Utah's Early Intervention Reading Software Program

2016-2017 K-3 Program Evaluation Results

Submitted to the Utah State Board of Education
September 2017

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

All correspondence should be directed to:
Jon Hobbs, Ph.D.
jhobbs@eticonsulting.org

Table of Contents

List of Figures

List of Tables

Acronyms

BOY	Beginning-of-Year
C	Control group/non-program group
EISP	Early Intervention Software Program
EOY	End-of-Year
ES	Effect Size
ETI	Evaluation and Training Institute
ITT	Intent to Treat
LEA	Local Education Agency
NS	Non-significant statistical coefficient
OLS	Ordinary Least Squares analyses
OPTI	Optimal Use sample
ROPT	Relaxed Optimal Use
Tr.	Treatment group/program group
USBE	Utah State Board of Education

Executive Summary

Evaluation Purpose

The Early Intervention Software Program (EISP) was established in 2012 to improve the literacy achievement of K-3 students in Utah through their use of adaptive computer-based literacy software programs. In 2016-2017, Local Education Agencies (LEAs) selected among seven unique software programs to use with their students. LEAs were expected to use the program in Grades K-1 with all students and as an intervention in Grades 2-3. As the EISP external evaluator, the Evaluation and Training Institute (ETI) studied two aspects of the EISP: 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 seven software programs (program-wide) and between each software vendor (vendor-specific).

Program Implementation Findings

Each software program vendor provided Local Education Agencies (LEAs) with dosage (program use) recommendations to increase student literacy achievement. In the 2016-2017 program year, meeting the program vendors usage recommendations continued to be a challenge; however, there was an increase in students' use from last year. The percentage of students who met the total weeks recommendations varied widely among vendors: slightly more than half of the vendors (4 out of 7) had between 65-72 percent of their students meet the vendor recommendations for total weeks of use, while the remaining three vendors had only 30-41 percent meet their recommendations for weeks of use.

Program-wide Impacts Findings

The strength of the program effects increased with students' program use, and, when examining outcomes for students who had the highest program use, our findings show that students across all four grade levels (K-3) had statistically significant differences in literacy achievement when compared to a similar group of non-program ("control") students (**Table 1**). While literacy achievement was higher for students who used the program than those who did not, the overall strength of program effect was highest in kindergarten (ES=.2), with diminishing returns in other grades. We also found that the program was more effective for students with specific characteristics or school environments, including those who were female, low-income, classified as special education, from Title I schools or English Language Learners (ELL).

Table 1. Predicted Means of EOY Composite for Matched Treatment (Tr.) and Control (C), Program-Wide, by Dosage Level

	Kindergarten			1st Grade			2nd Grade Intervention			3rd Grade Intervention		
	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES
Highest Use	N=7,126			N=9,238			N=1,772			N=1,322		
	156	141	**.2*** **	214	201	.13***	174	160	**.18*** **	273	261	**.14** **
	15,466			19,732			5,172			4,714		
	147	138	.11***	196	191	.05***	156	150	.08**	NS		
Lowest Use	27,482			27,020			7,662			8,618		
	144	138	.06***	NS			NS			NS		

* p ≤ .05. ** p ≤ .01. *** p ≤ .001.

Note. NS (not significant) in a cell means the program did not have a significant effect. ES: Effect Size (based on Cohens D). Low use: includes all students; the second highest use: students met at least 80% of vendors recommended dosage; Highest use: students met vendors' recommendations for at least 80% of the weeks it was used, and used it for the total weeks recommended by vendors.

Note: ES's greater than .13, the average for similar intervention programs, are highlighted in bold.

Vendor Impacts Findings

Our analyses of the relationship between hours of use and literacy scores supported the program-wide findings. More vendors produced statistically significant positive effects in the earlier grades levels (four to five vendors) than in upper grade levels (one to two vendors), however, the strength of these findings varied by vendor and grade, and we need to be cautious when comparing vendors because of issues related to small sample sizes for specific grades. When we compared program students who met a minimum program dosage threshold to a matched control group of students with similar characteristics, we found that multiple vendors had an impact on increased literacy achievement in kindergarten as measured by their mean literacy composite scores, but only a few vendors had an impact in the upper-early grade levels. To measure the strength of these effects, we looked to the average effects sizes produced by similar education intervention programs. In kindergarten, four vendors had stronger effects than the average for similar programs (Waterford, Core5, Imagine Learning, and MyOn), one vendor in first grade (Waterford), two in second grade (Waterford; Imagine Learning); and one in third grade (MyOn).

Overall Conclusions & Recommendations

The 2016-2017 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. When reviewing our current evaluation results with those from

previous years, it is easy to recommend that the program be continued for kindergarten students. It is more difficult to endorse the program's use with students in 1st through 3rd 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. Program-wide (all vendors combined) also showed the program can be effective for upper-early grades, particularly when students met or exceeded a threshold of minimum program use. Schools are doing a better job implementing the program according to vendors' recommendations, but there is still room for improvement. This last issue is particularly important, because our results this year showed a direct and positive relationship between higher levels of program use leading to stronger student outcomes. We believe that if schools could continue improving program implementation, students' benefits would also improve.

2016-2017 Early Intervention Reading Software Program Evaluation Report

Evaluation Purpose & Evaluation Questions

The Utah state legislature established the Early Intervention Software Program ("EISP") to aid in the development of Utah students' literacy skills through computer-based, adaptive reading software programs designed to meet students' unique learning needs. The Evaluation and Training Institute (ETI) conducts an annual evaluation of how the reading software programs were used and how students' participation is associated with literacy achievement, including the combined impact of all the software programs and a comparison of the relative effects on literacy achievement of each of the software providers ("vendors"). This report includes findings from the 2016-2017 academic year, the EISP's fourth year of implementation. These findings are intended to help the Utah State Board of Education (USBE) and Local Education Agencies (LEAs) understand how well the program is working, identify potential areas of improvement in program implementation, and make informed decisions about the future direction of the program.

Research questions are often used to guide the direction of an evaluation and can be useful for facilitating a shared understanding of the overarching evaluation goals. We answer the following general research questions in this report:

1. Did students use the software as intended?
2. Did the program have an overall effect across all vendors?
3. Did the program effects differ based on student or school characteristics?
4. Were there differences in treatment effects among vendors?

The EISP annual reports are disseminated to a wide-audience of stakeholders, including educators, researchers, policy staff and non-technical reviewers, and we structured this report for all types of stakeholders to understand.

We include a brief description of the EISP and 2016-2017 enrollment information in the next section, and then present a streamlined summary of our research methods. We address each research question in the findings section of this report, which we organized based on the two study objectives: program implementation and program impacts. Finally, we summarize the key findings across all the objectives and research questions and discuss the study limitations.

