EDUCATION

THE UTAH STATE BOARD OF EDUCATION

Report to the Education Interim Committee

Early Intervention Reading Software Program Report

November 2019

Melanie Durfee
Digital Teaching and Learning Specialist
melanie.durfee@schools.utah.gov

Todd Call
Coordinator for Digital Teaching and Learning
todd.call@schools.utah.gov

Jennifer Throndsen
Director of Teaching and Learning
jennifer.throndsen@schools.utah.gov

Darin Nielsen
Assistant Superintendent of Student Learning
darin.nielsen @schools.utah.gov

STATUTORY REQUIREMENT

U.C.A. Section 53F-4-203 requires the State Board of Education and the contracted independent evaluator to report annually on the results of the evaluation to the Education Interim Committee. The independent evaluator is required to (i) evaluate a student's learning gains as a result of using the provided early interactive reading software; (ii) for the evaluation, use an assessment not developed by a provider of early interactive reading software; and (iii) determine the extent to which a public school uses the early interactive reading software.

Early Intervention Reading Software Program Report

EXECUTIVE SUMMARY

The Early Intervention Reading Software Program encourages literacy growth and achievement in students in grades K-3. The program addresses early reading through the use of computer-based literacy software which provides individualized instruction designed to supplement students' classroom learning. During the 2018-2019 school year, these software programs were used in 88 local education agencies (LEAs) and by approximately 124,378 students. The schools use the software to build literacy skills for all students in kindergarten and first grade, as well as for intervention with students in second and third grade. The independent evaluation for the 2018-2019 school year is attached.

Utah's Early Intervention Reading Software Program

2018-2019 Program Evaluation Findings

Submitted to the Utah State Board of Education
September 2019

Evaluation and Training Institute
100 Corporate Pointe, Suite 387
Culver City, CA 90230

ADA Compliant 5.19.2020

Table of Contents

List of Figures

List of Tables

Acknowledgements

The Evaluation and Training Institute (ETI) thanks Todd Call (Coordinator, Digital Teaching and Learning) and Melanie Durfee (Specialist, Digital Teaching and Learning) from the Utah State Board of Education (USBE) for their ongoing collaboration throughout this evaluation project.

We also acknowledge Kristen Campbell and Malia Mcllvenna at USBE for all their efforts preparing and transferring the student data used for the analyses.

Finally, the software vendor representatives played a key role in helping us understand their software programs, sharing their data, and working patiently with us to prepare the data in a consistent and streamlined format. In particular, we give special thanks to Nari Carter from Imagine Learning, Haya Shamir from Waterford, Sarah Franzén from Lexia, Corey Fitzgerald and Joel Kongaika from Curriculum Associates. Each of these individuals provided necessary data from their products that were used to complete the evaluation project.

Acronyms

BOY	Beginning-of-Year
Cntrl	Control group/non-program group
EISP	Early Intervention Software Program
EOY	End-of-Year
ES	Effect Size
ETI	Evaluation and Training Institute
IS	Insufficient Sample
LEA	Local Education Agency
NS	Non-significant statistical coefficient
OLS	Ordinary Least Squares analyses
Tr.	Treatment group/program group
USBE	Utah State Board of Education

Executive Summary

Evaluation Purpose

The Early Intervention Software Program (EISP) was designed to increase the literacy skills of all students in K-1 and struggling readers in Grades 2-3 through the use of adaptive computer-based literacy software programs. For the 2018-2019 school year, the program provided Utah's Local Education Agencies (LEAs) with an option to select among four adaptive computer-based literacy software programs: Imagine Learning, Curriculum Associates (i-Ready), Lexia® (Core5), and Waterford. The Evaluation and Training Institute (ETI), the external evaluator of EISP, studied two aspects of the program: 1) students use of the program during the school year ("program implementation"); and, 2) the effects the program had on increasing students' literacy achievement ("program impacts"), including program effects across all four software programs (program-wide) and between each software vendor (vendor-specific).

Program Enrollment and Implementation Findings

During the 2018-2019 school year, EISP was implemented in 88 Local Education Agencies (LEAs) and by 124, 378 students throughout the state of Utah. Core5 was used by the most students (65,109), followed by Imagine Learning (41,305), i-Ready (10,016), and Waterford (7,948). State-wide program implementation set the stage for large numbers of students to receive program benefits, however, it was important for students to meet minimum usage requirements (set by program vendors) in order for the program to impact students' literacy achievement. To that end, program vendors provided LEAs with recommendations on how many minutes per week students should have used the program, on average, as well as the total number of weeks the program should be used. Recommendations for average weekly and weeks of use was known as the minimum recommended program dosage. The implementation study was designed to determine the extent to which students met each vendors' recommendations for minimum average weekly use and total weeks of use. A majority of students (63-83%) using three of the four software programs (Core5, Imagine Learning and Waterford), met these requirements for total weeks of use, which ranged from 18-30 weeks, and was an indication of students consistent use of the software. Although a majority of students across programs used the software for the recommended total weeks, fewer students met their respective vendors recommended minutes per week. Among the four vendors, Core5 was the only vendor with more than 50 percent of their students who met the minimum recommended weekly minutes of use, on average.

Program-wide Impacts Findings

The program had a positive impact on students' literacy development in kindergarten, regardless of their program dosage, and in kindergarten and first grade for students who used the program based on vendors' recommended dosage, as measured by students' Acadience Reading composite scores. Kindergarten and first grade program students who met vendors recommendations also scored higher than their control group counterparts on all literacy skill

areas measured for their grade level. However, the mean score differences for individual literacy skills were small (less than five points across literacy skills) and the effect sizes were less than .26, which is smaller than the average effect size seen in similar intervention programs. There were no statistically significant positive effects for students in second or third grade for Acadience Reading composite scores or for any literacy skill area.

In addition to studying program impacts for all students using the software, we examined how the program may have affected certain subgroups of students. We found statistically significant program effects in kindergarten, but not for other grade levels. Kindergarten EISP students who were identified as low-income, special education, English Language Learners (ELL), and those who attended a Title 1 school all had higher predicted means scores than non-EISP low-income, special education, ELL, and Title 1 matched control students. Special education program students had the highest difference in mean scores, with special education program students scoring approximately 8 points higher than special education non-program students on the Acadience Reading composite scores.

We also explored the relationship between program dosage and literacy outcomes in the program-wide section of the report by examining how the predicted end-of-year mean composite scores deviated from the lowest, middle, and highest dosage groups. In general, the effectiveness of the program increased in strength as dosage increased from the lowest to highest dosage group, but the biggest jump in mean scores occurred from the lowest to middle group and not the middle to highest group. This trend was observed for all three grade levels.

Vendor Impacts Findings

We also studied the impacts of individual program vendors on students' literacy achievement. All four vendors had a positive impact on students end-of-year composite scores in kindergarten (Waterford, Imagine Learning, Core5, and i-Ready), followed by two vendors in first grade (Waterford; i-Ready), one vendor in second grade (Waterford), and two vendors in third grade (Core5, i-Ready). To measure the strength of these effects, we looked at the average effect sizes produced by similar education intervention programs (ES benchmark: .26). In kindergarten, first and second grade the effects were stronger than those found in similar intervention programs for two vendors: i-Ready in K-1 and Waterford in Grade 2.

Discussion & Recommendations

The 2018-2019 program had a positive effect in kindergarten (both looking at the program as a whole, and for a majority of specific vendors), and had mixed effects on students in 1st through 3rd grade, depending on the software vendor, analyses method, and literacy domain. Based on these findings we recommend the program continue to be used in kindergarten. It is more difficult to endorse the program's use with students in first through third grade due to mixed results from year-to-year and the complexities involved with making vendor comparisons (e.g. differences in vendor sample sizes, etc.). With select vendors, however, there were indicators that students in these upper-early grades benefited from the program, so we are recommending

that more data be collected and results reviewed for future cohorts. Future research is needed to increase our understanding of the conditions which lead to improvements in literacy achievement for specific vendors and students. We propose studying additional implementation details and their link to program outcomes in order to make targeted recommendations to improve the efficacy and impacts of the program. For example, past research indicates there may be a link between active teacher involvement in the program and literacy outcomes (Best Practices for Improving Early Intervention Software Programs in Utah Schools, 2017), but we would need to collect data on how schools implement the program beyond dosage to better understand the impact of this link.

Evaluation Purpose & Evaluation Questions

Utah passed legislation in 2012 (HB513) to supplement students' classroom learning with additional reading support in the form of computer-based adaptive reading programs. The intent of the legislation was to increase the number of students reading at grade level each year, and to ensure that students were on target in literacy achievement prior to the end of the third grade. The legislation provided funding to use for the programs with students in kindergarten and in first grade, and as an intervention for students reading below grade level in second and third grade. To participate in the Early Intervention Software Program (EISP), Local Education Agencies (LEAs) submitted applications to the USBE requesting funding for the use of specific reading software programs prior to the start of each school year. Four software vendors provided software and training to schools through the EISP in 2018-2019. The four vendors were (in alphabetical order): Curriculum Associates ("i-Ready"), Imagine Learning, Lexia® ("Core5®"), and Waterford.

The Evaluation and Training Institute (ETI) contracted with the Utah State Board of Education (USBE) to study how the reading software programs were used by schools and the impact they had on students' literacy development. The evaluation included results for the combined impact of all the software programs used in Utah schools ("program-wide" impacts) and a comparison of the relative effects on literacy achievement for each of the software providers ("individual vendor impacts").

This report included findings from the 2018-2019 academic year, the EISP's sixth year of implementation. These findings were intended to help the USBE and Local Education Agencies (LEAs) understand how the program was working, to identify potential areas for program improvement, and to make evidence-based decisions about future iterations of the program.

The following research questions were used to guide our evaluation and organize the findings in this report:

1. Did students use the software as intended?
2. Did the program have an overall effect across all vendors?
3. What interactions between student characteristics and school type effect program impacts?
4. Were there differences in treatment effects among vendors?

In this report we included a description of the EISP and 2018-2019 program enrollment, findings related to each research question and the two study objectives (program implementation and program impacts). A detailed summary of our research methods were included in **Appendix A**. Finally, we discussed the key findings and the study limitations.

Program Enrollment

In 2018-2019, the four EISP software vendors were used in 88 LEAs and 438 schools and by 124,378 students. While the EISP was intended for second and third grade students reading below grade level (referred to as "intervention students" throughout the report), some educators implemented the program with their entire class, and in these instances, students reading at grade level ("non-intervention students") also had access to the software programs. As a result, we have provided enrollment information for both non-intervention and intervention students in second and third grade in **Tables 1-2**, so readers may understand how the program was implemented in practice and as intended. As depicted in **Table 1**, Core5 was the most frequently used program (188 schools, 65,000+ students), while Waterford was used with the fewest students among the four vendors (7,948 students).

Table 1. 2018-2019 Program Enrollment Overview

Program	LEAs	Schools	Students	
			All K 3	All K 1 & 2 3 Intervention
Waterford	28	67	7,948	7,155
Imagine Learning	50	191	41,305	37,135
Core5	42	223	65,109	39,384
i-Ready	15	43	10,016	5,466
Total	88	438	124,378	89,140

Note. Some LEAs/schools used multiple vendors. Totals represent unique cases of LEAs and schools. Data source: vendor data in K-1 and vendor data merged with Acadience Reading data in Grades 2-3.

Table 2 presents 2018-2019 program enrollment by vendor and grade level. Student participation by grade varied by program. Imagine Learning and Core5 had a fairly even distribution of students across Grades K-3, while Waterford was used more frequently in earlier grades, and i-Ready was used more frequently in the upper-early grades.

Table 2. 2018-2019 Program Enrollment by Vendor and Grade

Program	Kinder	1st	2nd		3rd	
			All	Intervention	All	Intervention
Waterford	3,469	3,102	1,183	489	194	95
Imagine Learning	10,148	12,100	10,425	4,656	8,632	4,273
Core5	14,256	15,788	17,614	4,574	17,451	4,766
i-Ready	1,462	2,236	3,201	888	3,117	880
Total	29,335	33,226	32,423	10,607	20,762	10,014

Note. Data source: vendor data in K-1 and vendor data merged with Acadience Reading data in Grades 2-3. Grades 2-3 intervention students included those with scores below benchmark for their grade at the beginning of year.

The USBE contracted with vendors to implement their programs with specific grades, depicted in Table 3. Vendors allowed LEA's to use their software with students outside of the contracted grades, which resulted more students being exposed to the program.

Table 3. Grade Levels in Vendor Contracts

Program	Grades in Contract	Grades Served
Waterford	K-1	K-3
Imagine Learning	K-3	K-3
Core5	K-3	K-3
i-Ready	2-3	K-3

Program Implementation Findings

It is important for evaluators to study program implementation prior to measuring the program impacts on student learning, and with increased understanding of how a program was implemented, conclusions made about the program impacts can be more meaningful. For the EISP, the most important aspect of program implementation was dosage, which is how much of the program a student received during the school year, as students must use the program for a long enough period of time for it to have an impact on their literacy skill development. The program implementation findings in this section are reported for general education students in K-1, and intervention students in Grades 2-3. Intervention students were defined as being below benchmark reading levels at the beginning of the school year.

