers until it gave them the right answer. I think it would be nice that if the students got an answer wrong, they would have another question added to their assessment at the end. They would have to answer 8 questions correct, and they answered questions until they had answered a total of 8 correct. This would eliminate the rushing through it and clicking on answers because if they got it wrong, another question would be added to their assessment." Another suggested option was to require a set amount of time for reading an article before students could progress to the questions.

TEACHER RECOMMENDATIONS FOR IMROVEMENT

- Turn some of the weekend articles into 5-steps.
- Make the teacher answer keys more user-friendly.
- Make the point system more rewarding (offer students game time or prizes they can work towards).
- Make it easier to grade the thought question responses.
- Add videos to help with understanding for ELL and struggling readers.
- Add more articles in pop culture, current events, and sports.
- Include more visuals (maps, diagrams).
- Make the stretch article and activity available through the application.
- Allow students to digitally underline text.
- Add a summarize button so students can keep their comments instead of having to explain under "setting the purpose."
- Improve program glitches and issues with losing information, and answer-switching.
- Format the articles to be more printer friendly on one page.
- Have the bonus lesson count toward the number of activities.
- Not have the highlighting tool bar cover the text on the screen.
- Don't have answers that are so tricky (implied in the article, but not stated).
- Have more math problem support for literacy teachers.
- Make sure answer options are not too close (confusing for the student and the teacher).
- Have visuals for difficult vocabulary.
- Reduce the time to take the LevelSet.

Comparison Teachers' Perceptions of Their Literacy Programs

KEY FINDINGS:

Overall, most comparison teachers said their core literacy programs were *moderately useful* (60.87%) or *useful* (39.13%). Most teachers said their comparison programs positively impacted student achievement, but most teachers also said their materials were outdated and did not meet the needs of all levels of students.

On the spring comparison teacher online survey, comparison teachers were asked a series of questions about their perceptions of their ELA programs and those programs' effects

on students. As reported earlier, comparison teachers used various core literacy programs in addition to supplemental materials.

Comparison Teachers' Perceptions of Impacts on Student Learning

On the comparison teacher survey, teachers were asked to rate how well their comparison materials impacted student learning in literacy. Many comparison teachers found their programs to be beneficial for student learning and engagement and have seen students "make very good progress." For example, one teacher said "student literacy has improved" and another said the program has "increase[d] literacy and improve[d] critical thinking skills."

In addition to student literacy, comparison teachers mentioned that their programs improved student writing, student engagement, and helped students become better problem solvers. One teacher mentioned that student writing skills had improved, "especially in the areas of development, transitions and evidence." One teacher said the literacy program, "has affected student learning by a noticeable increase in students' engagement and understanding." One teacher was very positive about the comparison program and said, "It has helped with building academic language, encouraged students to become problem solvers and build their critical thinking skills, and encouraged students to relate to real world issues."

While most comparison teachers found their current literacy materials were effective for the average and above-average students, some teachers said there was not enough support for struggling readers, and the materials were too difficult for students who were below grade level. One teacher said, "It is always a struggle to move our students up from a 3rd grade reading level to 6th grade. We have many low reading levels, and the current materials I believe are insufficient." Another comparison teacher said, "The literature that we have available to us is outdated and makes it hard for the students to understand the language." One teacher said "It takes a lot of planning and scrounging of materials to meet the needs of all level [of students]." Some comparison teachers said that due to inefficiencies in their programs, they could manage student learning better with supplemental activities, and students were more engaged with supplemental activities and materials.

Comparison Teachers' Perceptions of Impacts on Student Interest

Comparison teachers were also asked to comment on how their ELA programs affected student interest in literacy. Most comparison teachers were very positive about their programs' impact on student interest. The majority of teachers said they used materials with high interest topics or nonfiction articles to keep student interest and curiosity high. For instance, one teacher said, "My students are very interested in the subjects they discuss and inadvertently are showing an interest in literacy." Multiple teachers said they like to give students choices whenever possible to keep interest high. One teacher said, "Some of the articles are really relevant and student engagement increases, depending on the topic. I have even given students the opportunity to bring in class appropriate texts and that has also worked." Another teacher said, "My kids have been very interested in our reading units. All of our units have to do with our state history, so they like to learn about these things." Lastly, one teacher reported: "Some of the stories we have read have really improved my students' interest. I have noticed even some students who struggle with paying attention have been super engaged in class readings and discussions."

Comparison teachers also were critical of their programs and said their programs were lacking in certain areas, which affected student interest and sometimes didn't meet the needs of students at all levels. Some comparison teachers said novels can be challenging for students and that they try to find interesting nonfiction articles, but it can take a lot of time finding interesting articles that align with a standard. One teacher said, "some of the stories and materials are interesting but I wish it would grab their attention more and be exciting to them." Some teachers said the materials were outdated, one commenting that "the literature that we have available to us is a bit outdated and hard for the students to relate to." Another teacher said students were not engaged because they don't get instant feedback. The teachers who expressed dissatisfaction with their ELA programs said their students had "minimal interest" or "interest is below adequate." Some comparison teachers said they had to supplement their programs with additional materials to increase the level of student interest.

Comparison Program Components that Teachers Particularly Liked or Disliked

On the spring survey, comparison teachers were asked to described the benefits of their literacy materials, and to report any challenges or difficulties with the comparison program or lack of a program. In addition to increasing literacy and student learning as reported in the previous section, comparison teachers also said their programs helped to strengthen additional skills such as figurative language, interpretation of text, word formation through study of roots and stems, author's purpose and rhetorical devices, language development, grammar, and critical thinking. Specifically, one teacher said, "The students have become very good at citing textual evidence from the core materials." Another teacher said, "I like how it can serve the wide range of materials offered and can meet the needs in my classroom."

In citing challenges with their programs, comparison teachers mentioned not having enough high quality materials and assessments to meet the needs of all students. The most common complaint was that materials were outdated (12 years old), that students lost interest, and the materials were too complex for some students. One teacher said, "The required texts are very outdated which made it difficult for students to understand the language, relate to the material, or identify with the themes or characters." Comparison teachers complained about not having enough materials and options to meet the needs of individual students and especially advanced readers and ESL students, or students whose reading levels are dramatically low. Another teacher said, "I would love it if our school provided us with the proper curriculum. As a first year teacher teaching 6th grade, it takes a lot of time to research my own materials."

Comparison teachers offered some suggestions for improvement. First, comparison teachers would like to see more practice for skills such as vocabulary, affixes, synonyms, antonyms, homonyms, homophones, resource materials (thesaurus, atlas), alphabetizing to the third letters, maps, charts, graphs, schedules, and grammar. Comparison teachers would also like to see their comparison programs incorporate the following: online components, practice books for each student to work in, a homework practice book for students to take home and practice the skills taught, aids to encourage parent involvement, and a variety of reading formats (magazines, newspapers, brochures, etc.).

Comparisons of Teacher Perceptions Regarding Achieve3000 and Comparison Programs

KEY FINDING:

More treatment teachers described Achieve3000 as having higher student engagement, and an appropriate amount and pacing of materials, than did teachers of comparison programs.

This section compares similar questions from the treatment teacher online weekly logs and the online comparison teacher survey to facilitate comparisons of results across study conditions. It is important to note that treatment teacher results were averaged over the 32 online weekly logs and the comparison teacher results are from a one-time online survey. Teachers in both study conditions were asked to describe student engagement, amount of material, program pacing, student needs, and students' skills.

Perceived Student Engagement in Achieve3000 and Comparison Programs

Treatment and comparison teachers rated student engagement based on their in-class observations. Both groups of teachers most often reported that their students exhibited "high engagement" or "average engagement." Visual examination of means suggests that on average, treatment teachers appeared more likely than comparison teachers to report that students were highly engaged (see Figure 28).