Program Background and Enrollment

Utah passed legislation in 2012 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 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 struggling students in second and third grade. To participate, LEAs (districts and charter schools) submitted applications to the USBE for the use of specific reading software programs prior to the start of each school year.

Seven[1] vendors provided software and training to schools through the EISP in 2016-2017. The seven vendors were (in alphabetical order): Imagine Learning, Istation, Pearson ("SuccessMaker"), Lexia Reading Core5® (Core5), MyOn, Reading Plus and Waterford. These software programs were used in 388 schools and by 86,723 students (**Table 2**). Core5 was the most frequently used program (157 schools, 40,000 plus students), while Istation was used by the fewest schools (7 schools; 889 students).

Tables 2-3 contain information on the 2016-2017 enrollment of LEAs and students who used each vendor. While the EISP is intended for second and third grade intervention students, some educators implemented the program with their entire class, and in these instances, non-intervention students also had access to the software programs. Our report focuses on intervention students in Grades 2-3 only, however, we have provided enrollment information for both types of students in **Tables 2-3** so readers may understand how the program was implemented in practice.

[1] During the 2016-2017 school year the USBE contract with CurriculumAssociates ("i-Ready") was cancelled. As the state no longer had the contract available to the program for future licensing and reporting requests, they were not evaluated this year.

Table 2. Program Enrollment and Overview

Program	LEAs	Schools	Students	
			K-3	K-1 / 2-3 Intervention
Istation	3	7	889	613
Waterford	22	54	7,286	6,415
Imagine Learning	37	152	28,861	21,340
SuccessMaker	7	23	2,080	1,488
Core5	29	157	40,308	23,832
Reading Plus	5	23	1,875	185
MyOn	11	31	5,424	2,801

*Note. Schools could use multiple programs for different grades. Grades 2-3 intervention students included those with scores below benchmark for their grade at beginning of year. We excluded students with usage of less than five minutes.

The percent of participants per grade varied by program, and three vendors had a greater percentage of students who used the program in the third grade than other grades (**Table 3**).

Table 3. Program Enrollment by Vendor and Grade

Program	Kinder	1st	2nd		3rd	
			All	Intervention	All	Intervention
Istation	N=176 20%	199 22%	206 23%	95 16%	308 35%	143 23%
Waterford	3,159 43%	2,998 41%	1,129 15%	258 4%	N/A	N/A
Imagine Learning	7,529 26%	9,478 33%	7,110 25%	2,382 11%	4744 16%	1,951 9%
Success-Maker	283 14%	768 37%	630 30%	234 16%	399 19%	203 14%
Core5	7,563 19%	10,173 25%	11,736 29%	2,958 12%	10,836 27%	3,138 13%
ReadingPlus	N/A	N/A	292 16%	1 (0%)	1,583 84%	184 100%
MyOn	567 10%	1,253 23%	1,667 31%	22,770 26%	1,937 36%	19,807 23%
Total	19,277 22%	24,869 29%	22,770 26%	6,344 11%	19,807 23%	6,184 11%

*Note. Grades 2-3 intervention students included those with scores below benchmark for their grade at beginning of year.

Usage Recommendations

Each vendor provided recommendations for using the software programs in order for it to have an impact on student 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 12 to 28 weeks. To encourage LEAs to implement the program as it was intended, at least 80 percent of their students had to meet a relaxed version of the vendors dosage recommendations within two years of implementation[2].

Table 4. Vendor 2016-2017 Minimum Dosage Recommendations

Program	Kindergarten ALL Student	First Grade ALL students	Second Grade Intervention Students	Third Grade Intervention Students	Suggested Minimum Weeks
Istation	60 min/week	60 min/week	60 min/week	60 min/week	28 weeks
Waterford	60 min/week	80 min/week	80 min/week	80 min/week	28 weeks
Imagine Learning	45 min/week	60 min/week	60 min/week	60 min/week	20 weeks
Success-maker	45 min/week	45 min/week	60 min/week	60 min/week	15 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
Reading Plus	45 min/week	45 min/week	45 min/week	45 min/week	15 weeks
MyOn	45 min/week	45 min/week	45 min/week	45 min/week	20 weeks

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

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

Evaluation Methods

We provide an overview of our research methods, samples and data sources that were used to answer each research question. We present information about our methods in question format for clarity. **Appendices A-C** provide additional details on our methods, data processing procedures and sample.

What sources of data were used in our analyses?

We collected data from ten different sources to create our master dataset for the EISP analyses. The data sources included: seven program vendors, who provided us with usage information for each student who used their programs; state Dynamic Indicators of Basic Early Literacy skills (DIBELS Next) testing data from two online reporting systems (DMG and AMPLIFY); and student information system (SIS) demographic data provided by the Utah State Board of Education (USBE).

Which instruments did we use to measure literacy achievement?

We measured literacy achievement using the DIBELS Next, which was administered in schools throughout the state in Grades K-3. The DIBELS Next measures were used throughout Utah, and are strong predictors of future reading achievement. DIBELS Next 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), DIBELS Oral Reading Fluency (DORF), and DAZE. In addition to scores for the six subscale measures described above, we used overall composite scores and benchmark levels, or criterion-reference target scores that represent adequate reading progress.

How did we create our analytic samples?

Our samples changed based on the specific analyses goals, or out of necessity in response to barriers found with the data, such as small enrollment numbers for specific vendors. 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.

Program-Wide Samples. We created three usage groups to study the effects of increased program use on students' test scores across vendors. Each program vendor provided schools with a recommendation for how much time students should use the program before benefits are observed. This minimum use recommendation was an important predictor of literacy achievement, and we wanted to determine how student use characteristics effected their outcomes. We operationally defined the combination of

weekly use and weeks of use as "program dosage". We created three program-wide samples to determine the effects of program dosage on students' achievement:

- The **intent to treat** (ITT) sample was comprised of all students who used the program for any amount of time, and shows how effective the program was irrespective of use. Students in this sample had the lowest average program dosage.
- The **relaxed optimal** (ROPT) use sample was comprised of students who used the program greater than or equal to 80% of vendors' recommended use. Students in this sample had the second highest average program dosage.
- The **optimal use** (OPTI) sample was comprised of students who met the vendors recommended use (in minutes) [3] for at least 80% of the weeks the software was used. In addition, students must have used the software for at least the minimum number of weeks suggested by each program vendor. Students in this sample had the highest program dosage.