Each vendor provided recommendations for using the software program in order for it to have an impact on students' literacy achievement (**Table 4**). Recommended weekly use ranged from 20 minutes to 80 minutes of use per week, and suggested weeks of use ranged from 18 to 30 weeks. For LEAs to continue to receive program funding, the state required that at least 80 percent of the students within a school meet 80% of vendors' average use or weeks of use recommendations within two years of implementation[1].

Table 4. Vendor 2018-2019 Minimum Dosage Recommendations

Program	Kindergarten	First Grade	Second Grade	Third Grade	Suggested
Waterford	60 min/week	80 min/week	80 min/week	80 min/week	28 weeks
Imagine Learning	40 min/week	45 min/week	45 min/week	45 min/week	18 weeks
Core5	20 minutes to 60 min/week*	20 minutes to 60 min/week*	20 minutes to 60 min/week*	20 minutes to 60 min/week*	20 weeks
i-Ready	30 minutes	45 minutes	45minutes	45minutes	26-30 weeks

Note. Core5 based its usage recommendations on student performance, and students who were working below grade level were assigned usage recommendations that were greater than those who worked at or above grade level.

In the following sections, we explored the differences in usage across software programs and grade levels in order to better understand the nuances of program implementation based on these factors. We used the recommendations provided by each program vendor on average weekly use and total weeks of use to determine if students were using the program as it was intended. A more detailed summary of student use is included in **Appendix F**.

[1] ETI submitted a separate report to the USBE on school level fidelity.

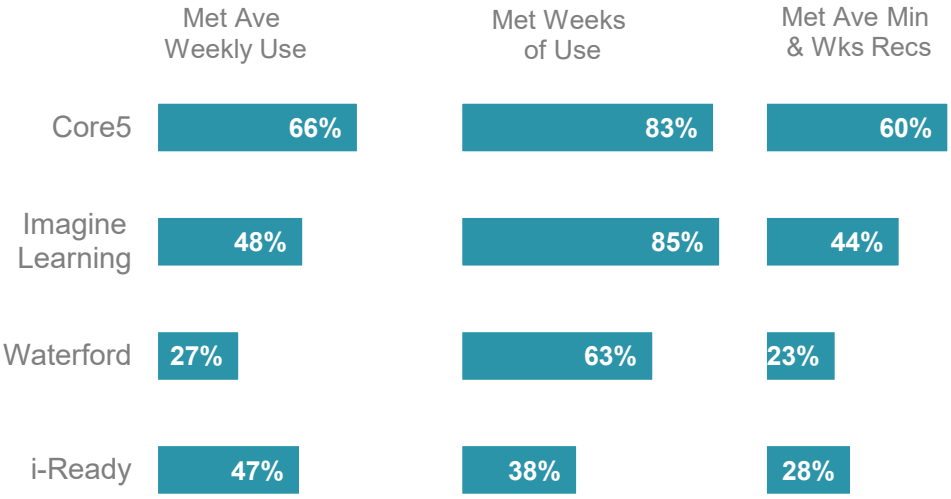
Did students use the software as intended?

Figure 1 depicts the percentage of students who met vendors average weekly minutes of use, total weeks of use, and both minutes and total weeks recommendations[2]. More students met the weeks of use targets (63-83%) versus the average weekly minutes targets (27-66%) or combined targets (23%-60%) among three out of the four vendors: Waterford, ImagineLearning and Core5. This finding suggests that most LEAs were facilitating students' use of the software on a weekly basis and for the minimum number of weeks that vendors' recommended.

While LEAs made sure that their students used the software regularly, it was more difficult for them to meet vendors' weekly minutes of use targets.[3] Among the four vendors, there was one vendor, Core5, in which at least half of their students used the software for the recommended minutes per week, on average. i-Ready was the only vendor in which a higher percentage of students met the average weekly targets over the minimum weeks of use targets.

The percentage of students who met vendors' recommendations for both average minutes and total weeks is presented in the last column of **Figure 1**. These students used the programs as intended on both aspects of dosage: weekly minutes and total weeks. Over half of the students who used Core5 met both recommendations, and almost half of Imagine Learning reached this goal.

Figure 1: Students who met vendors minimum dosage recommendations



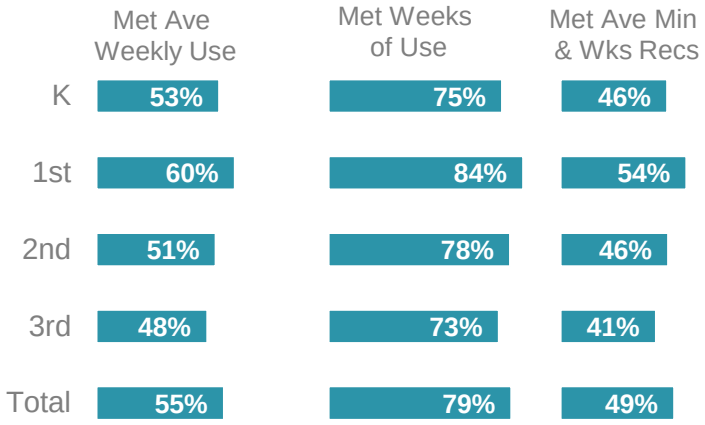
N: i-Ready (5,466); Waterford (7155); IL (28,957); Core5 (39,384)
Data source: vendor data in K-1 and vendor data merged with Acadience Reading in Grades 2-3.

[2] Vendor recommendations for total weeks of use ranged from 18-30 weeks.

[3] Vendor recommendations for average minutes per week ranged from 45-80 minutes. Core5 had lower recommendations for non-intervention students: 20 minutes per week.

Figure 2 provides an overview of program use within each grade. Forty-eight to 60 percent of students met the average minutes recommendations across grades, while 73 to 84% met the minimum weeks requirements. Fewer students met the average weekly use, weeks of use, and both average and total weeks of use recommendations in third grade among all the grades, while first grade students had the most students meet all three dosage recommendations.

Figure 2: Students who met the dosage recommendations by grade



N: 80,602: K (29335); 1st (33226); 2nd (9287); 3rd (8754).
Data source: pre-merged data in K-1 and data merged with Acadience Reading in Grades 2-3.

Program Impacts on Literacy Achievement

We studied the effectiveness of the program on literacy achievement by comparing groups of students who used the program to groups of students who did not. We present our findings in two sections: 1) Program-wide impacts, and 2) Individual vendor impacts. The first section includes findings on the impact of the EISP across all four software programs, providing a global view of how the program performed as it was used across the state, while in the second section, we explore the relative impacts of each program vendor. We have included a detailed methods section for technical reviewers in **Appendix A**. The program impact findings in this section are reported for general education students in K-1, and intervention students in Grades 2-3. Intervention students were defined as being below benchmark reading levels at the beginning of the school year.

Program-Wide Impacts

We began the program-wide analyses studying the program impacts for three samples representing different levels of program use (from lowest to highest use). This analysis helped illustrate the relationship between program effects and program use (or dosage) and depicted program effects for literacy composite scores for each grade. We then compared the program impacts for students who used the software across all levels of program use as well as for students who used the software for the minimum amount of time recommended by software vendors. We completed our analyses with an examination of program effects for specific groups of students. Program impact results reported in this section include students who used the program in K-1 and intervention only students in Grades 2-3.

How does program usage effect program impacts?

To determine how dosage affected outcomes, we split our treatment sample into three groups based on the range of participants' total minutes of use: Low: 6-1,086 minutes; Middle: 1,087 – 1,770 minutes; and High: 1,771 to the maximum number of minutes. As seen in **Table 5**, the statistically significant program-wide effects on Acadience Reading[4] end-of-year (EOY) composite scores increased with dosage, and the more a student used the program the better his/her EOY outcomes. For example, mean scores rose from ten to 17 points from the lowest to highest dosage categories across grades. The largest mean score difference occurred in kindergarten (17 points), followed by second grade (13 points), first grade (12 points), and third grade (10 points). Across all grade levels, the largest increase in mean scores occurred from the lowest to the middle dosage category versus the middle to highest dosage group.

Table 5. Predicted Means of Acadience Reading Composite Scores for Program-wide Treatment, Highest to Lowest Quantile Samples

	Kindergarten		1st Grade		2nd Grade Intervention		3rd Grade Intervention	
	N	Mean	N	Mean	N	Mean	N	Mean
High	5,299	159	11,944	195	2,921	157	2,163	253
Middle	7,832	152	9,574	193	2,462	150	2,407	252
Low	9,695	142	7,238	183	2,406	144	3,025	243

Data source: K-3 vendor data merged with Acadience Reading Outcome data and SIS data.
All data points displayed in the table were statistically significant at $p \leq .05$.

[4] Acadience Reading was formerly known as the Dynamic Indicators of Basic Early Literacy Skills® (DIBELS).

Did the program have an overall effect across all vendors?

We examined program impacts for students based on the following two analytic samples: 1) students who used the program, irrespective of their usage, which we identified as our Intent to Treat (ITT) sample; and, 2) students who met the program vendors dosage recommendations for average weekly minutes of use and total weeks of program use (MRD sample). The ITT analyses showed how the program affected all students throughout the state (in our sample), and the MRD analyses showed how the best usage was related to effects.

Table 6 presents the predicted mean scores and effect sizes of the matched treatment and control sample for the ITT group and the group of students who met vendors dosage recommendations (MDR). As shown below, there were statistically significant treatment effects in kindergarten (ITT and for the MDR groups) and in first grade (MDR Group). Effect sizes (ES) described the magnitude of the difference between two groups on an outcome and are often interpreted as meaningful if they reach a certain minimum threshold. For the purposes of this report, we defined this threshold as any effect size equal or greater to .26, which is the average effect size seen in similar intervention programs (Lipsey et. al, 2012). While the effect sizes increased from the ITT to the MDR group, they were all below the .26 threshold.

Table 6. Predicted Means of Acadience Composite Scores for Matched Treatment and Control, Program-wide, ITT and Met Recommendations Groups

	Kindergarten			1st Grade			2nd Grade Intervention			3rd Grade Intervention		
	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES
Met Dosage Recs	N=21,262			N=24,296			N=6,250			N=5,320		
	163	151	.12	199	194	.04	164	159	NS	262	260	NS
ITT	31,048			26,084			8,036			9,114		
	152	146	.05	191	190	NS	159	157	NS	255	258	NS

Note. NS (not significant) in a cell means the program did not have a statistically significant effect. ES: Effect Size (based on Cohens D). ES's greater than .26, the average for similar intervention programs, are highlighted in bold.
Data source: Matched K-3 ITT sample and matched MRD sample.
All data points displayed in the table were statistically significant at $p \leq .05$.

What impacts does EISP have on literacy skills as measured by the Acadience Reading?

We examined the program's benefits on specific literacy skill development (**Table 7)** by comparing Acadience Reading mean scores between treatment and control students. This analysis gave stakeholders a view into how the software changed students' test scores in specific skill areas. Program students had higher mean scores than their control group counterparts across all grade levels and literacy measures, although these differences were small (from 1 to 4 points;

scores presented were statistically significant unless otherwise specified). The largest difference in mean scores was observed for developing kindergarten and first grade students' alphabetic principles and basic phonics skills (NWF: CLS) as well as letter naming fluency (LNF) in kindergarten, with program students scoring 4 points higher, on average, than the control group. In the upper-early grades there were no significant findings for any literacy skill area. While there were slight differences between treatment and control group mean scores in K-1, none of these differences produced effect sizes greater than the .26 effect size benchmark.

Table 7. Predicted Means of EOY Acadience Reading Literacy Domains for Matched Treatment and Control, MRD Sample

Acadience Scale	Kindergarten N=20,888 21,262			1st Grade N=23,837 24,296			2nd Grade Intervention N=6,249 6,250			3rd Grade Intervention N=5,318 5,320		
	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.
First Sound Fluency (FSF)	41	39	2		N/A			N/A			N/A	
Letter Naming Fluency (LNF)	55	51	4		N/A			N/A			N/A	
Phoneme Segmentation Fluency (PSF)	56	54	2		N/A			N/A			N/A	
Nonsense Word Fluency-CLS	51	47	4	88	85	4		N/A			N/A	
Nonsense Word Fluency-WWR	9	9	1	27	26	1		N/A			N/A	
Oral Reading Fluency (ORF)		N/A		66	65	1		NS			NS	
DAZE		N/A			N/A			N/A		13	13	0

Note. NS (not significant) in a cell means the program did not have a statistically significant effect. N/A: measure not administered in grade.
Data source: Matched K-3 MRD sample.
All data points displayed in figure were statistically significant at $p \leq .05$.