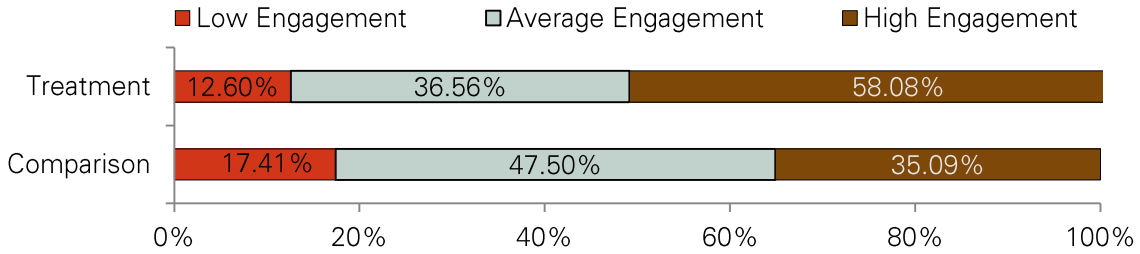


Figure 28. Teachers' reports of percentages of students engaged in Achieve3000 or comparison programs.

Perceptions of Achieve3000 and Comparison Materials

Treatment and comparison teachers also reported their perceptions about the amount of materials to cover. On average, treatment teachers were most likely to report that the Achieve3000 program was *just right* (75.25%). For comparison teachers, most teachers (52.17%) said there were *not enough* materials to cover, followed by *too much* to cover (39.13%). Based on a visual examination of the averages reported in Figure 29, it appears that treatment teachers were more likely than comparison teachers to report that their program offered the right amount of material.

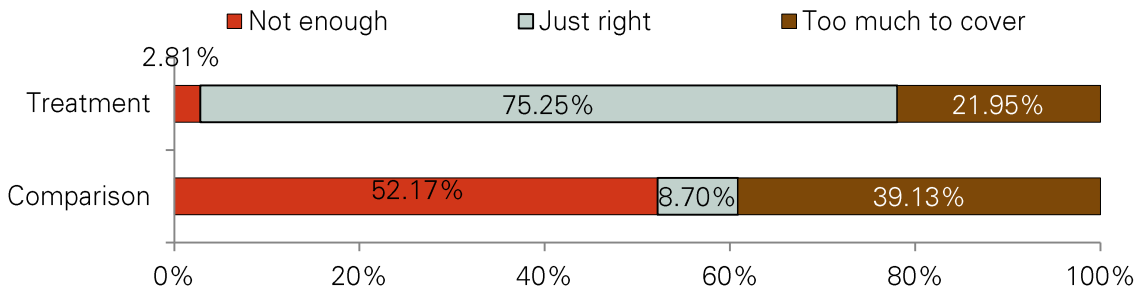


Figure 29. Treatment and comparison teachers' perceptions of the amount of material to cover.

Perceptions of Pacing of Achieve3000 and Comparison Programs

Study teachers were also asked to describe the overall pacing of the Achieve3000 and comparison programs. Most treatment (80.76%) and comparison teachers (60.87%) said their programs were *reasonably paced*. The results displayed in Figure 30 suggest that more treatment than comparison teachers judged their programs to be *reasonably paced*.

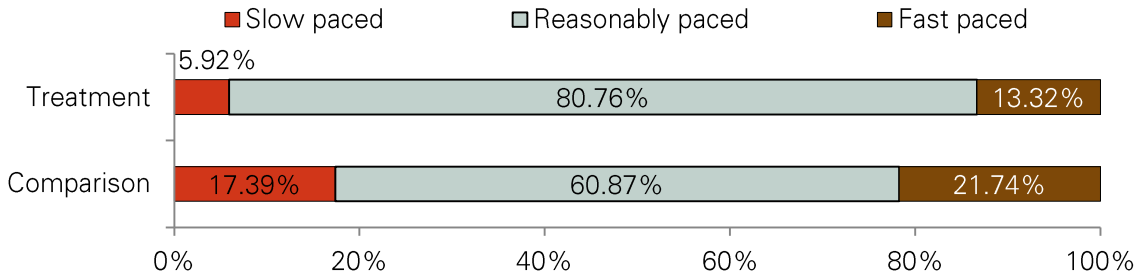


Figure 30. Treatment and comparison teachers' description of the overall pacing of their programs.

Perceptions of Achieve3000 and Comparison Programs at Meeting Student Needs

KEY FINDING:

Overall, Achieve3000 teachers appeared more likely than comparison teachers to report that their program was *adequately* or *very adequately* meeting the needs of students.

Lastly, treatment and comparison teachers also rated how adequately their reading instruction met the needs of all students in their classes. The majority of treatment teachers said the Achieve3000 program was *adequately* meeting the needs of all students, while the majority of comparison teachers said their programs were *somewhat adequately* meeting the needs of all students. More comparison teachers than treatment teachers said their program was *not adequately* meeting students' needs (see Figure 31).

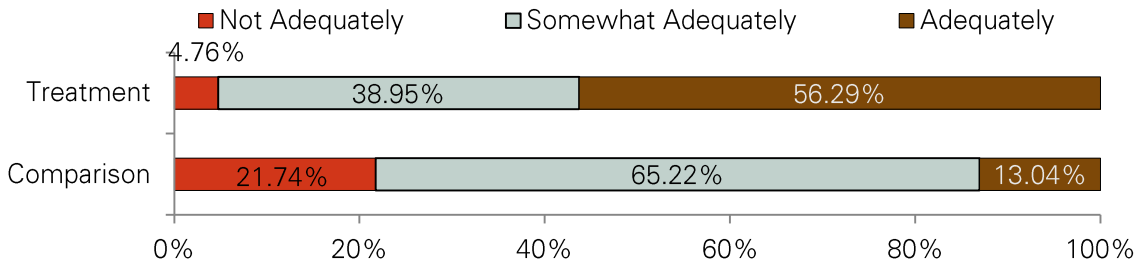


Figure 31. Treatment and comparison teachers' perceptions of reading programs effectiveness in meeting the needs of all students in their classes.

KEY FINDING:

More Achieve3000 than comparison teachers said their program *adequately* or *very adequately* supported students at all levels. More Achieve3000 teachers also said their program helped students build academic vocabulary, comprehend complex text, and critically evaluate informational text. More comparison teachers said their program effectively supported reading fluency.

Treatment and comparison teachers rated their perceptions about how well their programs met the needs of specific subgroups of students. For "below-level" readers, most treatment students said the program was either *adequately* or *very adequately* meeting student needs, while most comparison teachers (52.17%) said that their programs were *neither inadequate nor adequate*. For "on-level" and "advanced-level" readers, most teachers (treatment and comparison) reported that their programs were *adequate* or *very adequate* at meeting students' needs (see Table 28). Overall, treatment teachers appeared more likely than comparison teachers to report that the Achieve3000 program was *adequately* or *very adequately* meeting student needs.

Table 28. Treatment (T) and Comparison (C) Teachers' Perceptions about Meeting Needs of Specific Below-level, On-level, and Advanced-level Students

Student Needs		Very Inadequate	Inadequate	Neither Inadequate nor Adequate	Adequate	Very Adequate	Not applicable
Below-level Readers	T	0.00%	7.58%	14.99%	44.65%	31.96%	0.08%
	C	0.00%	0.00%	52.17%	8.70%	39.13%	0.00%
On-level Readers	T	0.00%	0.66%	7.06%	52.71%	38.26%	1.31%
	C	0.00%	8.70%	0.00%	82.61%	8.70%	0.00%
Advanced Readers	T	0.00%	0.66%	7.77%	43.80%	45.12%	2.64%
	C	0.00%	34.78%	4.35%	52.17%	4.35%	0.00%

For "English language learners" and "special education students," most treatment teachers reported that the Achieve3000 program was *adequately* meeting student needs, while comparison teachers most often reported that their programs were *inadequate* at meeting student needs. It appears that treatment teachers rated their program as more *adequately*

meeting the needs of "special education" and "English language learners" than comparison teachers (see Table 29).