Individual Vendor Samples. For the individual vendor analyses in which a *matched control group* was used, our individual vendor samples were not large enough to study the program effects of students who met vendors' exact usage recommendations (e.g. optimal usage), and we studied a subset of students who met a relaxed version of vendors' recommendations instead (ROPT). We used the ITT sample (all students, regardless of use) when we had a low ROPT sample for certain vendors and grades, and we wanted to see if any effects could be found with a larger sample of students. We identified all instances in which the ITT samples were used in our findings. In addition, we used the ITT samples for all seven vendors when we analyzed the relationship between hours of program use and literacy outcomes, which did not require a control group or for students to meet a minimal usage criteria.

Both program-wide and individual vendor samples were matched to groups of students who did not use the program ("control group") 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, achievement level, gender, race, and poverty status). 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. We created three matched samples for each usage group (ITT, ROPT, OPTI) for the program-wide analyses. Similar to our program-wide approach, we created seven matched samples for each program vendor, which allowed us to have tightly matched control groups for each program vendor. *(see **Appendix A** for more information on CEM and how it was used to match students).*

[3] "Met the vendors recommended use (in minutes)" is equal to 80% of the recommended weekly minutes. For example, if a vendor recommended 60 minutes, the student must have used the program for at least 48 minutes.

What statistics do we provide in our results?

Where appropriate, we provided mean scores for our treatment and control groups, which are meaningful when comparing treatment and control groups from the same sample. We also provided treatment effect sizes (ES; based on Cohen's Delta[4], 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.

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). For the purposes of this study, we have chosen to contextualize our findings using similar instructional programs as our benchmark. The mean effect size for similar instructional programs is **.13**, 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.[5] More information on how we selected our ES benchmark is provided in **Appendix E.**

How did we study program use?

Vendors provided us with usage data, including: software use (in minutes) for each week the program was used from the beginning to the end of the school year, and total minutes of use. Having usage data reported by week enabled us to identify the number of weeks a student used the software, and calculate average weekly use. A student met fidelity if, on average, he or she used the software for at least 80% of the vendors recommended average minutes of use or 80% of the total weeks of use (see **Table 4**, Vendor Dosage Recommendations).

How did we study the impacts across all vendors?

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. A two-level random intercept statistical model with school as the level-2 predictor was used to predict student outcomes. 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

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

[5] 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).

student-level predictors, and, at the school-level, controlling for Title 1 status. In general, non-significant predictors were removed from statistical models to increase the variance we could explain with the significant predictors of achievement.

How did we study individual vendor impacts?

We conducted three types of analyses to determine the impacts of each software program on student literacy achievement:

1. We conducted a **usage effects analysis**, and measured the relationship between students' program use (in hours) and DIBELS Next composite scores and literacy scales for an ITT treatment sample;
2. We conducted a **between group mean score analysis** for treatment and control group students in each vendors' ROPT sample (and used the ITT sample when ROPT samples were too small to detect program effects for certain programs and grades). We used an Ordinary Least Squares (OLS) regression model to predict the differences in mean scores while controlling for demographic characteristics and baseline scores.
3. We conducted **benchmark analyses** for treatment and control group students in each vendors' ROPT sample, or the ITT sample when ROPT samples were too small using descriptive statistics.

Findings

We evaluated two facets of the EISP: program implementation and its impacts on student learning. Our program implementation findings focused on program usage in relationship to its intended use, as described through vendors dosage recommendations. The impact findings included an examination of the program as whole (all vendors combined) as well as impacts for individual vendors. We present the findings for both facets of our findings by research question for ease of interpretation and clarity.

Program Implementation

It is often important for evaluators to study program implementation prior to measuring program impacts. 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 is students' dosage, or program use, as students must use the program for long enough for it to have an impact on their literacy skill development. We explored the differences in usage across software programs

Key Finding: Although our findings highlight a need for continued improvement in students' fidelity of program use, students' use has improved since last year: 41% of students met vendors average weekly use last year compared to 55% this year, while 60% of students met the total weeks recommendations last year, compared to 77% of students this year.

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. In addition, we relaxed the dosage to 80 percent of vendor recommendations in order to adjust vendors' recommendations to the competing activities during the school year (including student vacation time, interruptions in computer lab access due to state testing, and other factors). A more detailed summary of student use is included in **Appendix B**.

Did students use the software as intended?

A higher percentage of students were able to use the program for the total number of weeks recommended by vendors, but students still struggled to use the programs for the average recommended minutes per week. Less than half of the students in the program met the vendors' average weekly use recommendations for their software (42% of students overall), but more than half met vendors' recommendation for total minimum weeks of use (67% of students overall). Using the "real world" relaxed minimum use approach, 60 percent of students met vendors average minutes of use recommendations, and 77 percent met the weeks of use recommendations (**Figure 1**). When combining the two usage recommendations, only 53% of students met both the relaxed average use and relaxed total weeks of use recommendations (**Figure 2**).

When we review the findings for each of the program vendors, we see a wide variety of variation among the students using each program. Core5 and SuccessMaker were the vendors with the highest percentage of students to meet vendors' average weekly use recommendations[6], with 52% and 43% of students using the programs as they were intended, respectively, and were also among the top three vendors whose students met the total weeks recommendations. (**Figure 1**). Imagine Learning had the highest percentage of students to meet the total weeks recommendations (72%). Istation, MyOn and ReadingPlus were the vendors with the fewest students to meet both categories of the dosage recommendations (average weekly use and total weeks).

[6] Program dosage recommendations varied by software vendor, and the percentages depicted in these figures were based on the students who met vendors specific recommendations. These figures were not designed to show students with the highest overall program use.