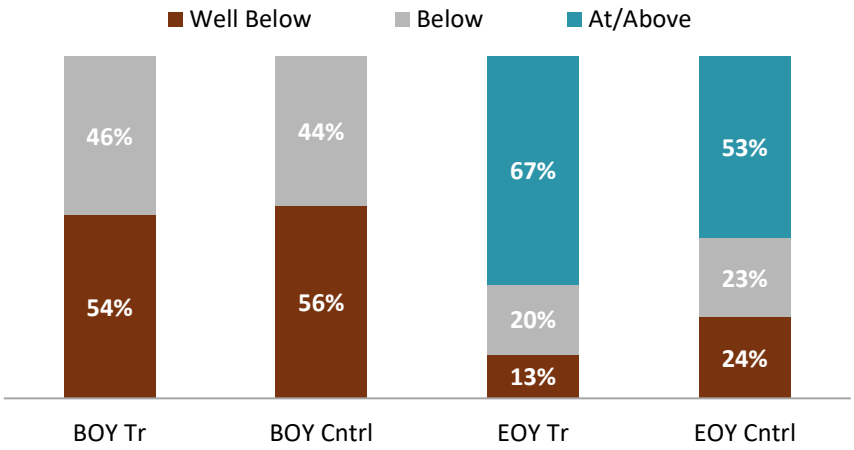
What were the differences in treatment and control group outcomes for at-risk students across all vendors?

We studied benchmark reading levels to determine if the program helped students improve their benchmark classification. Acadience Reading benchmark levels served as an indicator of students' reading level. Benchmark categories were designated as At or Above Benchmark, Below Benchmark, and Well Below Benchmark. Students with Acadience Reading composite scores below At or Above Benchmark for their grade level may be at-risk compared to their peers. To determine how the program affected the outcomes of at-risk students, we depicted the percent of students who started the year Well Below Benchmark or Below Benchmark for their grade, and followed their change in reading status in comparison to their non-program counterparts (**Figures 3-6)**. The two bars on the left of each figure portray the percentage of students who began the year Below or Well Below benchmark in the treatment and control group ("BOY Tr" vs. "BOY Cntrl"), and the two bars on the right portray the percentage of students who ended the year in each benchmark category ("EOY Tr" vs. "EOY Cntrl"). Similar to the trends found in the regression analyses, descriptive analyses showed that program students had the highest growth compared to their comparison group counterparts in kindergarten and first grade. We described the findings for each grade level in more detail in the following paragraphs.

Kindergarten: In kindergarten, 7,045 EISP and comparison students in the matched MRD sample began the school year below grade level based on their beginning of year reading Acadience Reading scores[5]. Of these, slightly over half began the year well below benchmark in both the treatment and control groups. Students with scores in the well below benchmark category had a 10-20% likelihood of achieving subsequent reading goals without intensive support outside of core curriculum (Dynamic Measurement Group, 2016). By the end of the school year, over half of the students who began the year reading below grade level had caught up. Moreover, 14 percent more treatment students ended the year reading at grade level compared to the control students (67% vs. 53%; Figure 3).

[5] The matched MRD dosage group was matched on beginning of year composite scores and therefore the percentage of students in each benchmark category is not equal.

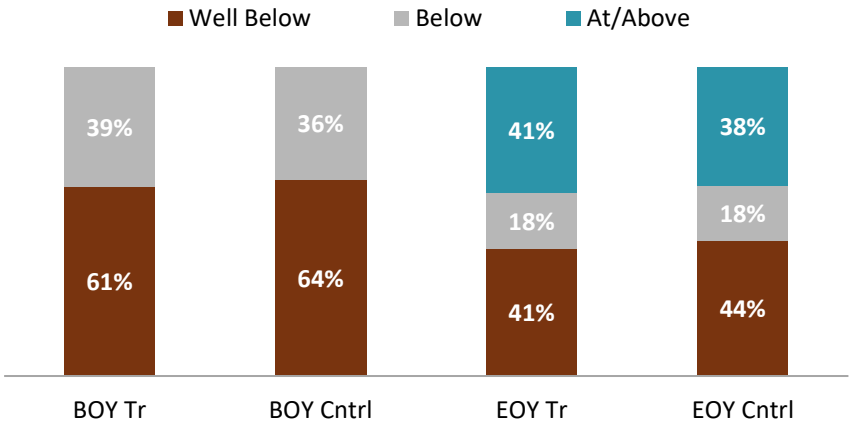
Figure 3. % Change in Benchmark Status from BOY to EOY, Kindergarten



Data source: Subset of students reading below benchmark at BOY from the matched kindergarten MRD sample. N: 7,045

First Grade: There was a 20 percent drop in students reading well below grade level from beginning to the end-of-year for both treatment and control groups in the first grade (**Figure 4**). In addition, there was a 21 percent drop for treatment students who were reading below benchmark versus an 18 percent drop for control students (a difference of 3 percent in favor of the treatment group).

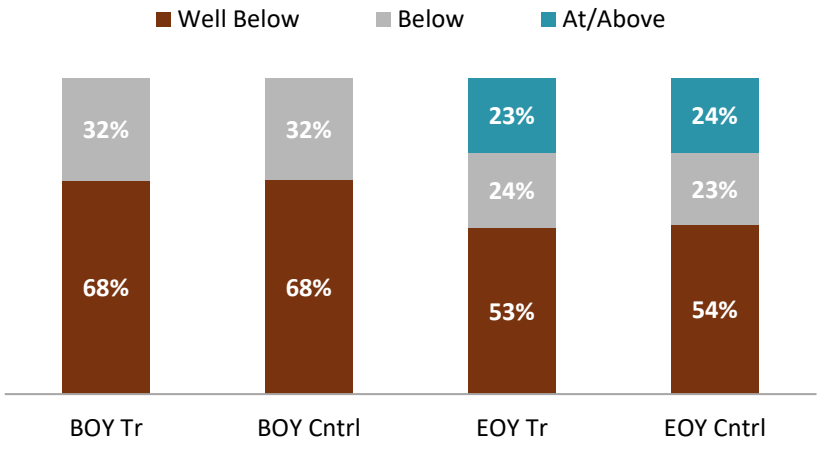
Figure 4. % Change in Benchmark Status from BOY to EOY, 1st Grade



Data source: Subset of students reading below benchmark at BOY from the matched kindergarten MRD sample. N: 9,360

Second Grade: Both treatment and control groups had high levels of Well Below Benchmark at BOY, an alarming trend in students entering into second grade. By EOY, both treatment and control students had a drop in the number of Well Below Benchmark students- a positive occurrence for both groups- and the treatment group had a minor advantage of one percent more students leaving Well Below Benchmark (**Figure 5**). Despite these positive findings, approximately half of the students in both groups still fell within the Well Below benchmark category at EOY, which underscores a need for additional supports for these at-risk students.

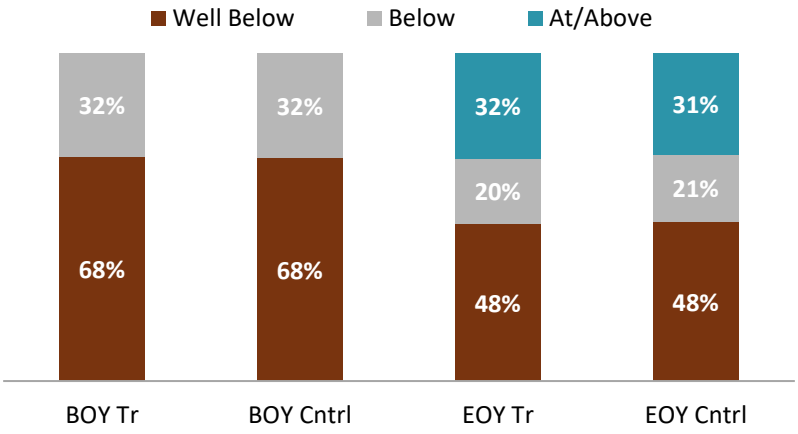
Figure 5. % Change in Benchmark Status from BOY to EOY, 2nd Grade



Data source: Students reading below benchmark at BOY, matched 2nd MRD sample. N: 6,250

Third Grade: Similar to the second grade findings, there was a 1% difference in growth among the percentage of students reading at grade level compared to non-program students (**Figure 6**). Thirty-two percent of program students and 31 percent of non-program students in the matched MRD sample identified as Below or Well Below Benchmark at the beginning of the school year reached At/Above benchmark status by year end.

Figure 6. % Change in Benchmark Status from BOY to EOY, 3rd Grade



Data source: Students reading below benchmark at BOY, matched 3rd Grade MRD sample. N: 5,320

What interactions between student characteristics and school type effect program impacts?

Table 8 presents the mean score differences in Acadience Reading composite scores at program exit for certain subgroups of program students. Program students who were identified as low-income, special education (SPED), English Language Learners (ELL), and those who attended a Title 1 school had higher predicted means scores than their low-income, special education (SPED), ELL, and Title 1 non-program counterparts in kindergarten. These differential treatment effects were the most pronounced for special education student: in kindergarten they scored 8.2 points higher than non-program special education students. There were no statistically significant differences in Grades 1-3.

Table 8. Mean Score Differences on EOY Acadience Reading Composite Scores by Grade and Subgroup, MRD Sample

	Kindergarten	1st Grade	2nd Grade Intervention	3rd Grade Intervention
Special Education (SPED)	8.2	NS	NS	NS
ELL	7	NS	NS	NS
Low-income	4.1	NS	NS	NS
Title I Schools	5.6	NS	NS	NS

Note. NS (not significant) in a cell means the program did not have a significant effect.
Kindergarten (N= 21,262); 1st Grade (N= 24,296); 2nd Grade (N= 6,250); 3rd Grade (N= 5,320)
Data source: Matched K-3 MRD sample.
All data points displayed in figure were statistically significant at p≤ .05.

Individual Vendor Impacts

The vendor-specific analyses were designed to help program stakeholders understand the effectiveness of the individual programs and make informed decisions. The vendor-specific findings in this section included a mean comparison between each program and a matched control group that showed program effects on overall literacy scores. We also presented a vendor-specific benchmark analyses, in which we demonstrated the change in benchmark status between treatment and control students from beginning to end-of-year. Finally, we concluded the section with an analysis of Acadience Reading Pathways to Progress measures, a tool used to examine growth rates between the treatment and control group. We also distinguish the grade levels included in vendor contracts from grade levels that were not awarded funding in the vendor results.

What were the differences in treatment and control group outcomes among vendors?

Table 9 presents the predicted means and mean score differences of program and non-program students in the matched sample of students who met at least 80 percent of vendors dosage recommendations. Findings that were not statistically significant were identified as "NS", not significant. All four vendors had a positive impact on students in kindergarten (Waterford, Imagine Learning, Core5, and i-Ready), followed by two vendors in first grade (Waterford; i-Ready), one vendor in second grade (Waterford), and two vendors in third grade (Core5, i-Ready). In kindergarten and first grade, the average predicted Acadience Reading composite means for both program and non-program students fell within or above the At Benchmark range for their grade (119-151 in kindergarten and 155-207 in first grade), which signifies a 70-85% likelihood of achieving subsequent reading outcomes (Dynamic Measurement Group, 2016). Second and third grade students who began the year reading below grade level and with whom received program benefits were still at risk based on their end-of-year reading level: predicted mean scores fell within the Well Below Benchmark range (0-179 in second grade and 0-279 in third grade) at end-of-year. Second and third grade findings suggest that, while certain vendors helped students improve compared to those who did not receive the program, "the probability of achieving later reading goals is low unless intensive support is provided" (Dynamic Measurement Group, 2016).