Table 29. Treatment (T) and Comparison (C) Teachers' Perceptions about Meeting Needs of English Language Learners and Special Education Students

Student Needs		Very Inadequate	Inadequate	Neither Inadequate nor Adequate	Adequate	Very Adequate	Not applicable
English Language Learners	T	0.00%	7.14%	19.68%	40.77%	16.90%	15.51%
	C	8.70%	47.83%	21.74%	13.04%	0.00%	8.70%
Special Education Students	T	0.00%	13.60%	17.25%	34.66%	23.05%	11.44%
	C	17.39%	43.48%	21.74%	13.04%	0.00%	4.35%

Perceived Impacts of Achieve3000 and Comparison Programs on Students' Skills

On the online weekly logs and the one-time comparison teacher survey, evaluators also asked treatment and comparison teachers to rate their program's effectiveness at increasing students' skills in reading, using a 5-point Likert scale (*very ineffective, ineffective, neither ineffective nor effective, effective, very effective*). The majority of treatment and comparison teachers rated their programs as *effective* at improving students' skills (see Table 30).

Table 30. Treatment (T) and Comparison (C) Teachers' Perceptions about Program Impacts on Students' Skills

Student Skills		Very Ineffective	Ineffective	Neither Ineffective or Effective	Effective	Very Effective	Not applicable
Reading Fluency	T	0.00%	5.05%	27.96%	56.31%	0.00%	10.68%
	C	0.00%	30.43%	8.70%	56.52%	4.35%	0.00%
Building Academic Vocabulary	T	0.00%	2.86%	16.95%	77.57%	0.00%	2.63%
	C	0.00%	17.39%	17.39%	60.87%	4.35%	0.00%
Comprehending Complex Text	T	0.00%	0.95%	15.51%	81.62%	0.00%	1.90%
	C	0.00%	17.39%	17.39%	56.52%	8.70%	0.00%
Critically Evaluating Informational Text	T	0.00%	1.66%	19.00%	77.20%	0.00%	2.14%
	C	0.00%	13.04%	30.43%	47.83%	8.70%	0.00%

Summary of Teacher Perceptions

Overall, treatment teachers found Achieve3000 program components useful and described many benefits to the program including differentiation, less time for lesson preparation, and positive effects on student engagement and student growth. However, some treatment teachers were frustrated with the monotony of the program routine, the amount of time the program took away from their core curriculum, the brevity of the training, program navigation, and technology issues. Many teachers offered suggestions for improvement such as improving teacher tools, adding visuals for vocabulary, improving various digital components, and navigation features.

Many comparison teachers relied heavily on supplemental materials because core materials were unengaging or outdated. Though comparison teachers struggled with finding interesting materials to meet all students' needs, they were generally happy with their students' achievements.

A comparison of teachers' perceptions of program effectiveness suggests that Achieve3000 had higher student engagement; the appropriate amount of materials to cover; more adequate support for students on all levels; and greater support for building academic vocabulary, comprehending complex text, and critically evaluating informational text. Teachers' perceptions suggest that comparison programs more effectively supported reading fluency.

Summary and Discussion

This randomized control trial studied the efficacy of Achieve3000 at improving reading achievement among third-, sixth-, and ninth-grade students. The study also examined the degree to which teachers implemented the program with fidelity, as well as their perceptions regarding Achieve3000. Magnolia Consulting conducted this independent evaluation study across 16 schools in four districts during the 2014/15 school year.

Implementation measures for this study included teacher self-reported online weekly implementation logs, observation data collected by evaluators, and student usage data compiled by the Achieve3000 program. Overall, treatment teachers reported on their weekly logs that they: (a) implemented the Achieve3000 program on average 1.86 days per week; (b) implemented the program for 88.43 minutes each week; and (c) implemented at least one Achieve3000 lesson per week. Based on the implementation fidelity calculations from the weekly log data, observations, and student usage reports, the implementation grand mean for this study was 71% (out of a possible 100%, which would indicate perfect fidelity). This indicated that treatment teachers implemented the program 29% less than the minimum requirements specified in the implementation guidelines. For this study, implementation was affected by teacher challenges with the program, such as technology issues and insufficient training, as well as various real-world implementation challenges (i.e. competing curriculum requirements, school activities, sick days, holidays, and weather delays) which were reported on 23.50% of the weekly logs.

Evaluators compared treatment teacher implementation of Achieve3000 to comparison teacher implementation of their regular ELA programs. Comparison teachers reported using various literacy programs for more days per week than treatment teachers. However, although they used their programs for more days per week, they used them for less time per week than treatment teachers reported using Achieve3000. Comparison teachers reported planning and preparing for a longer period of time than treatment teachers, and they reported using more supplemental materials.

Evaluators used several types of analyses to examine treatment students' learning gains over the study period. These included multilevel modeling analyses and calculation of effect sizes to examine learning gains as evidenced by the GMRT-4 and LevelSet, as well as other parametric tests to explore outcome data provided by the LevelSet. Multilevel modeling analyses indicated that on average, students who used Achieve3000 during the study period demonstrated statistically significant and substantively important gains as evidenced by their performance on the GMRT-4 Vocabulary, Reading Comprehension, and Total Tests, as well as the LevelSet Lexile assessment. More than 50% of the treatment students met or exceeded their expected Lexile levels. By the end of the study, Achieve3000 users were generally more likely to be classified as on or above the LevelSet college and career benchmark than they were at the beginning of the study. Additionally, analyses revealed positive relationships between treatment students' performance on Achieve3000 activities and learning gains on the GMRT-4 Reading Comprehension, GMRT-4 Total Reading and LevelSet Lexile levels, but no statistically significant relationship between their afterschool use of Achieve3000 and learning gains.

Evaluators also examined the degree to which teachers' implementation fidelity of Achieve3000 was associated with learning gains among their students. Findings revealed that although teachers who implemented the program with higher fidelity generally had students who scored higher on all posttest assessments, the relationship was only statistically significant for the LevelSet Lexile gains. Therefore, overall, when teachers implemented Achieve3000 with relatively higher fidelity, they tended to have students who demonstrated greater Lexile gains over the study period. It is important to note that this finding applies to implementation fidelity ranges observed in this study (35.99% to 96.25% out of a possible 100%). It is unclear how other implementation levels would relate to student learning.

One of the main purposes of this study was to determine the impact of Achieve3000 on reading achievement by comparing end-of-study GMRT-4 performance among students who used Achieve3000 and students who used their schools' typical ELA programs. First, evaluators conducted multilevel modeling analyses across grades, and these were considered the main analyses. Next, evaluators conducted exploratory analyses to examine impacts within each grade separately. The main analyses that compared reading performance by study condition revealed that Achieve3000 did not have a statistically significant impact on posttest GMRT-4 Vocabulary performance, but it did have a statistically significant positive impact on posttest Reading Comprehension and Total Reading scores. Although none of the effect sizes (i.e., 0.12, 0.22, and 0.20) was considered substantively important based on the WWC standards, the effect sizes for Reading Comprehension and Total Reading approached the 0.25 threshold for substantive importance. The WWC improvement indices indicate that for this study, an average Achieve3000 user would rank 5 percentile points higher than an average comparison group student on the GMRT-4 Vocabulary posttest, 9 percentile points higher than an average comparison group student on the GMRT-4 Reading Comprehension test, and 8 percentile points higher than an average comparison group student on the GMRT-4 Total Reading test.