Figure 1: Students who met vendors minimum dosage recommendations

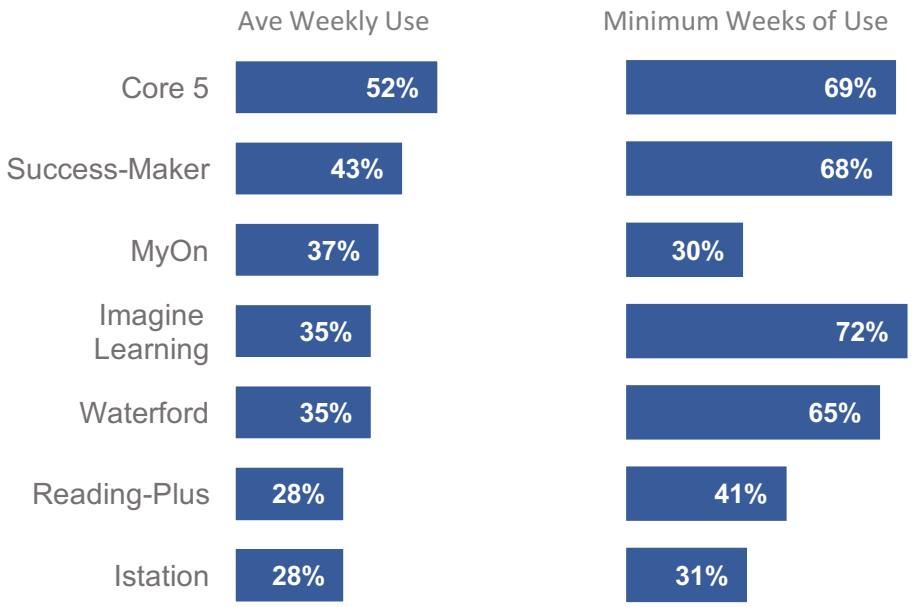
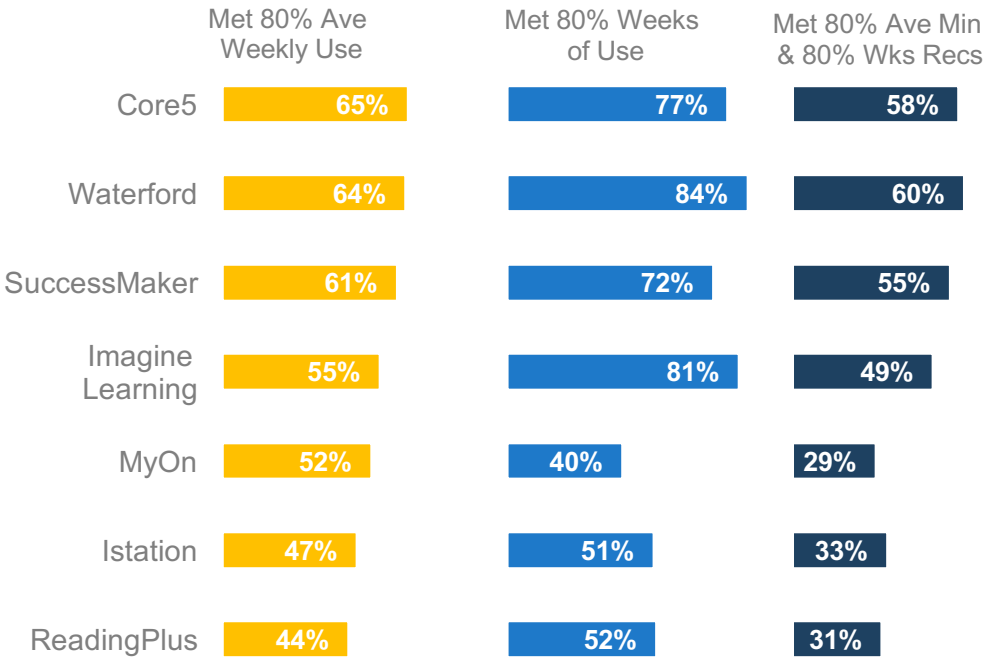


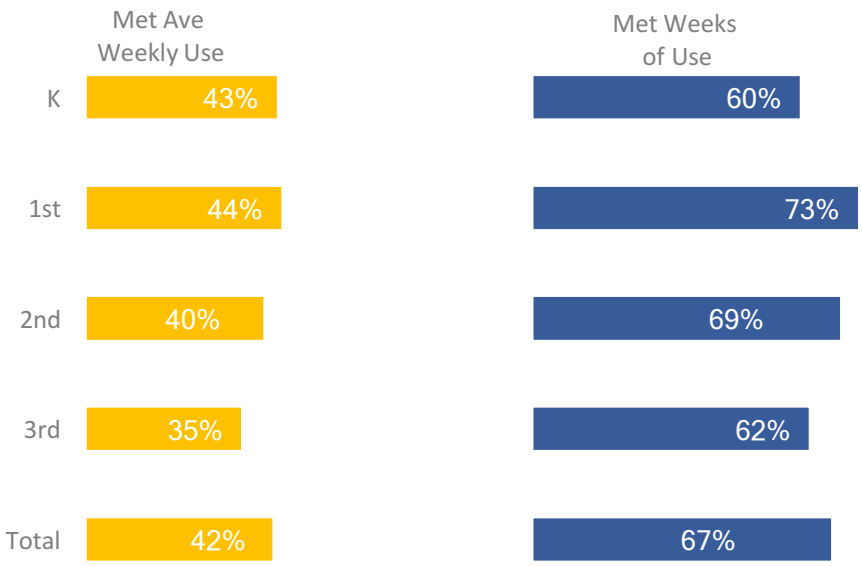
Figure 2 depicts students who met a relaxed dosage recommendation. Similar to our previous findings, Core5 and SuccessMaker were among the top three vendors whose students met either the average weekly use or total weeks of use requirements. However, after relaxing the dosage recommendations, Waterford was the vendor with the highest percentage of students (60%) who met both types of dosage recommendations.

Figure 2: Students who met at least 80% of the minimum dosage recommendations



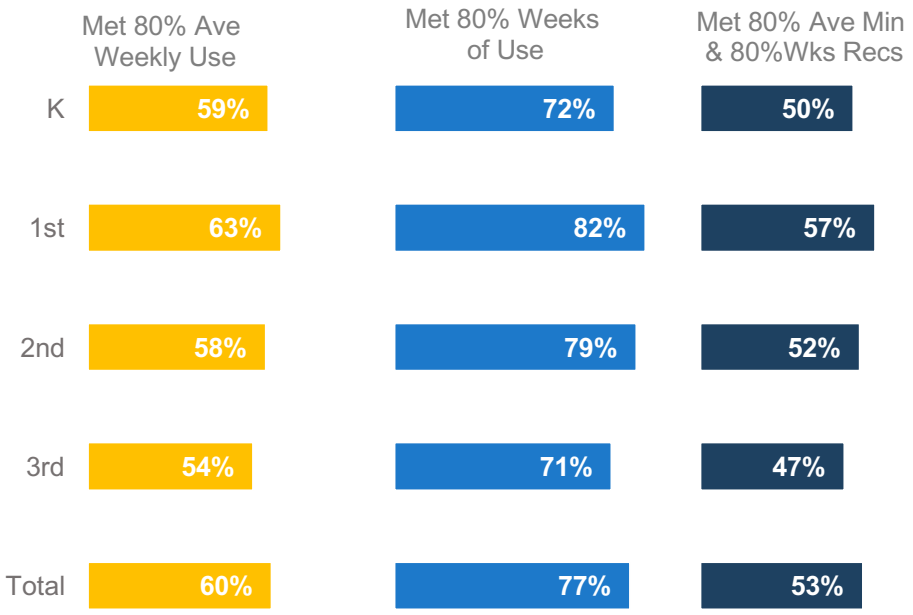
Additionally, we examined program usage by grade in Figures 3 and 4. In **Figure 3** we can see that there were very few differences in fidelity of use among grades. However, we found that more first grade students met both the average minutes and weeks recommendations (57%), compared to other grade levels, and that kindergarten and third grade students were the grade levels with the fewest students who met the average weekly use or total weeks recommendations.