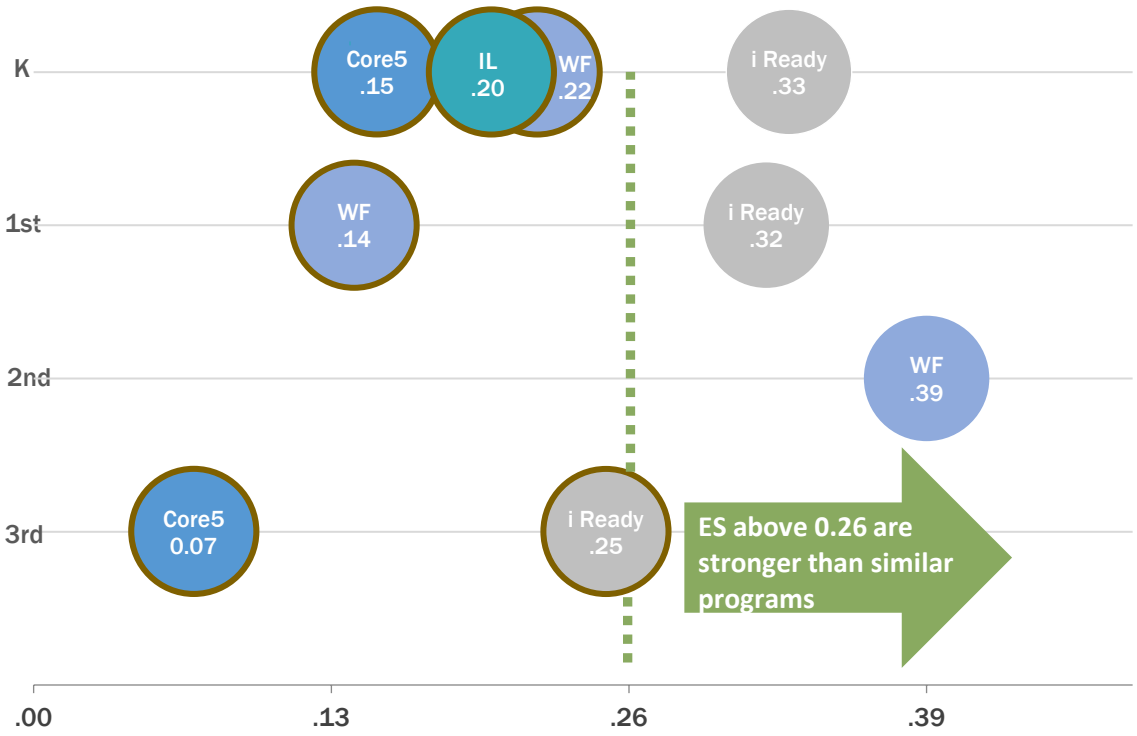
Table 9. Predicted Means of EOY Acadience Reading Composite for Matched Treatment and Control, by Vendor, OLS Regression Model

	K			1st			2nd			3rd		
	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.
WF		**N=3,470**			**2,202**			144			IS	
	149	**140**	**9**	**189**	**180**	**9**	151	128	23			
IL		**8,722**			14,302			3,750			2,852	
	155	**147**	**8**		NS			NS			NS	
Core5		**15,362**			16,960			**4,660**			**4,612**	
	160	**154**	**6**		NS		**157**	**162**	**-5**	**262**	**258**	**5**
i-Ready		1,376			1,924			670			**698**	
	160	147	13	207	188	19		NS		**272**	**256**	**16**

Notes. Data highlighted in bold identifies the grade levels included in vendors' contracts.
Model covariates were gender, Hispanic, special education, school Title I status, low-income, ELL and BOY Composite score. IS: Insufficient sample. NS (not significant) in a cell means the program did not have a significant effect.
Data source: Matched tr and cntl K-3 students who met at least 80% of vendors dosage recommendations.
All data points displayed in figure were statistically significant at $p \leq .05$.

Figure 7 below presents the effect sizes of the above analyses. As shown, i-Ready had a meaningful effect on student learning in kindergarten (ES: .33) and first grade (ES: .32). In second grade, Waterford had substantive effects on student learning (ES: .39). The remaining vendors had positive effects as well, but none were greater than the effect size benchmark of .26.

Figure 7. Impact of Individual Vendors on Acadience Reading Composite Scores, Effect Sizes by Grade



Note. Data points with borders identify grade levels included in vendors' contracts.
IS for WF 3rd grade (n=8).
All data points displayed in figure were statistically significant at p≤ .05.
Data source: Matched tr and cntl K-1 and 2-3 intervention students who met at least 80% of vendors dosage recommendations.

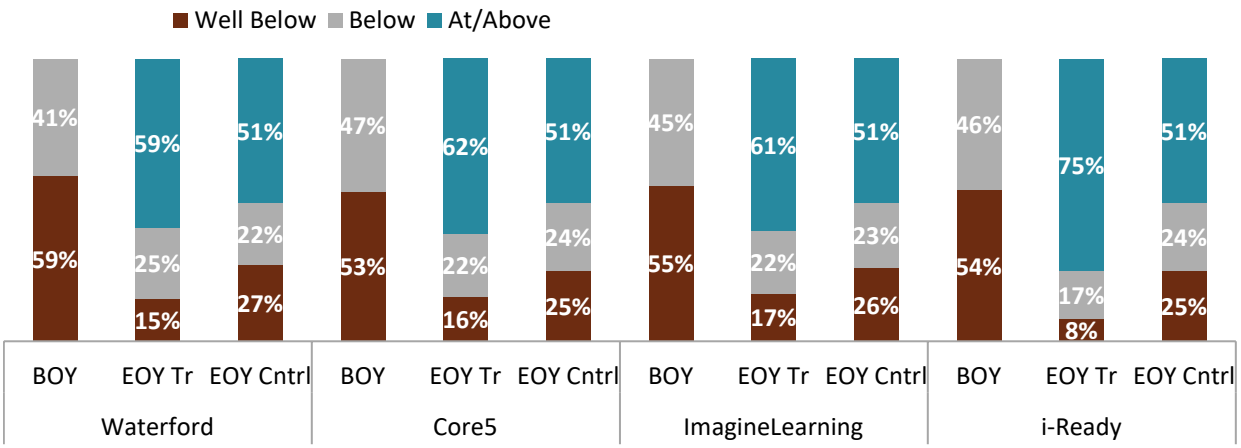
What were the differences in treatment and control group outcomes for at-risk students within each vendor?

Acadience Reading benchmark levels serve as an indicator of students' reading level. Benchmark categories are designated as At or Above Benchmark, Below Benchmark, and Well Below Benchmark. Students with Acadience Reading composite scores may be at-risk compared to their peers if their literacy composite scores were below At or Above Benchmark for their grade level. To determine how programs affected the outcomes of at-risk students, we compared the positive growth of program and non-program students who started the year below grade level based on their benchmark status.

Kindergarten

Over 50 percent of the students who began the year reading Below or Well Below Benchmark ended the year At or Above Benchmark in kindergarten. A higher percentage of treatment students ended the year At or Above Benchmark compared to control students for all four software vendors. The highest difference in growth from beginning to the end-of-year occurred for i-Ready program participants (24% T-C group difference), followed by Core5 (11%), ImagineLearning (10%), and Waterford (8%).

Figure 8. % Change in Benchmark Status from BOY to EOY, Kindergarten Students reading below benchmark at BOY

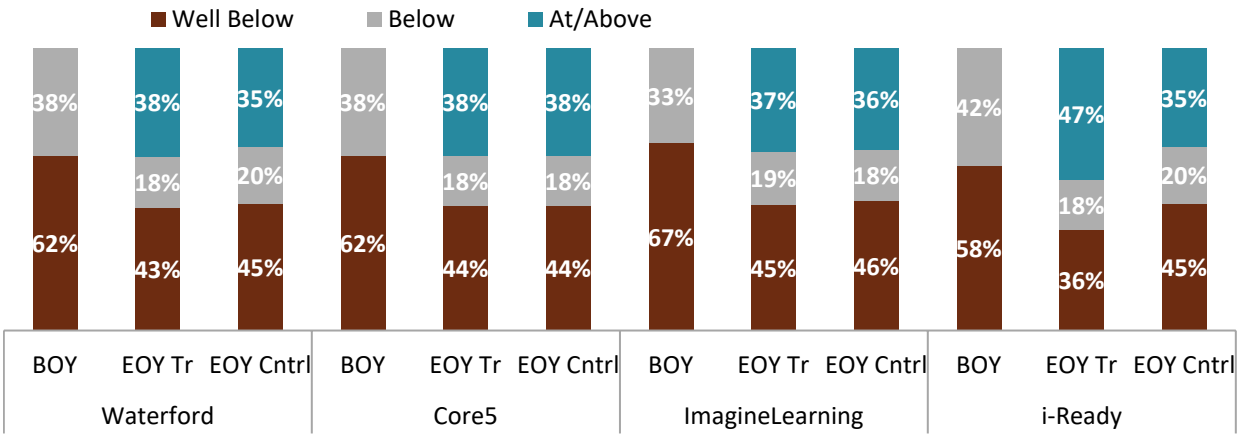


Note. Data source: Subset of students reading below benchmark at BOY from matched kindergarten sample. WF, Core5, and IL were contracted with the state to be used in kindergarten.

First Grade

As shown in **Figure 9**, students using three of the four vendors (i-Ready; Waterford; ImagineLearning) had higher growth compared to control students at the end-of-year, as measured by their percent change in benchmark status. i-Ready was the only vendor with a difference in growth higher than 5 percent, with a difference in growth of 12 percent.

Figure 9. % Change in Benchmark Status from BOY to EOY, First Grade Students reading below benchmark at BOY

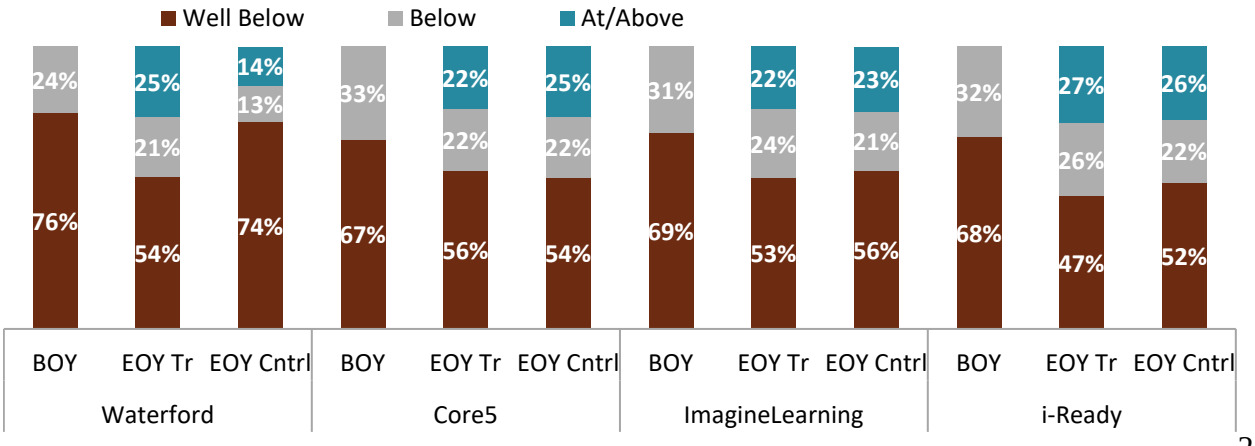


Note. Data source: Subset of students reading below benchmark at BOY from matched 1st grade sample. WF, Core5, and IL were contracted with the state to be used in 1st grade.

Second Grade

In second grade, two of the four vendors had a higher percentage of students scoring at or above benchmark compared to control students at the end of the school year: Waterford (11%) and i-Ready (1%).

Figure 10. % Change in Benchmark Status from BOY to EOY, Second Grade Students reading below benchmark at BOY

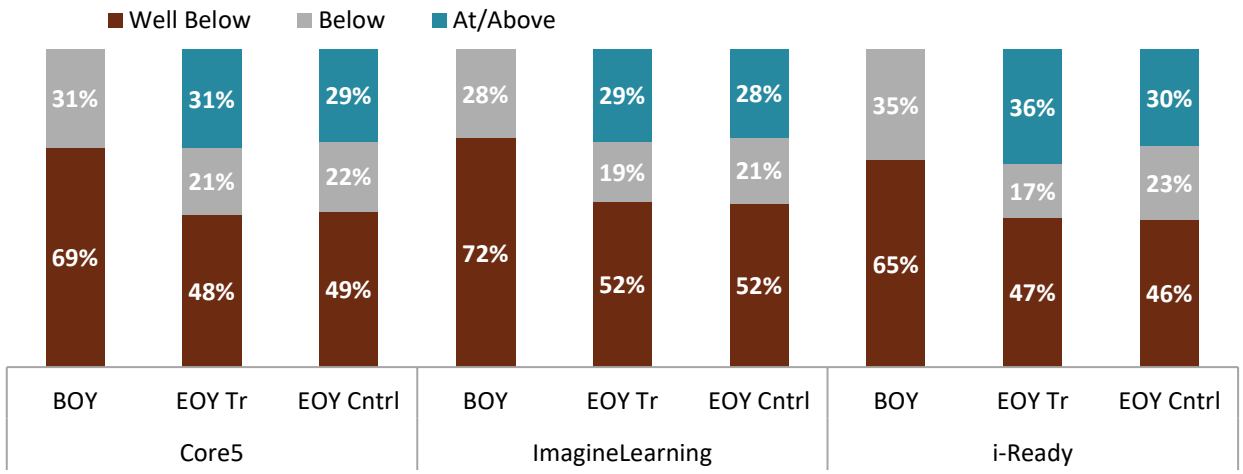


Note. Data source: Subset of students reading below benchmark at BOY from matched 2nd grade sample. Core5, IL and i-Ready were contracted with the state to be used in 2nd grade.