The within-grade exploratory analyses suggested that impacts varied by grade. (These subgroup analyses had less statistical power to detect program effects than the main analyses.) More specifically, for third grade, there were no statistically significant or substantively important impacts (effect sizes were -0.02, 0.02, and 0.06). The improvement indices suggested that a typical third-grade treatment-group student would rank 1 percentile point lower on the GMRT-4 vocabulary test, 1 percentile point higher on the GMRT-4 Reading Comprehension test, and 2 percentile points higher on the GMRT-4 Total Reading test than an average comparison group student. For sixth grade, none of the impacts were statistically significant, and the effect sizes of 0,21, 0.22 and 0.22 all approached the WWC threshold for being substantively important. Additionally, the WWC improvement indices indicate that this study's average sixth-grade Achieve3000 user would rank 8 percentile points higher than an average sixth-grade comparison group student on the GMRT-4 Vocabulary posttest, 9 percentile points higher than an average sixth-grade comparison group student on the GMRT-4 Reading Comprehension test, and 9 percentile points higher than an average sixth-grade comparison group student on the GMRT-4 Total Reading test. Ninth-grade findings showed that although the differences by study condition were not statistically significant (likely because of the reduced sample sizes for these subgroup analyses), all of the corresponding effect sizes (i.e., 0.28, 0.51, and 0.44) exceeded the WWC threshold for substantive importance. The WWC improvement indices indicate that this study's average ninth-grade Achieve3000 user would rank 11percentile points higher than an average ninth-grade comparison group student on the

GMRT-4 Vocabulary posttest, 19 percentile points higher than an average ninth-grade comparison group student on the GMRT-4 Reading Comprehension test, and 17 percentile points higher than an average ninth-grade comparison group student on the GMRT-4 Total Reading test. The substantively important effect sizes suggest that ninth-grade students who used Achieve3000 during the study period out-performed comparison students who used their schools' typical literacy programs.

Exploratory analyses examining the impact of Achieve3000 on reading achievement among ELL students revealed no statistically significant differences or substantively important effect sizes by study condition. The effect sizes for Reading Vocabulary, Reading Comprehension, and Total Reading (i.e., -0.07, -0.06, and -0.05) corresponded to WWC improvement indices of -3, -2, and -2 percentile points, respectively. These findings, which should be interpreted with caution given the small sample size of this ELL subgroup, suggest that on average ELL students who used Achieve3000 performed similarly to comparison-group ELL students who used their schools' typical literacy programs.

In addition to measuring program implementation and findings regarding student learning, evaluators collected treatment teachers' perceptions of the program on weekly logs and the majority of the findings were positive. Treatment teachers found the Achieve3000 program components useful and comprehensive, and described many benefits to the program including differentiation, less time for lesson preparation, and positive effects on student engagement and student achievement. However, some treatment teachers had difficulties with program navigation, technology issues, and not being able to meet the needs of all students. Teachers suggested that the program could be improved by adding more teacher trainings earlier in the school year, additional program features, and reduced assessment time. Most teachers said they would use the program again next year, but many teachers said they would implement it differently.

To compare treatment teachers' perceptions to comparison teachers, evaluators collected a one-time comparison teacher survey. Analysis of treatment and comparison teachers' log and survey data revealed that Achieve3000 users had higher student engagement, an appropriate amount of materials to cover and more applicable pacing than comparison programs. According to study teachers, Achieve3000 more adequately or very adequately supported below-level, on-level, and advanced-level readers, English Language Learners and special education students than comparison programs. For student skills, Achieve3000 more effectively supported building academic vocabulary, comprehending complex text and critically evaluating informational text than comparison programs while comparison programs more effectively supported reading fluency.

Study Limitations

This evaluation study had many strengths, including its randomized design, use of multiple measures, inclusion of multiple school districts, and rigorous analyses. However, it also had limitations. First, schools were not eligible to participate in the study if they did not have adequate technology capacity and infrastructure, including classroom computers or laptops or sufficient availability of computer lab time to meet the required implementation guidelines of 90 minutes per week. Thus, findings from this study are only generalizable to schools that have sufficient technology access. Recruitment criteria also required at least two

ELA teachers per grade to enable random assignment of teachers within grades to study conditions. Many interested high schools only had one ELA teacher at the ninth-grade level, making them ineligible to participate. Additionally, during the study period, two teachers dropped out of the study because they felt overwhelmed by the technology requirements. Thus, it is unclear how teachers with less familiarity and comfort with technology might implement the program and if this would have impacted study findings. Additionally, the loss of these two teachers reduced the sample size and the statistical power to detect program effects. After the initial Achieve3000 training, many teachers requested follow-up trainings, but they were not provided with training in a timely manner. Therefore, it is possible that had teachers received additional training at an earlier date, their implementation fidelity might have increased, which could have impacted student learning. Finally, it is important that readers use caution when interpreting the within-grade subgroup findings because of the relatively small sample sizes that contributed to less statistical power to detect program effects compared to the study's main analyses.

Conclusions and Suggestions for Future Studies

Findings from this randomized control trial study of Achieve3000 showed that treatment teachers generally implemented the program with moderate fidelity and noted that their students used most of the Achieve3000 components and activities. However, teachers reported challenges with implementation on many of their logs and did not often use many administrative components and teacher materials. Results showed that students who used Achieve3000 demonstrated statistically significant and substantively important gains in reading. Furthermore, across grades, the study showed that Achieve3000 had a statistically significant impact (when compared to the comparison group) on GMRT-4 Reading Comprehension and Total Reading. Within grades, the study showed that Achieve3000 impacts on sixth-grade reading were not statistically significant but approached the WWC threshold for being substantively important, and that impacts on ninth-grade GMRT-4 Vocabulary, Reading Comprehension, and Total Reading were substantively important.

This evaluation yielded important findings regarding the efficacy of Achieve3000. It also provided insight into additional topics worth examining that were beyond the scope of this study. For example, because this study suggested that program impacts varied by grade, it would be important for future studies to examine program impacts by grade using larger sample sizes within each grade. It would also be informative to examine the program's impact at other grade levels. Additionally, it would be beneficial for future studies to include teacher interviews to permit a deeper understanding of their implementation and perceptions of the program.

References

Alberta Education. (2010). Making a difference: meeting diverse learning needs with differentiated instruction. Retrieved from http://education.alberta.ca/teachers/resources/cross.aspx

Achieve3000: The power of one. (2011). LevelSet: Online Lexile assessment. Retrieved from http://doc.achieve3000.com/article/LevelSet_Lexile_Assessment.pdf

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

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley.

Borman, G., Slavin, R., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2005). Success for all: First-year results from the National Randomized Field Trial. *Educational Evaluation and Policy Analysis*, *27*, 1–22.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implement Science*, *2*, 40.

Common Core State Standards Initiative. (2010). Standard RL.9-10.3. Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects. Council of Chief State School Officers and the National Governors Association. Retrieved from http://www.corestandards.org/the-standards/english-language-arts-standards/reading-literature-6-12/grade-9-10/.

Erikson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) New York, NY: MacMillan.

Graham, J. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology, 60,* 549–576.

Graham, J., Cumsille, P., and Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka and W. F. Velicer (Eds.), *Research Methods in Psychology* (pp. 87–114). Volume 2 of the *Handbook of Psychology* (I. B. Weiner, Editor-in-Chief). New York: John Wiley & Sons.

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

Huebner, Tracey A. (2010). What research says about differentiated learning. *Educational Leadership,* 67(5), 79-81.