Figure 3: Students who met the dosage recommendations by grade



When we relax the dosage recommendations, first grade continues to have the highest overall fidelity of use among all the grades (**Figure 4**). In addition, over half of the students in each grade met the average minutes and weeks recommendations, except for third grade, in which 47% of students met the recommendations.

Figure 4. Students who met 80% of the dosage recommendations by grade



Literacy Achievement Results

We evaluated the EISP's effectiveness by comparing the literacy achievement of groups of students who used the program to groups of students who did not use the program. Our evaluation results are presented in two sections: 1) Program-wide impacts, and 2) Individual vendor impacts. The program-wide analyses measured the impact of the EISP across all seven software programs, providing a big-picture view of how the program performed. In the individual vendor impacts section, we explore the relative impacts each program vendor had on literacy achievement.

Program-Wide Analyses

We begin the program-wide analyses studying the program impacts for three samples representing different levels of program use (from lowest to highest use). This analysis helps illustrate the relationship between program effects and program use (or dosage) and depicts program effects for literacy composite scores for each grade. Following this analysis, we examine the program effects on individual literacy subscales for the highest usage group, then determine how the program affects changes in students' benchmark status, an indication of students reading risk. We completed our analyses with an examination of program effects for specific groups of students.

Did the program have an overall effect across all vendors?

Dosage (or amount of software use) is the most important determinate in program-wide treatment effects. As seen in **Figures 4 - 7**, the statistically significant program-wide effects on DIBELS Next end-of-year (EOY) composite scores increase with dosage, and the more a student uses the program the better his/her EOY outcomes.

- In kindergarten, the treatment effects double when you move from the lowest dosage (Intent to Treat) to the second highest dosage, and triple when you go from the lowest to the highest dosage usage groups.
- In first grade, students in the highest dosage group have slightly more than three-fold the effects size when compared to the second highest dosage group (ROPT).
- In second grade, students in the highest dosage group have double the increase in the treatment effect size when compared to the second highest dosage group (ROPT).
- In third grade, only the highest dosage group produced a statistically significant effect.

Students with the highest program dosage in kindergarten and second grade had the highest treatment effect sizes overall, as measured by their average DIBELS Next Composite scores (ES: .2 and .18, respectively).

Figure 5. Kindergarten: Means of EOY Composite for Matched Program & Non-Program Means, by Usage Level

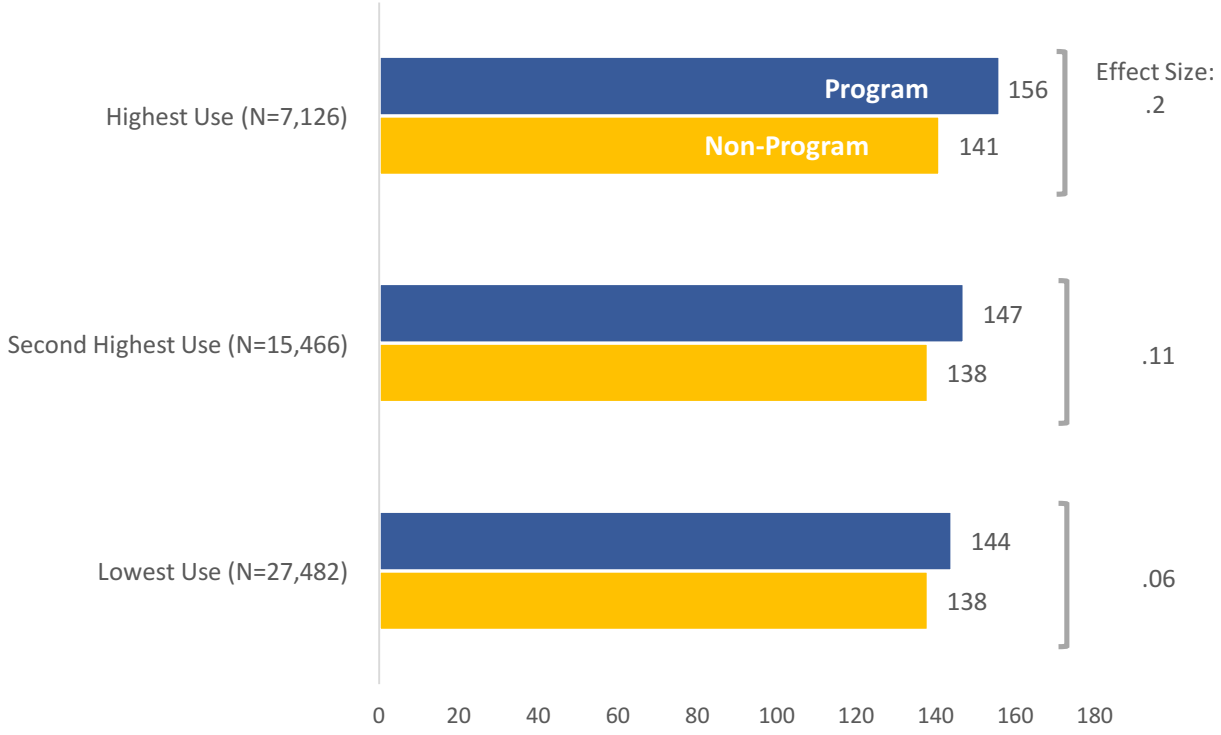


Figure 6. First Grade: Means of EOY Composite for Matched Program & Non-Program Means, by Usage Level