Third Grade

A higher percentage of third grade students moved from well below or below benchmark to at or above benchmark for all three vendors, compared to their control student counterparts. Thirty-six percent of students who used i-Ready were reading at or above benchmark compared to 30 of the control students, a difference of 6 percent, while the remaining two vendors had a difference in benchmark status between treatment and control groups of less than 5 percent.

Figure 11. % Change in Benchmark Status from BOY to EOY, Third Grade Students reading below benchmark at BOY



Note. Data source: Data source: Subset of students reading below benchmark at BOY from matched 3rd grade sample. Core5, IL, and i-Ready were contracted with the state to be used in 3rd grade.

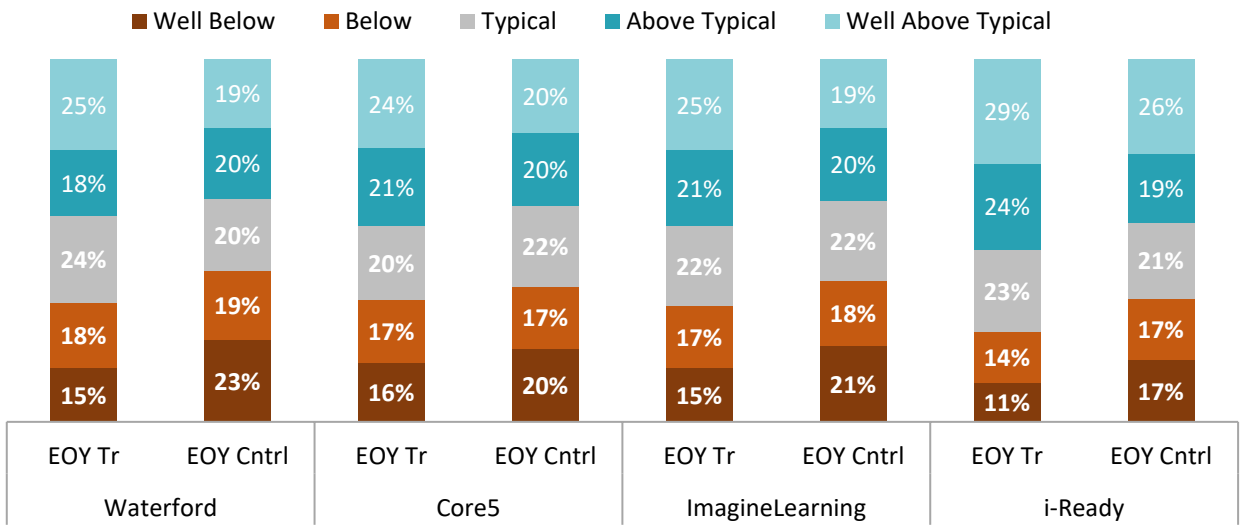
What were the differences in treatment and control group growth rates for students within each vendor?

Growth rates were important indicators of student learning trajectories, because they showed how quickly a student learns beyond just a single outcome score. Acadience Reading also offered a Pathways to Progress growth measurement tool, which are based on student growth percentiles. The pathways measure classified students' growth at the end-of-year, from Well Below Typical to Well Above Typical, compared to like peers from students beginning-of-year Acadience Reading data. Students' rates of growth were grouped into five categories: Well Below (below 20th percentile), Below (20th to 39th percentile), Typical (40th to 59th percentile), Above Typical (60th to 49th percentile) and Well Above Typical (80th percentile and above) growth (Dynamic Measurement Group, 2018). The figures below depicted end-of-year growth percentiles for the matched treatment and control group for each software vendor and grade.

Kindergarten

In kindergarten, students across all four software vendors experienced higher rates of growth compared to their non-program counterparts. I-Ready had the highest percentage of participants experiencing above typical or well above typical growth compared to the comparison group (8%), followed closely by Imagine Learning (7%), Core5 (5%), and Waterford (4%).

Figure 12. % of Students in each Pathways of Progress Category at end-of-year, kindergarten

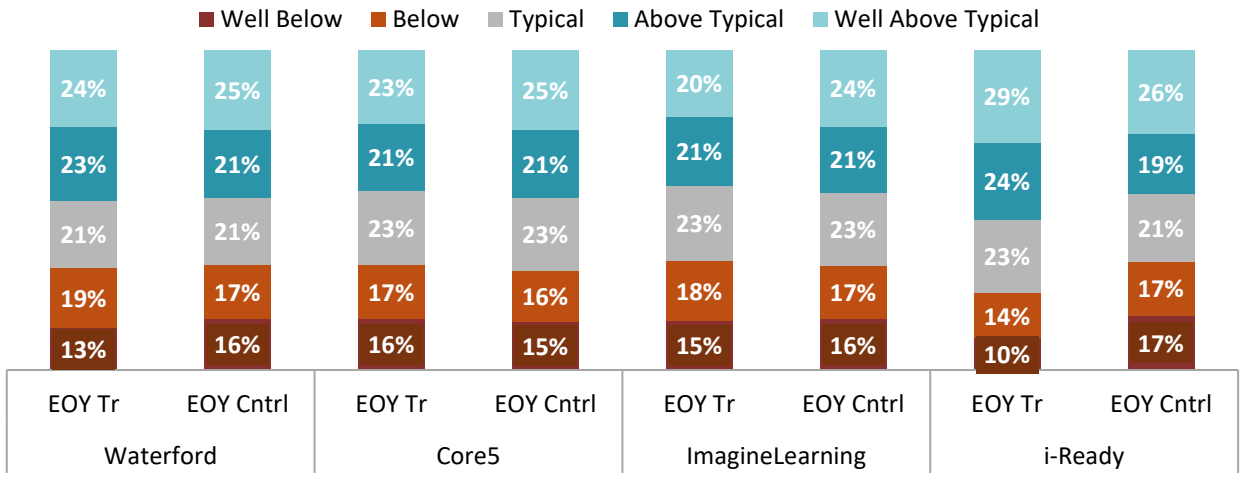


Data source: Matched tr and cntl K-3 students who met at least 80% of vendors dosage recommendations. WF, Core5, and IL were contracted with the state to be used in kindergarten.

First Grade

In first grade, fewer program students experienced higher growth rates compared to non-program students. As shown in **Figure 14**, only i-Ready and Waterford had a higher percentage of students with above to well above typical growth compared to non-program students, with 8% and 1% differences in growth, respectively.

Figure 13. % of Students in each Pathways of Progress Category at end-of-year, kindergarten

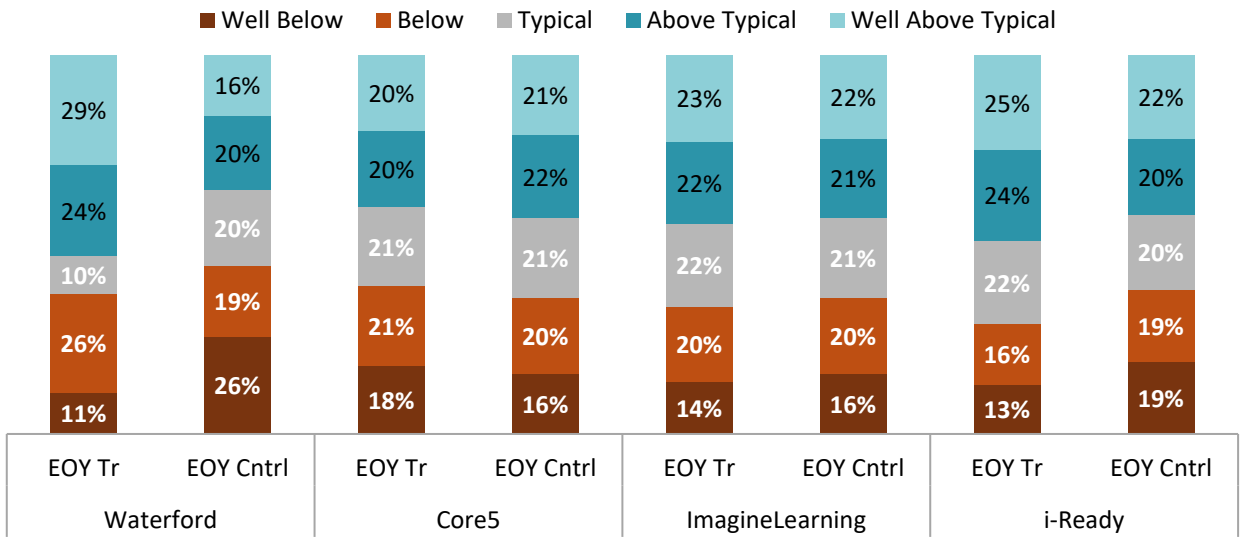


Data source: Matched tr and cntl K-3 students who met at least 80% of vendors dosage recommendations. WF, Core5, and IL were contracted with the state to be used in 1st grade.

Second Grade

For three out of four vendors, growth rates were higher for program participants than non-program participants in the second grade. Waterford had the highest difference in growth, with 17% more students who used the program experiencing above to well above typical growth rates, compared the comparison group of students. I-Ready (7%) and ImagineLearning (2%) also had higher growth rates compared to control group students.

Figure 14. % of Students in each Pathways of Progress Category at end-of-year, First Grade

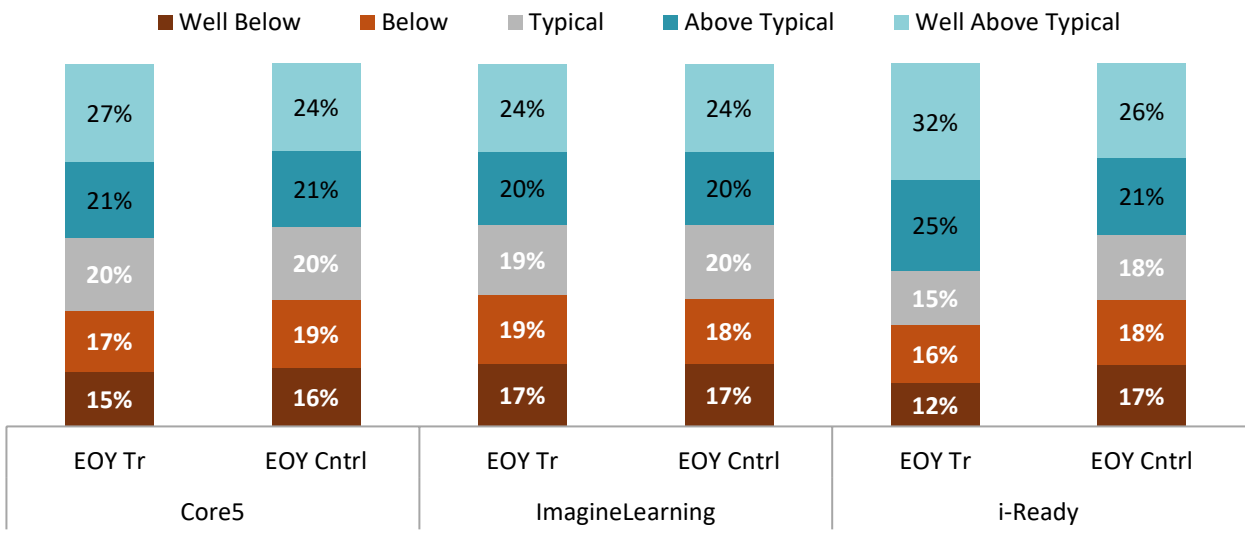


Data source: Matched tr and cntl K-3 students who met at least 80% of vendors dosage recommendations. Core5, IL and i-Ready were contracted with the state to be used in 2nd grade.

Third Grade

In third grade, a 10% difference in growth occurred for i-Ready students who fell within the above to well above growth rate categories, compared to non-program students. Core5 students also experience slightly higher growth rates in comparison to their non-program peers (a 3% difference in growth).

Figure 15. % of Students in each Pathways of Progress Category at end-of-year, Third Grade



Data source: Matched tr and cntl K-3 students who met at least 80% of vendors dosage recommendations. Core5, IL, and i-Ready were contracted with the state to be used in 3rd grade.

Discussion, Limitations and Recommendations

There were two primary evaluation goals: 1) to study program implementation in relation to vendors dosage recommendations; and, 2) to determine the impacts of the program on students' literacy achievement. We summarized the key findings for both goals in this section, presented a brief set of recommendations to help improve the program, and discussed the research limitations and their implications for interpreting the findings.

Program Implementation

Program implementation continues to improve year to year, but there is still much room for improvement. Successful program implementation hinged on using the software according to vendor recommendations. Program vendors set minimum program use recommendations for LEA's to receive maximum program benefits for their students. The recommendations included minimum average weekly login time and total weeks of program use. The specific dosage recommendations varied by vendor and grade level and ranged from 20 to 80 minutes per week and from 18 to 30 weeks of program use. In general, LEAs were able to meet either the average weekly use or minimum weeks of use recommendations, but not both. Less than half of the students who used three of the four vendors met both types of recommendations: ImagineLearning (44%), Waterford (23%), i-Ready (28%).