Lesnick, J., George, R., Smithgill, C, & Gwynn, J. (2010). *Reading on grade level in third grade: How is it related to high school performance and college enrollment?* Chicago, IL: Chapin Hall at the University of Chicago.

Mackenzie, N., and Hemmings, B. (2014) Predictors of success with writing in the first year of school. *Issues in Educational Research* 24(1), 41-54.

Moody, S., Vaughn, S. (1997). Instructional grouping for reading. *Remedial and Special Education*, *18*(6), 347-357.

National Institute for Literacy (2008). *Developing early literacy: A report of the National Early Literacy Panel.* Retrieved from: https://www.nichd.nih.gov/publications/pubs/documents/NELPReport09.pdf

Osborne, J. & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, *9*(6). Retrieved August 6, 2014 from http://PAREonline.net/getvn.asp?v=9&n=6.

Parsons, S., Malloy, J., Vaughn, M., and La Croix, L. (2014) A longitudinal study of literacy teacher visioning: Traditional program graduates and Teach for America members. *Literacy Research and Instruction*, 53(2),134-161.

Puma, M., Olsen, R., Bell, S., and Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Rasinski, T., Padak, N., Mckeon, C., Wilfong, L., Friedauer, J., & Heim, P. (2005). Is reading fluency a key for successful high school reading? *Journal of Adolescent & Adult Literacy, 49*(1), 22-27.

Raudenbush, S. Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods.* (2nd ed.). Newbury Park, CA: Sage.

Reis, S., McCoach, D., Little, C., Muller, L. and Kaniscan, R. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Education Research Journal, 48*(2*), 462-501.*

Santangelo, T. and Tomlinson, C. (2012) *Teacher educators' perceptions and use of differentiated instruction practicies: An exploratory investigation.* Journal of the Association of Teacher Educators, 34(4), 309-327.

Schumm, J., Moody, S., & Vaughn, S. (2000). Grouping for reading instruction: Does one size fit all? Journal of Learning Disabilities, 33(5), 477-488.

Schochet, Peter Z. (2008). *Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations* (NCEE 2008-4026). Washington, DC:

National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

U.S. Department of Education, National Center for Education Statistics. (2015). Common core of data. Retrieved from http://nces.ed.gov/ccd/districtsearch/ Watson, S. & Watson, W. (2011). The role of technology and computer-based instruction in a disadvantaged alternative school's culture of learning. *Computers in the Schools, 28*(1), 39-55.

What Works Clearinghouse, U.S. Department of Education, Institute of Education Sciences (2014). What Works Clearinghouse procedures and standards handbook (Version 3.0). Washington, DC: Author. Retrieved from: http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (Expanded 2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Appendix A: Data Preparation

Evaluators followed specific data cleaning and preparation protocols to ensure the most accurate and complete data possible. These protocols included addressing missing data, calculating and examining descriptive statistics, and identifying outliers. As shown in Tables C1 and C2 in Appendix C, missing data rates were moderate for this study. Therefore, evaluators addressed missing data by using multiple imputation. After calculating and examining descriptive statistics, evaluators identified outliers and examined patterns to determine potential causes of outliers. There were a few outliers in the GMRT and LevelSet outcome data, and examination of these outliers indicated that they were legitimate data points and did not have a major impact on analyses. The usage data contained several legitimate outliers as well, and evaluators determined that they accurately represented the variability of students' program use. Therefore, based on research-based recommendations for handling outliers, evaluators included these data in analyses (Osborne & Overbay, 2004).

Appendix B: Achieve3000 Implementation Guidelines

Throughout the study period, teachers should implement *Achieve3000* for at least 90 minutes per week. These 90 minutes should be broken out into two 45-minute sessions per week. Implementation will include a hybrid of student independent work and teacher-directed instruction. The instruction can take place in the regular classroom and/or computer lab. Table 1, below provides an overview of implementation.

Table 1. Overview of implementation guidelines

Targeted Users	• Grades 3, 6 & 9 • All students in study classrooms should participate
Instructional Model	• Student independent and teacher-directed instruction, classroom lab, home study
Usage Model	• Minimum of 2 lessons weekly for a total of 90 minutes per week • Year-end count of about 52 lessons (this accounts for school holidays)
Monitoring and Evaluation Model	• Teachers will run reports to monitor performance and usage • Teachers will review Lexile data monthly • Schools will participate in Achieve3000's contests

Appendix C: Procedures for Calculating Implementation Fidelity

Implementation fidelity scores were comprised of teachers' weekly log reports, observations, and student usage reports. Each of these measures was evaluated against the implementation guidelines for this study by using the minimum requirements as a denominator for each usage variable.

First, evaluators calculated fidelity scores for teacher's weekly implementation log self-reports by comparing them to the study implementation requirements for the following variables:

- Number of days teachers used Achieve3000 each week (minimum two days),
- Number of minutes teachers used Achieve3000 each week (minimum 90 minutes),
- Number of lessons teachers covered each week (minimum of two lessons), and
- If the teacher completed the multiple choice activity question(s) each week.

Second, for the teacher observation data, evaluators used 25 items from the following five constructs to calculate an implementation fidelity score:

- teacher-student interactions,
- equipment and technology,
- procedures associated with the use of Achieve3000,
- Achieve3000 program components, and
- student engagement.

Third, evaluators calculated fidelity scores for usage reports, by comparing scores to the study implementation requirements for the following activities:

- Total valid activities (minimum two activities per week), and
- Passing activities (minimum two activities per week).

To calculate the overall implementation fidelity score for each teacher, teachers' implementation scores from the weekly log reports, observations, and student usage reports were equally weighted (33.33%). To calculate the grand-mean implementation score for this study, the teacher fidelity scores were averaged.

Appendix D: Observation Scores

Table D1: Treatment Teacher Observation Scores

Indictor	Average	Min	Max
Teacher-Student Interactions			
Talk is centered on what students are learning rather than on controlling behavior.	2.43	1	3
Teacher language and encouragement reflect high expectations for students and positive reinforcement.	2.33	1	3
Teacher provides instructional support for the students.	2.38	0	3
Equipment and Technology			
There is enough equipment (such as computers) in the room for each student.	3.00	3	3
The equipment is in working order.	2.95	2	3
There are no problems or difficulties with technology.	2.52	2	3
Procedures Associated with Use of Achieve3000			
Students know and follow established routines for launching their Achieve3000 session.	2.95	2	3
Teacher provides individualized instruction, as necessary.	2.24	0	3
Teacher monitors students as they use Achieve3000 (e.g., the teacher walks around the room to be sure students are on task).	2.76	1	3
Students are able to navigate the Achieve3000 program with little help from their teacher.	3.00	3	3
Teacher provides support and assistance as needed while students use Achieve3000 (e.g models strategies on more rigorous text, responds to student questions, helps students navigate the program)	2.48	0	3
The Achieve3000 session lasts for an appropriate amount of time.	2.57	1	3
Achieve3000 Program Components			
Students respond to the "Before reading poll"	2.56	2	3
Students read the article	3.00	3	3
Students complete the multiple choice activity questions.	3.00	3	3
Students respond to the "After reading poll"	2.33	0	3
Students respond to the "Thought Question"	2.84	0	3
Students complete the 5 step literacy routine	2.95	2	3
Teachers support struggling or gifted & talented readers	1.95	0	3
Teacher discusses/reviews lesson vocabulary	1.76	0	3
Teacher completes whole group discussion	1.89	0	3
Teacher administers an assessment	3.00	3	3
Student Engagement			
Students follow the lessons/activities and transition to and from Achieve3000 activities appropriately.	2.81	2	3
Students are focused on the lesson/activity approximately 90-100% of the session (most students taking part and on task throughout the session).	2.48	1	3
Students show interest in the lesson, materials, and activities.	2.57	2	3

Appendix E: Missing Data Rates

The following tables show missing data rates for each assessment variable and at each time point.