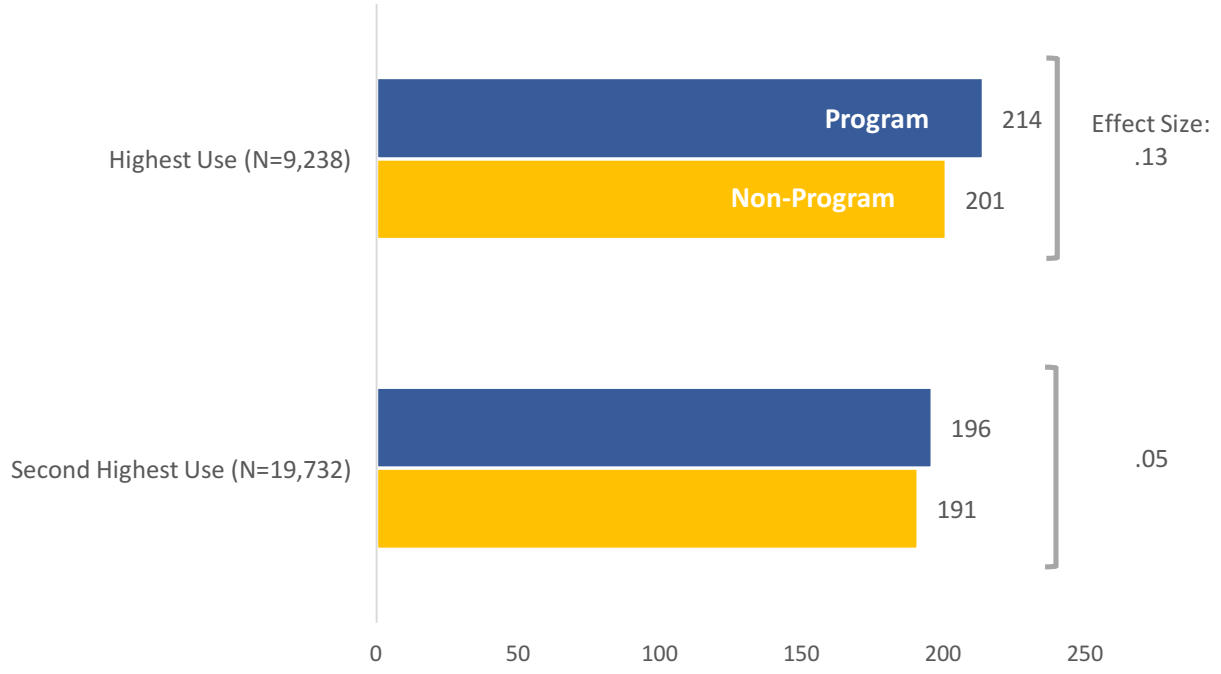


Figure 7. 2nd Grade: Means of EOY Composite for Matched Program & Non-Program Means, by Usage Level

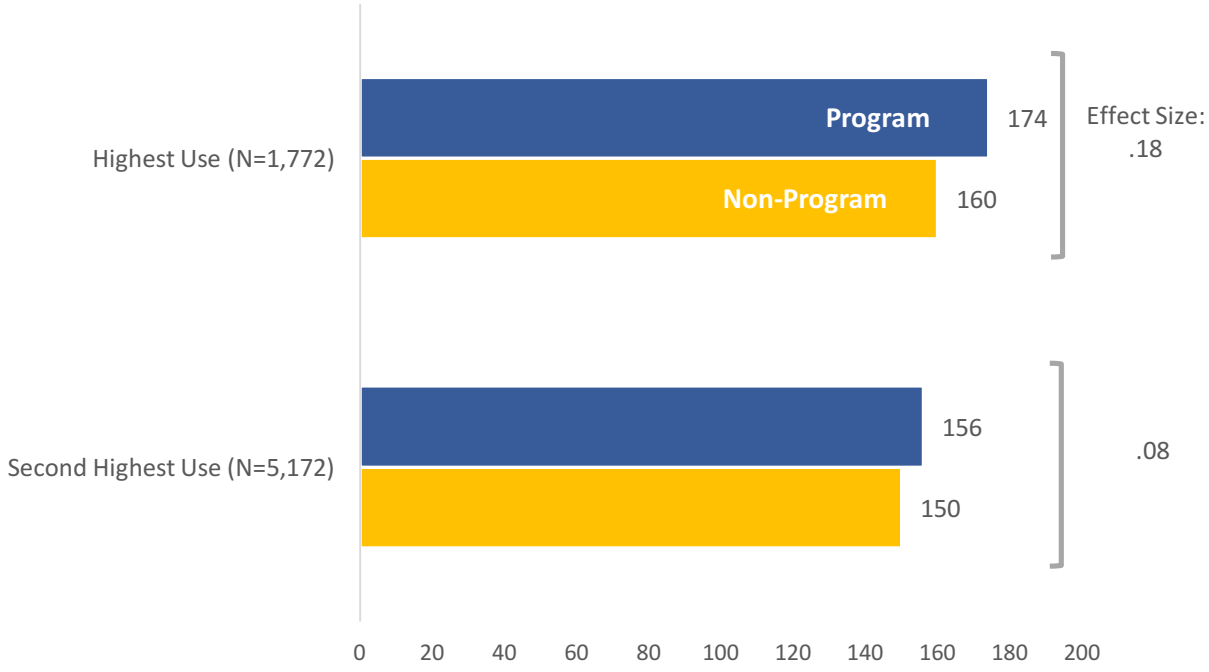
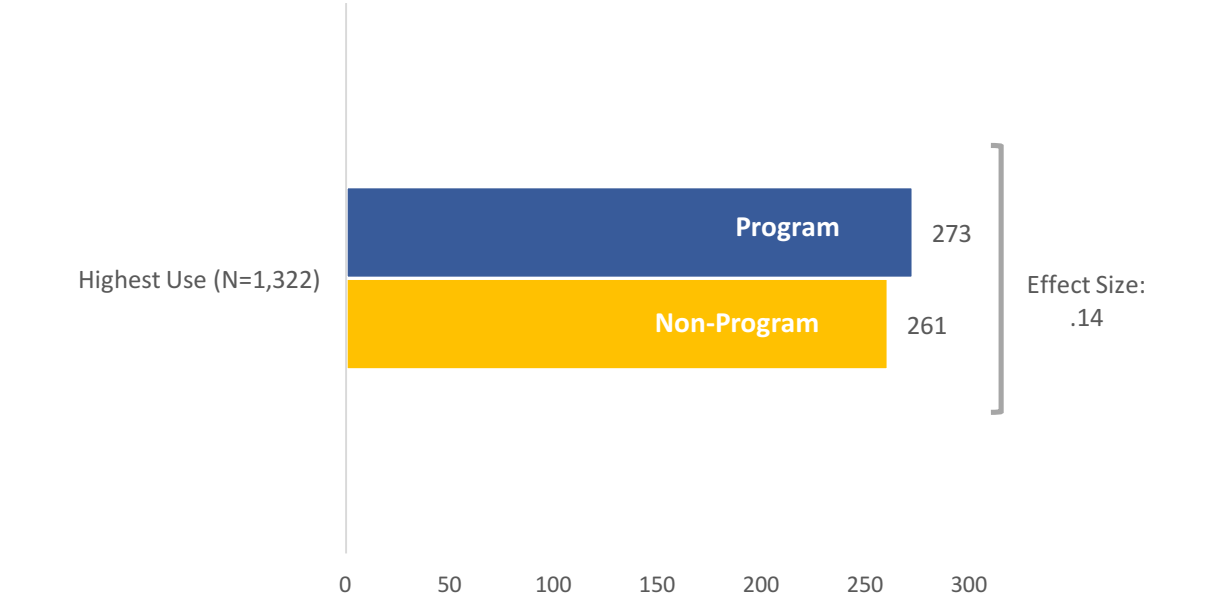