Program-wide Impacts on Literacy Achievement

Overall Effects. The program had strong effects in kindergarten which diminish through the more advanced grades. Program effects on students' literacy achievement were studied for Grades K-3. We compared literacy achievement outcomes between treatment and matched control students for two groups of treatment students: all students who used the software in our sample (known as ITT or intent to treat) and treatment students who used the program for at least the minimum average minutes and total weeks recommended by software vendors (known as the MDR or met dosage requirements). MDR students had stronger effects on literacy achievement than ITT in K-1, but there were no effects found in Grades 2-3.

Dosage Dependent Effects. Higher program use resulted in better literacy achievement outcomes, however, these effects did not scale linearly and there was a slight diminishing return to literacy growth at the higher levels of use. Program dosage effects on literacy outcomes were studied using a quantile analysis. We examined how treatment group mean scores changed with different usage levels (from low to high). We found an increase in students' mean achievement scores as total minutes of use increased within each dosage group. Interestingly, the biggest jump in mean scores occurred from the lowest (6-1,086 minutes) to middle dosage group (1,087 – 1,770 minutes), and not the middle to highest (1,771 to maximum minutes).

Effects on Specific Literacy Skills. We did not find program effects that benefitted individual Acadience Reading skill areas above others. When we studied the Acadience literacy scales, our findings for these analyses mirrored those in previous analyses in that significant results were reported in K-1, but no relevant findings were observed in Grades 2-3. Further, the findings

showed only minor mean score differences between treatment and control group mean scores across skill areas (less than 5 points).

Student Subgroup Effects. We found that the program was more effective for specific types of learners in kindergarten. Students who were identified as low-income, special education (SPED), English Language Learners (ELL), and those who attended a Title 1 school had higher predicted means scores than their non-program counterparts with the same characteristics. In addition, the descriptive benchmark analyses indicated the program was effective for students who read below grade level in kindergarten: nearly 14% more program students ended the year reading at grade level compared to non-program students. The differences in benchmark status between treatment and control group students in Grades 1-3 were less than 5 percent, indicating that the program had very little effect on this subgroup of students in these grades.

Vendor Impacts on Literacy Achievement

Effect Size Comparison. Vendor specific effects varied across vendor and grade. While each vendor had statistically significant effects, not all vendors had effects that exceeded our threshold for determining real world impacts. Only i-Ready (ES .33, kindergarten; ES .32 first grade) and Waterford (ES .39 second grade) had effect sizes larger than the .26 ES threshold.

Benchmark Growth Comparison. I-Ready was the vendor with the biggest change in benchmark status, compared to its control group and other vendors, in kindergarten, first and third grade, while Waterford participants had the highest percentage of change in benchmark status in second grade. Our final pathways of progress growth rate analyses further supported these findings in that the same programs with statistically significant impacts in the mean score analyses also had higher rates of growth in this descriptive analyses. This analyses added additional value in enabling us to contextualize findings in a easy to understand format.

Recommendations

The program has shown its effectiveness in kindergarten but had mixed results in first, second and third grade. Special education and English language learners were two groups of at-risk students that benefitted slightly more than their non-program peers. In addition, students in Title 1 schools had slightly better outcomes. These findings supported continued program use with these student groups.

Our results underscored how important it was for students to use the program according to vendors' minimum use recommendations. The below recommendations were intended to help the state improve program use across LEAs:

- To receive more program benefits, the state needs to hold LEAs accountable for using the program appropriately through annual fidelity of use reports. To do this we recommend that the USBE employ these strategies:
    - Highlight the importance of meeting both types of dosage recommendations across multiple modes of messaging to LEAs, including program enrollment communications, during vendor school trainings, etc.

- Encourage vendors to have consistent communication with key LEA staff about their progress towards meeting both types of dosage recommendations. For example, submitting usage reports to principals and teachers.
- Follow-up with LEAs with high levels of usage to discuss other possible strategies and best practices for encouraging appropriate levels of program use.

Evaluation Limitations

To understand the effect of the program on literacy achievement we compare program students to a group of similar non-program students. In recent years, we understand that LEAs have been increasing their use of digital technology intervention programs in the state, and it is possible that some of our control students used similar intervention programs, which may underestimate the strength of the program impacts. It is also possible that some LEAs used the same reading interventions with their students using a non-EISP funding source. For future evaluations, it would be useful for the USBE and vendors to track and share this information with evaluators.

At-risk control students may have had outside interventions that were unaccounted for by the evaluation, which removed any potential treatment effect because both groups received treatments. Program students in Grades 2-3 were classified as needing an "intervention" to improve their reading skills. The criteria for meeting this classification was reading at least one level (on the Acadience Reading Benchmark) below their peers. It stands to reason that when we matched a group of control students (who did not participate in the program) to these intervention program students, the controls were also likely to be identified by a teacher or school as needing help to bring their reading skills up to grade level benchmarks. It is possible that these at-risk control students received alternative reading interventions that were not able to be controlled for by our evaluation. If control students had alternative reading interventions, such as tutoring, after school programs, or other types of support, then we would not necessarily expect to see a treatment effect.

The results presented in this report were based on the outcomes of students who were included in our analytic sample and, while sampling treatment students from the population of program students is a standard practice, these results may differ from the population coefficients.

Waterford in 2nd grade had a small matched sample size of 144 students. Smaller sample sizes are sometimes associated with an increased margin of error, and as a result, we recommend this finding be interpreted with caution.

We have learned from previous evaluations that teachers were more or less active in supporting students use of the software in the classrooms, but we did not know to what extent teachers and schools were involved with program implementation among our sample of schools. Having this information could be helpful in the future to help us understand the link between different levels of program implementation beyond program dosage.

References

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Dynamic Measurement Group, Inc. (2016, September). *Acadience Reading Benchmark Goals and Composite Score.* https://Acadience.org/papers/AcadienceNextBenchmarkGoals.pdf.

Evaluation and Training Institute. (2017, October). *Best Practices for Improving Early Intervention Software Programs in Utah Schools.* Culver City, CA: Author

Evaluation and Training Institute. (2017, September). *Early Intervention Software Program Evaluation: 2016-2017 Results.* Culver City, CA: Author

Evaluation and Training Institute. (2016, September). *Early Intervention Software Program Evaluation: 2015-2016 Results*. Culver City, CA: Author

Evaluation and Training Institute. (2015, September). *Early Intervention Software Program Evaluation: 2014-2015 Results.* Culver City, CA: Author

Evaluation and Training Institute. (2014, October). *Early Intervention Software Program Evaluation: 2013-2014 Results.* Culver City, CA: Author

Hill, C. J., Bloom, H. S., Black, A. R. and Lipsey, M. W. (2008), *Empirical Benchmarks for Interpreting Effect Sizes in Research.* Child Development Perspectives, 2: 172–177. doi: 10.1111/j.1750-8606.2008.00061

Iacus, Stefano M., Gary King and Giuseppe Porro. 2008. *Matching for Causal Inference without Balance Checking*. http://gking.harvard.edu/files/abs/cem-abs.shtml.

IBM Corp. Released 2013. IBM SPSS Statistics for Mac, Version 22.0. Armonk, NY: IBM Corp

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington DC: Institute of Education Sciences.

Powell-Smith, K., Good, R.H., III, & Dewey, E.N., & Latimer, R.J. (2014). *Assessing the Readability of Acadience AD Oral Reading Fluency and Daze.* (Technical Report No.16). Eugene, OR: Dynamic Measurement Group.

Good, R.H., III, Powell-Smith, K., Kaminski, R.A., Stollar S., & Wallin J. (2011).

Acadience Reading Assessment Manual. Dynamic Measurement Group Inc. http://wenatchee.innersync.com/assessment/documents/Acadiencenext_assessmentmanual.pdf

Appendix A. Evaluation Methods

We provide an overview of our research methods, samples and data sources that were used to answer each research question. The methods are described for the two studies, the impact study of students' achievement outcomes and the implementation study of students' program use, that were used to inform the program evaluation. **Appendices A-C** provide additional details on our methods, data processing procedures and samples.

Which program participants were included in our study?

Implementation Study Evaluation Participant Samples

The goal of the implementation study was to examine the extent to which students used the software as intended by each program vendor. We included as many students who used the programs as possible to provide the most accurate depiction of students' program use, and the samples used for the implementation analyses were the most inclusive of all the samples. For K-1 students, we used the vendor data, and did not remove students with inaccurate SSIDs, students who used multiple software providers, or students with incomplete Acadience data. In Grades 2-3, our focus was on struggling readers, and we needed valid SSIDs in the vendor and Acadience data as well as beginning-of-year Acadience scores to identify the students reading below grade level.

Impact Study Evaluation Participant Samples

For the impact analyses, we selected a group of student participants (students who used the software) within the larger pool of program students to create an "analytic sample," which is the group of students with whom we ran our statistical analyses *(see **Appendix B** for descriptive statistics of the students included in our samples).* Our analytic samples changed based on the specific combination of vendor and grade of students being analyzed. In second and third grade, the program was designed to target intervention students only (students performing below grade benchmark literacy levels), and we constrained our samples to include participants who were below grade level literacy benchmarks at the beginning of the year across all analyses. Students needed to have accurate state student Ids (SSIDs) and complete Acadience data (outcome data) to be a viable case for our sample. We excluded students who may have used multiple software programs in order to study the individual impacts of each software vendor.

Control Student Matching Process. Our impact study analysis compared program students' literacy achievement outcomes, measured using literacy test scores, to non-program students' outcomes. This is known as a treatment-control comparison, and the comparison students allowed us to create a control condition and determine what impact the program had on learning achievement. Program students were matched to control students using Coarsened Exact Matching (CEM, Lacus et al., 2008). The students were matched on data from the beginning of the school year, and across several important characteristics (covariates used included: grade, beginning-of-year achievement scores, gender, race, English Language Learner status, and

poverty status). If no matches could be made, children were removed from the sample. CEM minimized differences between the two groups prior to enrollment in the program, creating groups of treatment and control student groups that were balanced across covariates.

Program-Wide Samples. For the program-wide analyses, we explored how program dosage impacted students' literacy skill development in two ways. First, we created three analytic samples of students with three levels of program dosage (Low, Medium, High) to study the effects of increased program use on students' test scores across vendors. To create these three groups, we separated the treatment sample into three equal groups, or quantiles, based on the range of participating students' total minutes of software use.

We also created two matched treatment and control samples based on two dosage thresholds. The first program-wide matched sample was comprised of students who used the software based on the vendors' recommended dosage[6]. The vendors' recommended dosage was based on vendors recommendations for how much time students should use the program before benefits are observed, and we wanted to determine how literacy outcomes were affected for students who met these recommendations. The second matched sample, the Intent to Treat (ITT) sample, included all students who used the program for any amount of time and showed how effective the program was for students, irrespective of use.

Individual Vendor Samples. For the individual vendor analyses, our goal was to create a sample of students who used the software long enough for improvements in literacy skill development to occur. If we created our sample from students who met the program vendors exact dosage recommendations for average minutes of use and minimum weeks of use, we would not have enough students to study each software program. Instead, we studied a subset of students who met a relaxed version of vendors' recommendations (students who used the software greater than or equal to 80% of vendors recommended use). We created one matched sample for each program vendor, which allowed us to have tightly matched control groups for each program vendor.

What sources of data were used in our analyses?

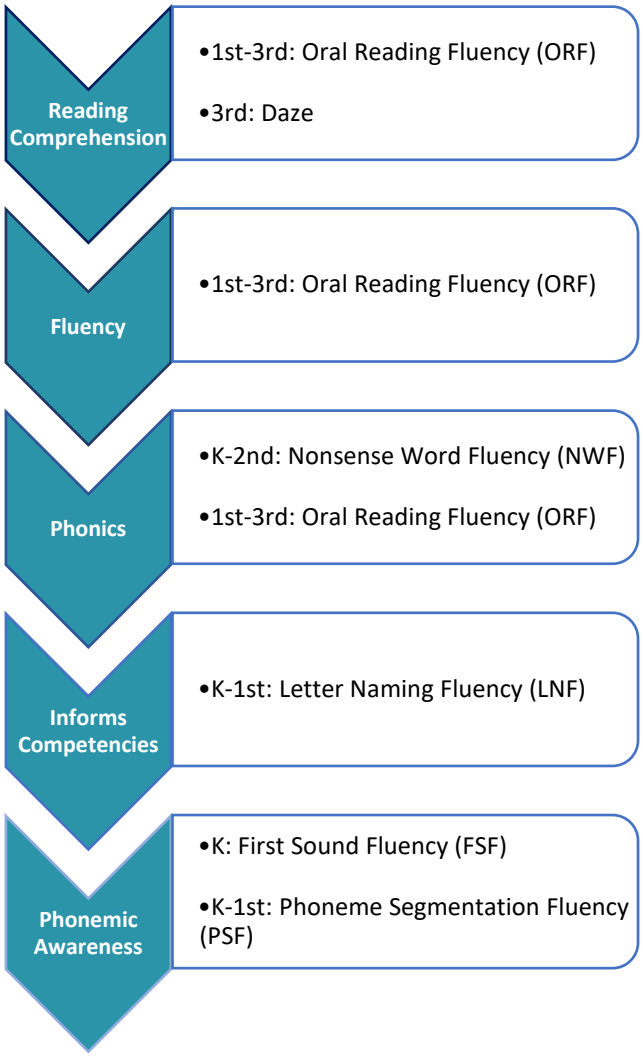
We collected data from nine different sources to create our master dataset for the EISP analyses. The data sources included: four program vendors, who provided us with usage information for each student who used their programs; state Acadience Learning (Acadience Reading) testing data; and student information system (SIS) demographic data provided by the Utah State Board of Education (USBE). See **Appendix C** for details on how we created our master dataset.