Table D1. Missing GMRT Data Rates by Variable and Time Point.

	Percent Missing
GMRT Vocabulary	
Pretest	4.64%
Posttest	4.05%
GMRT Comprehension	
Pretest	8.50%
Posttest	7.51%[10]
GMRT Total	
Pretest	9.19%
Posttest	8.40%[11]

Table D2. Missing LevelSet Data Rates by Variable and Time Point.

	Percent Missing
LevelSet Reading Lexile	
Pretest	0.78%[12]
Posttest	2.93%

[10] Missing six students' data due the wrong assessment form completed for the Comprehension subtest.
[11] Missing six students' data due the wrong assessment form completed for the Comprehension subtest.
[12] Missing four students' data due to manually adjusted data.

Appendix F: Supporting Tables for Student Performance Results

Unadjusted Pretest and Posttest Means by Study Condition and Grade

These tables show the unadjusted means for each assessment by variable, time point, and condition.

Table E1. Third Grade Students' Unadjusted GMRT Pretest and Posttest Means by Condition.

	Treatment			Comparison			Total		
	Third Grade (*N* =127)			Third Grade (*N* =143)			Third Grade (*N*=270)		
	n	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Vocabulary									
Pretest	127	440.04	40.27	143	442.14	39.29	270	441.15	39.69
Posttest	127	472.91	44.87	143	474.80	42.35	270	473.91	43.48
Comprehension									
Pretest	127	438.57	42.66	143	452.58	44.70	270	445.99	44.23
Posttest	127	473.24	42.61	143	476.83	44.32	270	475.14	43.48
Total									
Pretest	127	438.83	32.96	143	446.74	33.70	270	443.02	33.53
Posttest	127	471.76	38.78	143	474.80	38.59	270	473.37	38.64

Table E1. Sixth Grade Students' Unadjusted GMRT Pretest and Posttest Means by Condition.

	Treatment			Comparison			Total		
	Sixth Grade (*N* =263)			Sixth Grade (*N* =231)			Sixth Grade (*N* =494)		
	n	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Vocabulary									
Pretest	263	503.84	30.22	231	504.07	31.07	494	503.95	30.59
Posttest	263	521.38	37.42	231	515.33	37.02	494	518.55	37.32
Comprehension									
Pretest	263	498.13	31.34	231	499.12	34.93	494	498.59	33.04
Posttest	263	511.31	38.34	231	504.79	36.03	494	508.26	37.38
Total									
Pretest	263	501.43	26.62	231	502.05	28.98	494	501.72	27.73
Posttest	263	516.73	33.16	231	510.49	32.01	494	513.81	32.74

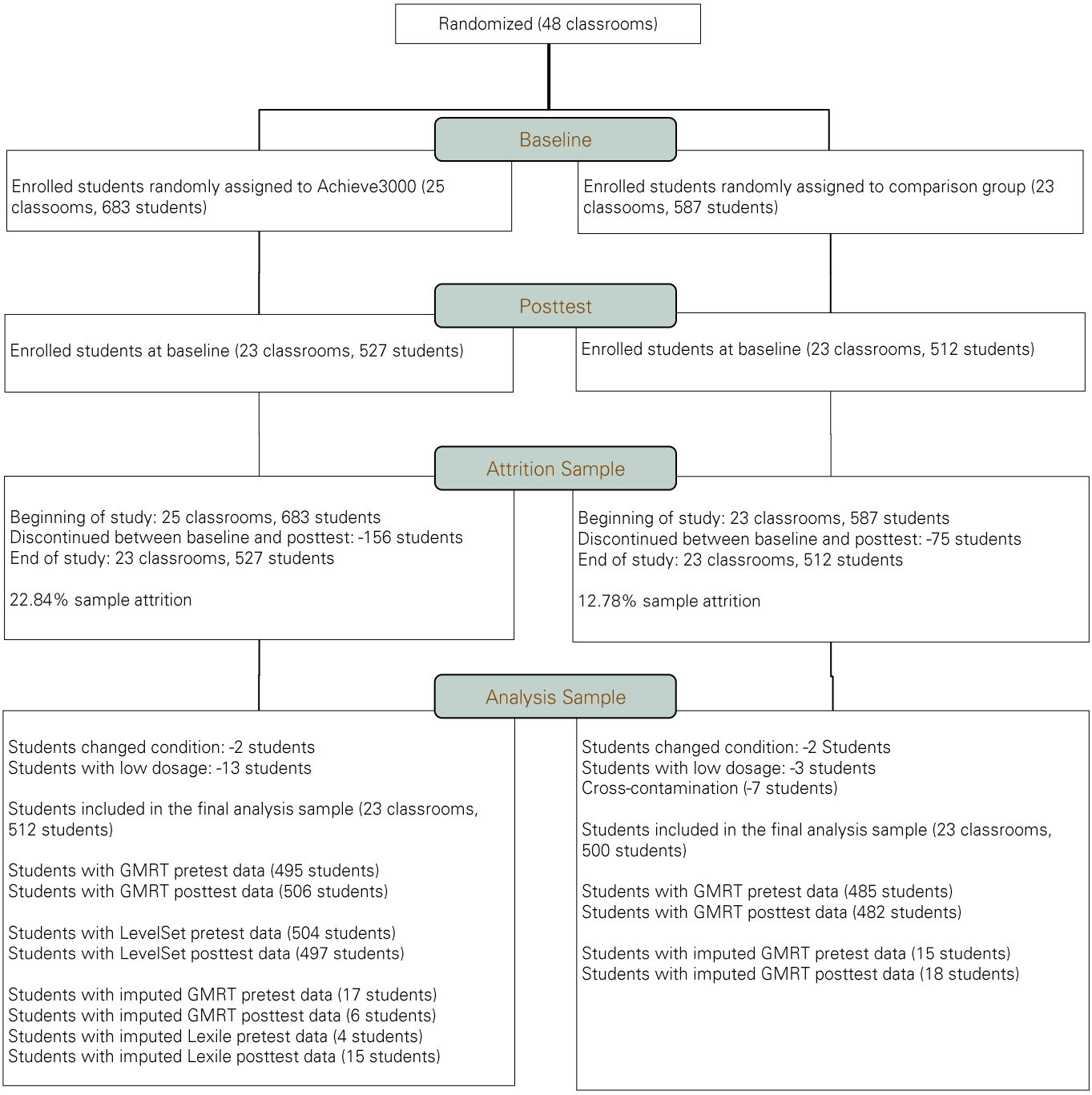
Table E2. Ninth Grade Students' Unadjusted GMRT Pretest and Posttest Means by Condition.

	Treatment Ninth Grade (*N* =122)			Comparison Ninth Grade (*N* =126)			Total Ninth Grade (*N* =248)		
	n	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Vocabulary									
Pretest	122	522.75	28.74	126	522.16	22.80	248	522.45	25.84
Posttest	122	532.85	29.14	126	527.35	32.52	248	530.06	30.97
Comprehension									
Pretest	122	508.64	33.72	126	515.37	34.77	248	512.06	34.36
Posttest	122	536.19	30.57	126	519.33	36.07	248	527.62	34.46
Total									
Pretest	122	518.34	26.23	126	521.14	25.57	248	519.77	25.88
Posttest	122	537.16	27.09	126	526.13	30.81	248	531.56	29.50

Table E3. Unadjusted LevelSet Reading Lexile Pretest and Posttest Means by Grade.