Figure 8. 3rd Grade: Means of EOY Composite for Matched Program & Non-Program Means, by Usage Level



The program-wide impacts for the literacy domains varied based on the specific measure and grade. The predicted mean scores in **Table 5** show that kindergarten students had higher test scores than their matched control student counterparts across all literacy domains, but that the differences between the average group mean scores were small. These trends do not hold up for all grades, however, where treatment effects were present only on certain subscales. The program produced positive effects for three of the five literacy subscales in first grade, one out of three in second grade, and two out of three in third grade. When interpreting the practical application of these findings, it should be noted that several subscales produced larger effects compared to similar intervention programs, including two reading fluency domains in kindergarten (LNF and NWF: CLS); oral reading fluency in first and second grade, and the DAZE scale in third grade.

Table 5. Predicted Means of EOY DIBELS Scales for Matched Treatment and Control, Program-Wide, Highest Use sample

DIBELS Scale	Kindergarten N=7,126			1st Grade N=9,238			2nd Grade N=1,772			3rd Grade N=1,322		
	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES
First Sound Fluency (FSF)	38***	36	0.08		N/A			N/A			N/A	
Letter Naming Fluency (LNF)	53***	49	**0.18**		NS			N/A			N/A	
Phoneme Segmentation Fluency (PSF)	53**	51	0.06		NS			N/A			N/A	
Nonsense Word Fluency-CLS	49***	42	**0.19**	91***	85	0.12		NS			NS	
Nonsense Word Fluency-WWR	8.75***	7.39	0.12	28***	26	0.08		NS			N/A	
Oral Reading Fluency		N/A		74***	70	**0.15**	59***	55	**0.16**	76*	74	0.11
DAZE		N/A			N/A			N/A		14***	13	**0.23**

*p ≤ .05. ** p ≤ .01. *** p ≤ .001.
Note. NS (not significant) in a cell means the program did not have a significant effect. ES: Effect Size (based on Cohens D).
Note: ES's greater than .13, the average for similar intervention programs, are highlighted in bold.

Benchmark Analyses

We conducted a review of DIBELS Next benchmark rankings for treatment and control students from beginning of the year (BOY) to end of the year (EOY). We focused on students who performed below grade level at BOY to determine the extent of their growth at EOY. **Table 6** shows the percent of students who started the year Well Below Benchmark (lowest category) or Below Benchmark for their grade (second lowest category), and followed their upwards growth compared to non-program students at year-end.

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

As shown in **Table 6**, kindergarten students who began the year with scores below their peers for their grade experienced greater upwards mobility by year-end compared to a matched control group (10% difference in growth). Although students in the upper grade levels also experienced greater growth compared to their non-program counterparts, the difference in growth was minimal (from 2-3% depending on the grade).

Table 6**.** Treatment and Control Comparison of Change in DIBELS Next Benchmark Levels for At-risk Students, Program-wide

	Group	Well Below to Below Bench		Well Below to Bench		Below to Bench		WB or Below to Bench	
K	Non-Program	422	24%	646	37%	752	54%	1398	45%
	Program	427	25%	851	50%	873	65%	1724	55%
	Tr-C Difference	5	1%	205	13%	121	11%	326	10%
1st	Non-Program	378	16%	658	28%	878	57%	1536	39%
	Program	442	19%	677	29%	932	59%	1609	41%
	Tr-C Difference	64	3%	19	1%	54	2%	73	2%
2nd	Non-Program	293	16%	144	8%	329	41%	473	18%
	Program	315	18%	167	9%	361	44%	528	20%
	Tr-C Difference	22	2%	23	1%	32	3%	55	2%
3rd	Non-Program	290	17%	221	13%	360	53%	581	25%
	Program	285	17%	274	16%	372	57%	646	27%
	Tr-C Difference	-5	0%	53	3%	12	4%	65	3%

Did the program effects differ based on student or school characteristics?

Certain student groups received slightly more benefit from the program, as evidenced by their slightly higher average composite test scores at program exit. Program students who were female, low-income, classified as special education, from Title I schools or English Language Learners (ELL), all did statistically stronger compared to a control group. These differential effects were the most pronounced in kindergarten. For example, special education program students had an average EOY composite test sore that was 17 points higher than their control student counterparts, but these advantages diminished in later grades.

Table 7. Matched Treatment and Control Group Differences on EOY Composite Scores by Subgroup and Grade

	Kindergarten	1st Grade	2nd Grade	3rd Grade
Females	+11	+4	+7	NS
Low Income	+13	+3.36	+6	NS
Special Education	+17	+7	+10	NS
Title I Schools	+8.44	+2.71	+4.74	NS
English Language Learners	+15	+2.92	+2.79	NS

Note. NS (not significant) in a cell means the program did not have a significant effect.

Vendor-Specific Analyses

The vendor-specific analyses were designed to help program stakeholders understand the effectiveness of the individual programs and make informed decisions. With this in mind, we have done our best to conduct comprehensive analyses in which readers may understand program effectiveness based on different aspects. We must also stress that differences within program vendors samples (e.g. sample size, types of students who used the programs, etc.) make it difficult to conduct a fair comparison among vendors. To help the reader understand these limitations, we indicate when different samples are used in our findings and discuss these limitations in the beginning of sections (where applicable) and at the conclusion of the report.

The vendor-specific findings in this section include: 1) a mean comparison between each program and a matched control group that shows program effects on overall literacy scores and subscales; 2) analyses of the relationship between time (in hours) and literacy outcomes; and, 3) a descriptive analysis in which we examine how the vendors effect at-risk students through upward movement in benchmark status.

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

Between Group Analyses

Table 8 presents the OLS regression results for each program and grade. A majority of programs had a positive impact on students in kindergarten (four of six), followed by two vendors in first, second and third grade. The vendors that produced effect sizes greater than our effect size benchmark were: Waterford (ES: .18), Imagine Learning (ES: .25), Core5 (ES: .28), and MyOn (ES: .29) in kindergarten, Waterford (.54) and Imagine Learning (ES: .15) in second grade, and MyOn in third grade (ES: .24). Waterford and Imagine Learning had an impact on three of the four grades.