[6] Students had to meet both vendors' recommended average minutes of use per week and minimum weeks of total use to be included in this sample.

Which instruments did we use to measure literacy achievement?

We measured literacy achievement using the Acadience Reading assessment, which was formerly known as the Dynamic Indicators of Basic Early Literacy Skills® (DIBELS)[7]. Acadience Reading was administered in schools throughout the state in Grades K-3. The Acadience Reading measures were used throughout Utah and are strong predictors of future reading achievement. Acadience Reading is comprised of six measures that function as indicators of critical skills students must master to become proficient readers, including: First Sound Fluency (FSF), Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), Oral Reading Fluency (ORF), and reading comprehension (DAZE). In addition to scores for the six subscale measures described above, we used reading composite scores and benchmark levels, or criterion-reference target scores that represent adequate reading progress. See **Appendix D** for additional detail on the Acadience Reading measures.

Figure 16. Acadience Indicator & Literacy Skill Measures



How did we study program implementation?

Our program implementation findings focused on program usage in relationship to its intended use, as described through vendors' dosage recommendations. Program usage data included the following: total minutes of software use, from log-in to logoff for each week the program was used during the school year; total weeks, and average weekly use. Program vendors supplied the usage data.

How did we study the program-wide impacts across all vendors?

Our study relied on three types of statistical analyses to measure program-wide impacts, which included hierarchical linear modeling, ordinary least squares (OLS) regression modeling, and a descriptive analyses of students benchmark scores.

[7] There were no changes made to the assessment when the DIBELS Next changed to Acadience Reading.

Hierarchical linear regression model. We studied the program-wide impacts by comparing a sample of treatment group students drawn from all vendors to a matched sample of control students. We determined that using a two-level regression model (also known as a "hierarchical linear regression model", or HLM) allowed us to study the differences in treatment and control group student outcomes, while controlling for other student-level predictors, and also allowed us to control for Title 1 status school effects. A two-level random intercept statistical model with school as the level-2 predictor was used to regress student outcomes on our predictor variables. Our independent variable was treatment group status (1/0), and we included other predictor variables to control for their effects in our models, including: beginning-of-year (BOY) test scores, gender, special education status, economic disadvantaged status, and ethnicity in the model to adjust for their influence on end-of-year reading scores. By accounting for these additional predictor variables, we increased our ability to show a causal link between program use and outcomes, while holding other factors unrelated to the program constant.

In addition, we studied how program participation impacted students with specific characteristics, such as English Language Learners, special education students, economic disadvantaged students, and students from Title 1 schools. We included students who met the vendors' recommended dosage in this analysis sample.

We also used OLS regression analyses to study the mean scores of program students with low, medium and high program dosage. These dosage groups were created by separating students who used the software into three groups, or quantiles. The models take sex, low-income status, special education status, English learner designation, Title I status, and beginning-of-year test scores into consideration. Means should be interpreted as predicted marginal mean end-of-year composite scores for each level of usage for the average member of the population of children (the independent variables are held constant at their means). This analyses shows how the end-of-year mean scores change as dosage increases.

Benchmark Outcome Visual Analyses. To present our findings in an intuitive and applicable context, we measured the change in treatment and control students reading proficiency at the beginning and end of the school year for students who began the year reading below grade level. Changes in students' reading proficiency benchmark levels were reported for a subset of students reading below grade level in the Met Recommended Dosage (MRD) matched sample. Although we used a sample in which students were similar on average, descriptive statistics did not allow us to control for pre-existing differences between groups and need to be interpreted with caution.

How did we study individual vendor impacts?

Ordinary Least Squares regression model. We used an Ordinary Least Squares (OLS) regression model to predict the differences in mean scores between treatment and control students while controlling for demographic characteristics and baseline scores. We controlled for students' beginning-of-year (BOY) reading scores, gender, special education status, economic disadvantaged status, ethnicity, English Language Learner status, and Title 1 school status in the models. Some covariates were dropped in certain models due to collinearity.

Benchmark Outcome Visual Analyses. Similar to the program-wide findings, we conducted a descriptive analyses of students' change in benchmark categories from beginning-to-end of the school year for each vendor and grade. Changes in students' reading proficiency benchmark levels were reported for a matched sample of students who met at least 80 percent of vendors dosage recommendations and who scored below grade level at the beginning of the school year.

Pathways of Progress Visual Analyses. Acadience Reading also provides Pathways of Progress measures, which use normed data to depict students rates of reading progress at the end-of-year based on students with similar scores at the beginning of the year[8]. Unlike the benchmark analyses, we did not constrain this sample to students who were reading below grade level at the beginning of the school year. The sample included matched treatment and control students who met at least 80 percent of vendors dosage recommendations for each software vendor. This analysis allowed us to compare the rates of growth for treatment and control students based on a normed sample of students.

[8] Pathways of Progress are calculated based on 2014-2015 Acadience Reading data for 2,395,969 students across grades K-6 (Dynamic Measurement Group, 2018).

What statistics do we provide in our results?

Where appropriate, we provided predicted mean scores and mean score differences for our treatment and control groups, which are meaningful when comparing treatment and control groups from the same sample. Statistical significance testing allowed us to determine the likelihood that a finding was a result of chance, or due to the treatment effect. We also provided treatment effect sizes (ES; based on Cohen's Delta[9], or "d") to help readers understand the magnitude of treatment effects. Presenting effect sizes enabled us to provide a standardized scale to compare results based on different samples and measure the relative strengths of program impacts. Descriptive statistics, such as percentages, were presented to describe students' program use and change in reading proficiency benchmark status.

When interpreting our findings, it is important to note that effect sizes can be used to measure the strength of program impacts in multiple ways. A commonly used method is Cohen's (1988) characterization of effect sizes as small (.2), medium (.5) and large (.8). However, recent studies have suggested using a more targeted approach for determining the magnitude of the program impacts. For example, Lipsey et. al (2012) suggested effect size comparisons should be based on "*comparable outcome measures from comparable interventions targeted on comparable samples*", and notes that effect sizes in educational program research are rarely above .3, and that an effect size of .25 may be considered large (pg. 4). In other words, the strength of an intervention should be measured based on whether its effect size is at, above or below those of similar programs. The challenge with using this method is that there are several different ways we could create a benchmark from averaging the effect sizes of similar programs, including creating a benchmark by outcome measure (Avg. ES: .25), intervention type (Avg. ES: .13), intervention target (Avg. ES: .40), or averaging all three methods (ES: .26) (Lipsey et. al, 2012).

For the purposes of this study, we have chosen to contextualize our findings using the average of all three methods as our benchmark. The mean effect size for similar instructional programs is **.26**, and we consider this the standard by which to compare our results. Effect sizes larger than this are stronger than average, which we note in our results.[10] More information on how we selected our ES benchmark is provided in **Appendix E**.

[9] Effect sizes are calculated by taking the difference in the two groups means divided by the average of their pooled standard deviations.

[10] This interpretation is based on a review of 829 effect sizes from 124 education research studies conducted by researchers at the Institute of Education Sciences (IES) (Lipsey et. al, 2012).

Appendix B: Analyses Samples

Tables B1 – B4 present the characteristics of the treatment group for each matched dosage sample used in our analyses. As a result of our CEM procedure, our matched controls had the same characteristics as the treatment group. For example, if the treatment group was comprised of 49% female, 9% SPED, 10% ELL, and had an average composite score of 35, then the control group was also comprised of 49% female, 9% SPED, 10% ELL, with an average composite score of 35.

Program-wide Analyses Samples

Table B1. Program-Wide Sample by Grade, Lowest Dosage

	N	Female	Caucasian	Hispanic	Other	SPED	Low-income	ELL	Ave Minutes	Ave Wks.	BOY Comp
K	9,695	49%	76%	16%	8%	9%	30%	10%	31	19	35
1st	7,238	50%	76%	16%	8%	12%	33%	10%	34	21	120
2nd	2,406	50%	66%	25%	9%	30%	50%	17%	33	20	63
3rd	3,025	48%	65%	26%	8%	32%	47%	19%	34	18	120

Table B2. Program-Wide Sample by Grade, Middle Dosage sample

	N	Female	Caucasian	Hispanic	Other	SPED	Low-income	ELL	Ave Minutes	Ave Wks.	BOY Comp
K	7,862	48%	79%	14%	8%	8%	26%	6%	49	29	38
1st	9,574	49%	80%	13%	8%	10%	28%	7%	49	30	128
2nd	2,462	50%	68%	23%	9%	23%	44%	15%	49	29	72
3rd	2,407	47%	66%	26%	9%	27%	47%	18%	49	29	125

Table B3. Program-Wide Sample by Grade, Highest Dosage sample

	N	Female	Caucasian	Hispanic	Other	SPED	Low-income	ELL	Ave Minutes	Ave Wks.	BOY Comp
K	5,299	48%	75%	16%	9%	8%	28%	9%	77	31	37
1st	11,944	48%	76%	15%	9%	11%	31%	8%	74	33	126
2nd	2,921	49%	66%	24%	10%	24%	46%	17%	75	33	72
3rd	2,163	46%	61%	27%	12%	28%	47%	21%	73	33	121

Table B4. Program-Wide Sample by Grade, ITT Sample

	N	Female	Caucasian	Hispanic	Other	SPED	Low-income	ELL	Ave Minutes	Ave Wks.	BOY Comp
K	15,524	49%	78%	15%	7%	9%	30%	6%	48	25	37
1st	13,043	49%	79%	14%	7%	11%	33%	7%	55	29	125
2nd	4,018	48%	76%	19%	5%	23%	45%	12%	54	28	72
3rd	4,557	48%	76%	20%	4%	27%	46%	12%	49	26	126

Table B5. Program-Wide Sample by Grade, Met Vendor Recommendations Sample

	N	Female	Caucasian	Hispanic	Other	SPED	Low-income	ELL	Ave Minutes	Ave Wks.	BOY Comp
K	10,631	47%	81%	13%	6%	7%	25%	5%	62	29	41
1st	12,148	49%	81%	13%	6%	10%	31%	6%	66	32	128
2nd	3,125	49%	77%	19%	4%	22%	43%	12%	69	31	74

3rd	2,660	46%	73%	23%	4%	26%	45%	15%	65	31	128

Individual Vendor Impacts Analyses Samples

Table B6. Vendor-specific Matched Sample by Grade

	Grade	N	Female	Caucasian	Hispanic	Other Ethnicity	SPED	Low-income	ELL	Ave Minutes	Ave Wks.	BOY Comp
Waterford	K	1,735	48%	77%	16%	6%	9%	41%	6%	68	31	31
	1	1,101	49%	87%	10%	4%	11%	39%	3%	78	32	118
	2	72	47%	82%	15%	3%	22%	42%	11%	79	32	57
	3	IS										
Imagine Learning	K	4,361	49%	81%	14%	5%	7%	30%	7%	50	28	37
	1	7,151	49%	82%	14%	4%	10%	33%	7%	55	30	122
	2	1,875	50%	77%	20%	3%	24%	47%	12%	54	29	70
	3	1,426	47%	73%	24%	3%	27%	49%	16%	52	28	124
Core5	K	7,681	47%	80%	13%	7%	7%	19%	5%	61	28	42
	1	8,480	49%	82%	12%	6%	9%	27%	5%	67	32	130

	Grade	N	Female	Caucasian	Hispanic	Other Ethnicity	SPED	Low-income	ELL	Ave Minutes	Ave Wks.	BOY Comp
	2	2,330	49%	74%	21%	5%	21%	42%	13%	70	32	74
	3	2,306	47%	72%	23%	5%	26%	43%	15%	64	31	128
i-Ready	K	688	47%	84%	10%	5%	7%	36%	4%	43	29	36
	1	962	49%	86%	10%	4%	8%	32%	3%	54	31	125
	2	335	47%	81%	16%	3%	20%	42%	9%	55	31	80
	3	349	45%	81%	15%	5%	26%	42%	11%	50	30	129

Note. IS: insufficient sample in cell indicates vendor had insufficient sample size to report findings.