	Third Grade (*N* =127)			Sixth Grade (*N* =263)			Ninth Grade (*N* =122)			Total (*N* =512)		
	n	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*
Pretest	127	214.28	264.35	263	610.03	252.36	122	831.46	174.29	512	564.63	325.02
Posttest	127	383.75	248.05	263	730.29	258.43	122	864.48	195.83	512	676.31	299.53

Appendix G: CONSORT



Appendix H. School-Level Characteristics

	District A			District B					
	School 1			School 1			School 2		
	Comparison	Treatment	Total	Comparison	Treatment	Total	Comparison	Treatment	Total
Number of Students									
Third Grade	15	17	32	47	25	72	-	-	-
Sixth Grade	14	15	29	-	-	-	47	57	104
Ninth Grade	-	-	-	-	-	-	-	-	-
School Totals									
Classrooms	2	2	4	2	1	3	2	2	4
Number of students	29	32	61	47	25	72	47	57	104
Gender Among Participants									
Female	37.93%	37.50%	37.70%	40.43%	56.00%	45.83%	65.96%	49.12%	56.73%
Male	62.07%	62.50%	62.30%	59.57%	44.00%	54.17%	34.04%	50.88%	43.27%
Ethnicity Among Participants									
Hispanic or Latino	65.52%	37.50%	50.82%	78.72%	72.00%	76.39%	76.60%	50.88%	62.50%
Not Hispanic or Latino	34.48%	62.50%	49.18%	21.28%	28.00%	23.61%	23.40%	49.12%	37.50%
Ethnicity Among Participants									
White	51.72%	40.63%	45.90%	78.72%	80.00%	79.17%	82.98%	57.89%	69.23%
Black or African American	27.59%	43.75%	36.07%	8.51%	16.00%	11.11%	10.64%	33.33%	23.08%
Asian	3.45%	6.25%	4.92%	8.51%	0.00%	5.56%	6.38%	5.26%	5.77%
Two or more races or Other	17.24%	9.38%	13.11%	4.26%	4.00%	4.17%	0.00%	3.51%	1.92%
Limited English Proficiency Among Participants									
LEP	58.62%	50.00%	54.10%	12.77%	24.00%	16.67%	40.43%	14.04%	25.96%
Non-LEP	41.38%	50.00%	45.90%	87.23%	76.00%	83.33%	59.57%	85.96%	74.04%
Special Education Among Participants									
Special Education	20.69%	6.25%	13.11%	4.26%	20.00%	9.72%	0.00%	15.79%	8.65%
Non-Special Education	79.31%	93.75%	86.89%	95.74%	80.00%	90.28%	100.00%	84.21%	91.35%
Free/Reduced Price Lunch Among Participants									
Free/Reduced Lunch	100.00%	100.00%	100.00%	85.11%	80.00%	83.33%	93.62%	89.47%	91.35%
Non-Free/Reduced Lunch	0.00%	0.00%	0.00%	14.89%	20.00%	16.67%	6.38%	10.53%	8.65%
Section 504 Among Participants									
Section 504	0.00%	3.13%	1.64%	0.00%	0.00%	0.00%	2.13%	0.00%	0.96%
Non-Section 504	100.00%	96.88%	98.36%	100.00%	100.00%	100.00%	97.87%	100.00%	99.04%

	District B			District C [13]					
	School 3			School 1			School 3		
	Comparison	Treatment	Total	Comparison	Treatment	Total	Comparison	Treatment	Total
Number of Students									
Third Grade	-	-	-	21	24	45	20	21	41
Sixth Grade	-	-	-	-	-	-	-	-	-
Ninth Grade	75	80	155	-	-	-	-	-	-
School Totals									
Classrooms	3	3	6	1	1	2	1	1	2
Number of students	75	80	155	21	24	45	20	21	41
Gender Among Participants									
Female	52.00%	38.75%	45.16%	47.62%	41.67%	44.44%	60.00%	47.62%	53.66%
Male	48.00%	61.25%	54.84%	52.38%	58.33%	55.56%	40.00%	52.38%	46.34%
Ethnicity Among Participants									
Hispanic or Latino	69.33%	63.75%	66.45%	14.29%	8.33%	11.11%	20.00%	33.33%	26.83%
Not Hispanic or Latino	30.67%	36.25%	33.55%	85.71%	91.67%	88.89%	80.00%	66.67%	73.17%
Ethnicity Among Participants									
White	80.00%	73.75%	76.77%	66.67%	75.00%	71.11%	25.00%	28.57%	26.83%
Black or African American	16.00%	17.50%	16.77%	19.05%	20.83%	20.00%	65.00%	47.62%	56.10%
Asian	4.00%	7.50%	5.81%	9.52%	4.17%	6.67%	5.00%	0.00%	2.44%
Two or more races or Other	0.00%	1.25%	0.65%	4.76%	0.00%	2.22%	5.00%	23.81%	14.63%
Limited English Proficiency Among Participants									
LEP	10.67%	8.75%	9.68%	9.52%	0.00%	4.44%	10.00%	14.29%	12.20%
Non-LEP	89.33%	91.25%	90.32%	90.48%	100.00%	95.56%	90.00%	85.71%	87.80%
Special Education Among Participants									
Special Education	1.33%	16.25%	9.03%	-	-	-	-	-	-
Non-Special Education	98.67%	83.75%	90.97%	-	-	-	-	-	-
Free/Reduced Price Lunch Among Participants									
Free/Reduced Lunch	77.33%	76.25%	76.77%	-	-	-	-	-	-
Non-Free/Reduced Lunch	22.67%	23.75%	23.23%	-	-	-	-	-	-
Section 504 Among Participants									
Section 504	2.67%	0.00%	1.29%	-	-	-	-	-	-
Non-Section 504	97.33%	100.00%	98.71%	-	-	-	-	-	-

[13] District C provided classroom level data only for Special Education, Free/Reduced Price Lunch and Section 504.

	District C [14]								
	School 3			School 4			School 5		
	Comparison	Treatment	Total	Comparison	Treatment	Total	Comparison	Treatment	Total
Number of Students									
Third Grade	40	40	80	-	-	-	-	-	-
Sixth Grade	-	-	-	18	20	38	19	46	65
Ninth Grade	-	-	-	-	-	-	-	-	-
School Totals									
Classrooms	2	2	4	1	1	2	1	2	3
Number of students	40	40	80	18	20	38	19	46	65
Gender Among Participants									
Female	42.50%	45.00%	43.75%	72.22%	50.00%	60.53%	47.37%	41.30%	43.08%
Male	57.50%	55.00%	56.25%	27.78%	50.00%	39.47%	52.63%	58.70%	56.92%
Ethnicity Among Participants									
Hispanic or Latino	12.50%	15.00%	13.75%	38.89%	50.00%	44.74%	47.37%	43.48%	44.62%
Not Hispanic or Latino	87.50%	85.00%	86.25%	61.11%	50.00%	55.26%	52.63%	56.52%	55.38%
Ethnicity Among Participants									
White	52.50%	42.50%	47.50%	50.00%	50.00%	50.00%	63.16%	65.22%	64.62%
Black or African American	25.00%	30.00%	27.50%	33.33%	35.00%	34.21%	21.05%	19.57%	20.00%
Asian	2.50%	12.50%	7.50%	0.00%	0.00%	0.00%	0.00%	4.35%	3.08%
Two or more races or Other	20.00%	15.00%	17.50%	16.67%	15.00%	15.79%	15.79%	10.87%	12.31%
Limited English Proficiency Among Participants									
LEP	2.50%	12.50%	7.50%	11.11%	15.00%	13.16%	5.26%	10.87%	9.23%
Non-LEP	97.50%	87.50%	92.50%	88.89%	85.00%	86.84%	94.74%	89.13%	90.77%
Special Education Among Participants									
Special Education	-	-	-	-	-	-	-	-	-
Non-Special Education	-	-	-	-	-	-	-	-	-
Free/Reduced Price Lunch Among Participants									
Free/Reduced Lunch	-	-	-	-	-	-	-	-	-
Non-Free/Reduced Lunch	-	-	-	-	-	-	-	-	-
Section 504 Among Participants									
Section 504	-	-	-	-	-	-	-	-	-
Non-Section 504	-	-	-	-	-	-	-	-	-

[14] District C provided classroom level data only for Special Education, Free/Reduced Price Lunch and Section 504.

	District C[15]								
	School 6			School 7			School 8		
	Comparison	Treatment	Total	Comparison	Treatment	Total	Comparison	Treatment	Total
Number of Students									
Third Grade	-	-	-	-	-	-	-	-	-
Sixth Grade	22	23	45	-	-	-	-	-	-
Ninth Grade	-	-	-	16	10	26	16	15	31
School Totals									
Classrooms	1	1	2	1	1	2	1	1	2
Number of students	22	23	45	16	10	26	16	15	31
Gender Among Participants									
Female	40.91%	39.13%	40.00%	75.00%	40.00%	61.54%	62.50%	46.67%	54.84%
Male	59.09%	60.87%	60.00%	25.00%	60.00%	38.46%	37.50%	53.33%	45.16%
Ethnicity Among Participants									
Hispanic or Latino	27.27%	30.43%	28.89%	62.50%	30.00%	50.00%	0.00%	26.67%	12.90%
Not Hispanic or Latino	72.73%	69.57%	71.11%	37.50%	70.00%	50.00%	100.00%	73.33%	87.10%
Ethnicity Among Participants									
White	72.73%	73.91%	73.33%	68.75%	80.00%	73.08%	56.25%	53.33%	54.84%
Black or African American	9.09%	8.70%	8.89%	25.00%	10.00%	19.23%	18.75%	26.67%	22.58%
Asian	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	12.50%	13.33%	12.90%
Two or more races or Other	18.18%	17.39%	17.78%	6.25%	10.00%	7.69%	12.50%	6.67%	9.68%
Limited English Proficiency Among Participants									
LEP	13.64%	8.70%	11.11%	12.50%	0.00%	7.69%	6.25%	0.00%	3.23%
Non-LEP	86.36%	91.30%	88.89%	87.50%	100.00%	92.31%	93.75%	100.00%	96.77%
Special Education Among Participants									
Special Education	-	-	-	-	-	-	-	-	-
Non-Special Education	-	-	-	-	-	-	-	-	-
Free/Reduced Price Lunch Among Participants									
Free/Reduced Lunch	-	-	-	-	-	-	-	-	-
Non-Free/Reduced Lunch	-	-	-	-	-	-	-	-	-
Section 504 Among Participants									
Section 504	-	-	-	-	-	-	-	-	-
Non-Section 504	-	-	-	-	-	-	-	-	-

[15] District C provided classroom level data only for Special Education, Free/Reduced Price Lunch and Section 504.

	District B[16]			District D					
	School 9			School 1			School 3		
	Comparison	Treatment	Total	Comparison	Treatment	Total	Comparison	Treatment	Total
Number of Students									
Third Grade	-	-	-	-	-	-	-	-	-
Sixth Grade	-	-	-	30	29	59	29	46	75
Ninth Grade	19	17	36	-	-	-	-	-	-
School Totals									
Classrooms	1	1	2	1	1	2	1	2	3
Number of students	19	17	36	30	29	59	29	46	75
Gender Among Participants									
Female	57.89%	41.18%	50.00%	40.00%	48.28%	44.07%	58.62%	34.78%	44.00%
Male	42.11%	58.82%	50.00%	60.00%	51.72%	55.93%	41.38%	65.22%	56.00%
Ethnicity Among Participants									
Hispanic or Latino	26.32%	11.76%	19.44%	3.33%	3.45%	3.39%	3.45%	8.70%	6.67%
Not Hispanic or Latino	73.68%	88.24%	80.56%	96.67%	96.55%	96.61%	96.55%	91.30%	93.33%
Ethnicity Among Participants									
White	52.63%	41.18%	47.22%	96.67%	79.31%	88.14%	79.31%	78.26%	78.67%
Black or African American	36.84%	41.18%	38.89%	3.33%	3.45%	3.39%	13.79%	10.87%	12.00%
Asian	0.00%	0.00%	0.00%	0.00%	6.90%	3.39%	3.45%	4.35%	4.00%
Two or more races or Other	10.53%	17.65%	13.89%	0.00%	10.34%	5.08%	3.45%	6.52%	5.33%
Limited English Proficiency Among Participants									
LEP	10.53%	0.00%	5.56%	0.00%	3.45%	1.69%	0.00%	4.35%	2.67%
Non-LEP	89.47%	100.00%	94.44%	100.00%	96.55%	98.31%	100.00%	95.65%	97.33%
Special Education Among Participants									
Special Education	-	-	-	33.33%	13.79%	23.73%	0.00%	4.35%	2.67%
Non-Special Education	-	-	-	66.67%	86.21%	76.27%	100.00%	95.65%	97.33%
Free/Reduced Price Lunch Among Participants									
Free/Reduced Lunch	-	-	-	13.33%	13.79%	13.56%	27.59%	19.57%	22.67%
Non-Free/Reduced Lunch	-	-	-	86.67%	86.21%	86.44%	72.41%	80.43%	77.33%
Section 504 Among Participants									
Section 504	-	-	-	3.33%	6.90%	5.08%	3.45%	0.00%	1.33%
Non-Section 504	-	-	-	96.67%	93.10%	94.92%	96.55%	100.00%	98.67%

[16] District C provided classroom level data only for Special Education, Free/Reduced Price Lunch and Section 504.

	District D School 3			Study Totals		
	Comparison	Treatment	Total	Comparison	Treatment	Total
Number of Students						
Third Grade	-	-	-	143	127	270
Sixth Grade	52	27	79	231	263	494
Ninth Grade	-	-	-	126	122	248
School Totals						
Classrooms	2	1	3	23	23	46
Number of students	52	27	79	500	512	1012
Gender Among Participants						
Female	46.15%	48.15%	46.84%	51.20%	43.36%	47.23%
Male	53.85%	51.85%	53.16%	48.80%	56.64%	52.77%
Ethnicity Among Participants						
Hispanic or Latino	3.85%	7.41%	5.06%	39.40%	34.77%	37.06%
Not Hispanic or Latino	96.15%	92.59%	94.94%	60.60%	65.23%	62.94%
Ethnicity Among Participants						
White	76.92%	88.89%	81.01%	70.00%	64.26%	67.09%
Black or African American	9.62%	7.41%	8.86%	18.40%	22.66%	20.55%
Asian	1.92%	3.70%	2.53%	3.80%	5.08%	4.45%
Two or more races or Other	11.54%	0.00%	7.59%	7.80%	8.01%	7.91%
Limited English Proficiency Among Participants						
LEP	3.85%	3.70%	3.80%	13.60%	11.52%	12.55%
Non-LEP	96.15%	96.30%	96.20%	86.40%	88.48%	87.45%
Special Education Among Participants						
Special Education	25.00%	22.22%	24.05%	10.36%	13.85%	12.07%
Non-Special Education	75.00%	77.78%	75.95%	89.64%	86.15%	87.93%
Free/Reduced Price Lunch Among Participants						
Free/Reduced Lunch	17.31%	22.22%	18.99%	62.14%	61.82%	61.98%
Non-Free/Reduced Lunch	82.69%	77.78%	81.01%	37.86%	38.18%	38.02%
Section 504 Among Participants						
Section 504	7.69%	7.41%	7.59%	2.91%	1.69%	2.31%
Non-Section 504	92.31%	92.59%	92.41%	97.09%	98.31%	97.69%