Appendix C. Data Processing & Merge Summary

We reviewed and cleaned data from six different sources in preparation of completing our analyses, including program usage data from four software program providers, student literacy achievement data, and demographic data (student information system, "SIS") data from the USBE. Throughout the different stages of data processing, a percentage of cases were dropped from each program vendor. In this Appendix, we show how our pool of treatment students shrank at each stage of the cleaning process, and describe how we cleaned the different types of data in the creation of the final datasets used our analyses.

Software Program Data

Each software program provider provided student level data with the time students spent in the software for each week of school. To help vendors provide quality data and ensure consistency across software program providers, vendors received an example data file, a description of the correct format for each variable, and a checklist to conduct a final review of their data. Our cleaning process for the program vendor data files included making sure all program schools that received licenses were included in the data, identifying and processing duplicate IDs within vendors' data, and formatting variables as needed, among other steps. We reviewed existing variables and created additional variables to use in our analyses, such as total weeks of use, average minutes of use, and other program fidelity measures.

When cleaning duplicate IDs within each vendors' data, we deleted cases that were the same student with different usage reported, and kept any unique cases after removing exact replicas. We did not count weeks, or include minutes, when there were fewer than five minutes recorded in a given week. After removing these instances, we updated the usage variables, such as total minutes, to reflect the change in use, and then removed students who had fewer than five minutes of total use from the data. After we cleaned and processed the vendors data, the total count of students went from 135,864 to 124,378 students. We used this data to study program implementation after identifying and removing students in Grades 2-3 who were reading on grade level at the beginning of year (n=89,140).

To create the vendor data used in our outcome analyses, we identified and removed duplicate IDs across vendors[11] (approximately 5,366 cases) and any IDs that did not comply with the state student ID (SSID) format (2,749 cases). The duplicate IDs across vendors indicated students used more than one software program, either because they moved to a different district, or because the LEA administered multiple programs to the same students. In either case, we did not include these students in order to report the individual impacts for each software provider.

SIS Data

We were provided SIS data for all students in Grades K-3. We reviewed the SIS data provided by the USBE to ensure that all LEAs who were listed as 2018-2019 participants were included in

[11] These IDs were also deleted from our pool of potential control students.

the data. The SIS data file consisted of 206,245 cases, of which approximately 3 percent were duplicate records. After cleaning the data of duplicates, our SIS data consisted of 200,864 records.

Acadience Reading Data

In 2018-2019, the USBE prepared and transferred an Acadience Reading data file (n=202,820). After cleaning the IDs (e.g. deleting missing IDs and IDs that were not in a valid format) and removing duplicates, we were left with a master Acadience file containing 188,276 cases. This master file contained outcome data for our pool of treatment and control cases.

Master Merged Data File

We merged the SIS data from the USBE into our master Acadience Reading file and were left with 188,275 cases. Next, we merged our master vendor data into the Acadience and SIS data, removed duplicate cases between vendors, missing data (e.g. beginning and end-of-year composite scores), and non-intervention students in Grades 2-3. Lastly, we identified (where possible) schools or students using one of the four program vendors through non-EISP funding and removed these cases from our pool of potential controls[12]. This included excluding students who used Imagine Learning through a separate state-wide grant[13] prior to reporting the program impacts for similar reasons (10,567 cases removed). After processing the data, our final, pre-matched dataset consisted of 106,226 cases, of which, 66,996 were treatment and 39,230 were potential controls.

Matched Data Files

Before we could run our analyses, the final step was to create our matched control groups. We needed to create a comparison group that matched the students in our program-wide sample, as well as for each individual vendor. We drew controls from a pool of non-program participants in the state of Utah, and in general, lost very few cases when creating our matched samples for individual vendors and the program-wide analyses which consisted of fewer students (e.g. the Met Recommended Dosage samples). However, for our largest sample of program students, the Intent to Treat (ITT) program-wide sample, there were more program students than control students. This automatically reduced the size of this particular sample.

12 We removed students from non-EISP funded schools who were using an EISP program based on information provided by vendors.

13 We excluded these students from our analyses using the SSIDs provided by Imagine Learning to identify students who used their reading software through this separate state-wide initiative.

Appendix D: Acadience Reading Measures

Acadience Reading is a statewide assessment used to measure students acquisition of early literacy skills at the beginning, middle, and end of the academic year. According to a technical report produced by the Dynamic Measurement Group (Powell-Smith, et al., 2014), *"The Acadience measures map on to the critical early reading skills identified by the National Reading Panel (2002) and include indicators of phonemic awareness, Alphabetic principle, vocabulary and oral language development, accuracy and fluency with connected text, and comprehension"*. **Table D1** provides a summary of the Acadience subscales used in our analyses.

Table D1. Acadience Reading Scales

Acadience Reading Scale	Description	Early Literacy Construct	Grade
Composite Score	Acadience Composite Score is a combination of multiple Acadience scores	Overall estimate of reading proficiency	K-6
First Sound Fluency (FSF)	A brief direct measure of a student's fluency in identifying initial sounds in words.	Phonemic Awareness	K
Letter Naming Fluency (LNF)	Assesses a student's ability to recognize individual letters and say their letter names.	Measure is an indicator of risk	K-1
Phoneme Segmentation Fluency (PSF)	Assesses the student's fluency in segmenting a spoken word into its component parts of sound segments.	Phonemic Awareness	K-1
Nonsense Word Fluency (NWF)	Assesses knowledge of basic letter sound correspondences and the ability to blend letter sounds into consonant-vowel-consonant and vowel-consonant words. Designed to measure alphabetic principle and basic phonics.	Alphabetic Principle and Basic Phonics	K-2
Oral Reading Fluency (ORF)	Students are presented with grade-level passages and are asked to read aloud and retell the passage. Measures advanced phonics and word attack skills, accuracy and fluency with connected text, reading comprehension.	Reading Comprehension Accurate and Fluent Reading of Connected Text	1-6
Daze (DAZE)	Students read a passage with every seventh word replaced by a box containing the correct word and two distractor words. Assesses student's ability to construct meaning from text using word recognition skills, background information and prior knowledge, and familiarity with linguistic properties (e.g., syntax, morphology).	Reading Comprehension	3-6

**Acadience Reading Manual: http://wenatchee.innersync.com/assessment/documents/Acadiencenext_assessmentmanual.pdf*

Appendix E: Determining Effect Size Benchmark

A commonly used metric for identifying the strength of treatment effects is Cohen's (1998) definition, in which effect sizes are categorized as small (0.2), medium (0.5), and large (0.8). Some studies have criticized the wide use of Cohen's categories, arguing for a more targeted approach in which the effectiveness of interventions is benchmarked against an average of the effect sizes generated from similar interventions, rather than Cohen's broad categories spanning many types of interventions (Lipsey et. al, 2012; Hill, Bloom, Black, Lipsey, 2007). In other words, the strength of an intervention should be measured based on whether its effect size is at, above or below those of similar programs.

One challenge to using this alternative approach is that there are several different ways to create a benchmark, including creating a benchmark based on interventions with similar outcome measures, intervention types, and intervention targets, to name just a few. Depending on which method is selected, the benchmark could look very different. For example, researchers at the Institute of Education Sciences (IES) reviewed 829 effect sizes from 124 education research studies conducted on K-12 students and reported an array of different effect size distributions that can provide insight into what constitutes a large or small effect relative to similar education evaluation studies (Lipsey et. al, 2012). They provide the following benchmarks to be used as normative comparisons:

- ***Benchmark by outcome measure***. IES researchers looked at the type outcome measures (i.e., did researchers use a self-developed outcome measure, a general standardized outcome measure like an IQ test, or a subject-specific standardized outcome measure like a reading or math test) by grade level and found that the average effect size for education research studies evaluating elementary students with a standardized subject test (like the Acadience Reading literacy tests) was **.25**.
- ***Benchmark by intervention type***. One metric for evaluating effect size was based on the type of intervention under investigation. Researchers sorted the interventions of reviewed studies into several broad categories (e.g., a whole school program, a teaching technique, a new instructional format, skill training, or an instructional program). EISP was closest to an instructional program. Average effect size for research studies that evaluated a comprehensive instructional program such as EISP was **.13.**
- ***Benchmark by intervention target***. A final yardstick to contextualize effect sizes focused on the targeted group of the intervention (e.g., individual students, small group, classroom, whole school, mixed.) that targeted individual students had average effect sizes of **.40**. Interventions that targeted individual students had the highest observed effect sizes, on average.

For the purposes of this report, we chose to compare the effect sizes in our study by averaging the three effect size benchmarks described above. The average effect size benchmark was .26.

Appendix F. Program Use by Vendor and Grade

Table F1 presents a comprehensive summary of usage for each vendor and grade. The table includes usage frequencies, such as average weekly minutes of use, average total minutes of use, average number of weeks of use, and the percentage of students who met vendors' recommendations for average minutes of use, total weeks of use, and a combination of average minutes and total weeks of use. We included information on student who met the dosage recommendations as vendors described, and those who met a relaxed version of their recommendations (e.g. 80% students who reached at least 80% of the recommendations).

Table F1. Program Use by Vendor and Grade

			Program Use			**Met Dosage Recs**			**Met Relaxed Version of Dosage Recs**					
	Grade	N	Ave Wkly Min.	Ave Total Min.	Ave Wks. of Use	% Met Wks. Recs	% Met Ave Min. Recs	% Met Min. & Wks. Recs	Met 80% Ave Min. Recs		Met 80% Wks. Recs		Met 80% Min./ 80% Wks. Recs	
Waterford	K	3469	55	1585	27	65%	35%	30%	2202	63%	2787	80%	2007	58%
	1	3102	63	1753	26	67%	21%	17%	1653	53%	2344	76%	1471	47%
	2	489	52	1299	23	43%	15%	13%	156	32%	297	61%	134	27%
	3	95	40	716	14	19%	1%	0%	10	11%	34	36%	7	7%
	Total	**7155**	**58**	**1627**	**26**	**63%**	**27%**	**23%**	**4021**	**56%**	**5462**	**76%**	**3619**	**51%**
Imagine Learning	K	12542	40	1033	24	82%	43%	40%	6439	63%	8811	87%	6012	59%
	1	15664	48	1341	27	90%	53%	50%	8801	73%	11285	93%	8409	69%
	2	4656	47	1259	26	83%	49%	46%	2345	70%	2913	87%	2223	67%
	3	4273	44	1058	23	73%	41%	35%	1952	65%	2396	80%	1703	57%
	Total	**37135**	**44**	**1192**	**25**	**85%**	**48%**	**44%**	**19537**	**68%**	**25405**	**89%**	**18347**	**64%**
Core5	K	14256	53	1367	24	75%	63%	55%	10481	74%	11775	83%	9486	67%
	1	15788	62	1847	29	89%	75%	69%	13573	86%	14504	92%	12721	81%
	2	4574	62	1814	28	86%	59%	54%	3520	77%	4140	91%	3283	72%
	3	4766	57	1623	27	83%	55%	50%	3546	74%	4195	88%	3299	69%

			Program Use			Met Dosage Recs			Met Relaxed Version of Dosage Recs					
	Grade	N	Ave Wkly Min.	Ave Total Min.	Ave Wks. of Use	% Met Wks. Recs	% Met Ave Min. Recs	% Met Min. & Wks. Recs	Met 80% Ave Min. Recs		Met 80% Wks. Recs		Met 80% Min./ 80% Wks. Recs	
	Total	**39384**	**58**	**1642**	**27**	**83%**	**66%**	**60%**	**31120**	**79%**	**34614**	**88%**	**28789**	**73%**
i-Ready	K	1462	37	748	19	43%	64%	39%	1215	83%	771	53%	749	51%
	1	2236	44	1099	24	38%	43%	27%	1459	65%	1377	62%	1095	49%
	2	888	43	1077	23	38%	40%	27%	540	61%	526	59%	426	48%
	3	880	42	989	23	29%	35%	17%	561	64%	496	56%	406	46%
	Total	**5466**	**41**	**984**	**22**	**38%**	**47%**	**28%**	**3775**	**69%**	**3170**	**58%**	**2676**	**49%**

Note. K-1 Data source: vendor usage data before cleaning invalid SSIDs, duplicates, missing data, contamination with other programs, etc.
Grades 2-3 Data Source: vendor usage data after cleaning invalid IDs and missing outcome data (in order to exclude students reading at grade level)

Evaluation and Training Institute
100 Corporate Pointe, Suite 387
Culver City, CA 90230
www.eticonsulting.org

For more information on the
Evaluation and Training Institute, contact ETI:

Jon Hobbs, Ph.D., President
Phone: 310-473 8367
jhobbs@eticonsulting.org