Table 8. Predicted Means of EOY Composite for Matched Treatment and Control, by Vendor, OLS Regression Model, Mixed samples

	Kinder			1st			2nd			3rd		
	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES
Istation	N=322‡		NS	N=179‡		NS	N=176‡		NS	N=236‡		NS
WF	N=2,484			N=2,314			N=142			N/A		
	144*	137	**.18**	198*	190	.13	162**	136	**.54**			
IL	N=6,162			N=8,880			N=2,098			N=1,360		
	144***	134	**.25**			NS	160**	151	**.15**	244*	236	.12
SM	N=252		NS	N=764		NS	N=118		NS	N=118		NS
Core5	N=6,610			N=9,712			N=2,858			N=2,844		
	153***	142	**.28**	207***	200	.12			NS			NS
RP	N/A			N/A			N/A			N=170‡		NS
MyOn	N=208‡			N=354			N=284‡			N=320		
	147*	134	**.29**			NS			NS	280*	261	**.24**

*p ≤ .05. **p ≤ .01. ***p ≤ .001.
Note*:* Model covariates were gender, Hispanic, special education, school Title I status, and BOY Composite score. †. Marginal predicted means generated from the ITT group. NS (not significant) in a cell means the program did not have a significant effect. ES: Effect Size (based on Cohens D).
Note: ES's greater than .13, the average for similar intervention programs, are highlighted in bold.

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

DIBELS Next 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 DIBELS Next 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.

Benchmark Analyses

Tables 9-12 present changes in students' benchmark status from beginning of year (BOY) to end of year (EOY) using DIBELS Composite scores. Gains made by treatment students were compared to control students ("T-C Difference"). The frequencies and percentages reported in each cell present the number and percentage of students who began the year below or well below (WB) benchmark and moved up one or two benchmark levels by the end of the year. The final column in the tables depict the number and percentage of students who started below their grade level (either below or well below), but ended the year at grade level (at or above bench).

For most vendors, the DIBELS Next benchmark analyses revealed higher rates of positive change in benchmark status from beginning-to-end-of-year among at-risk treatment students when compared to non-program students. We also observed similar trends in benchmark status growth rates between grade levels among most vendors: Kindergarten students experienced the highest difference in growth rates between treatment and control students (a 15% difference), while the T-C differences gradually decreased as students moved into upper grades (only a 4% difference by the 3rd grade).

We list the two vendors in each grade level with the highest growth among students who started the year below grade level (Below or Well Below Bench) and ended the year at grade level (At or Above Bench) relative to their control student counterparts:

- Kindergarten:
    - 59% of Core5 students moved from "Well Below Benchmark" or "Below Benchmark" to "At Benchmark" at EOY vs. 44% of non-program students (a 15% difference).
    - 55% of Imagine Learning program students moved from "Well Below Benchmark" or "Below Benchmark" to "At Benchmark" at EOY vs. 43% of non-program students (a 12% difference).
- 1st Grade:

- 52% of SuccesMaker students moved from “Well Below Benchmark” or “Below Benchmark” to “At Benchmark” at EOY vs. 43% of non-program students (a 9% difference).
- 43% of Waterford program students moved from “Well Below Benchmark” or “Below Benchmark” to “At Benchmark” at EOY vs. 37% of non-program students (a 6% difference).

- 2nd Grade:
  - 20% of Waterford students moved from “Well Below Benchmark” or “Below Benchmark” to “At Benchmark” at EOY vs. 13% of non-program students (a 7% difference).
  - 22% of Imagine Learning program students moved from “Well Below Benchmark” or “Below Benchmark” to “At Benchmark” at EOY vs. 18% of non-program students (a 4% difference).
- 3rd Grade:
  - 27% of Imagine Learning program students moved from “Well Below Benchmark” or “Below Benchmark” to “At Benchmark” at EOY vs. 23% of non-program students (a 4% difference).
  - 26% of Core5 program students moved from “Well Below Benchmark” or “Below Benchmark” to “At Benchmark” at EOY vs. 24% of non-program students (a 2% difference).

Table 9. Treatment and Control Comparison of Change in DIBELS Next Benchmark Levels for At-risk Kindergarten Students, by Vendor

Kindergarten		Well Below to Below Bench		Well Below to Bench		Below to Bench		WB or Below to Bench	
Waterford	Non-Program	65	25%	94	36%	100	48%	194	41%
	Program	71	28%	126	50%	120	55%	246	52%
	T-C Difference	6	2%	32	14%	20	7%	52	11%
Success-Maker	Non-Program	4	16%	9	36%	16	67%	25	51%
	Program	3	15%	9	45%	27	66%	36	59%
	T-C Difference	-1	-1%	0	9%	11	-1%	11	8%
MyON†	Non-Program	9	24%	16	43%	10	42%	26	43%
	Program	13	30%	23	53%	13	50%	36	52%
	T-C Difference	4	6%	7	10%	3	8%	10	9%
Core5	Non-Program	154	23%	255	38%	271	52%	526	44%
	Program	152	23%	342	52%	354	68%	696	59%
	T-C Difference	-2	0%	87	14%	83	16%	170	15%
Imagine-	Non-Program	183	24%	279	37%	307	51%	586	43%

Kindergarten		Well Below to Below Bench		Well Below to Bench		Below to Bench		WB or Below to Bench	
Learning	Program	204	26%	379	47%	367	66%	746	55%
	T-C Difference	21	2%	100	10%	60	15%	160	12%
Istation‡	Non-Program	12	19%	19	30%	22	65%	41	42%
	Program	12	20%	20	33%	23	66%	43	45%
	T-C Difference	0	1%	1	3%	1	1%	2	3%

‡ Indicates ITT Group

Table 10. Treatment and Control Comparison of Change in DIBELS Next Benchmark Levels for At-risk First Grade Students, by Vendor

1st		Well Below to Below Bench		Well Below to Bench		Below to Bench		WB or Below to Bench	
Waterford	Non-Program	46	17%	72	27%	92	52%	164	37%
	Program	48	17%	84	31%	106	63%	190	43%
	T-C Difference	2	0%	12	4%	14	11%	26	6%
Success-Maker	Non-Program	13	20%	17	27%	44	57%	61	43%
	Program	16	23%	26	38%	43	68%	69	52%
	T-C Difference	3	3%	9	11%	-1	11%	8	9%
MyON	Non-Program	1	3%	10	29%	15	60%	25	42%
	Program	7	18%	14	37%	8	42%	22	39%
	T-C Difference	6	15%	4	8%	-7	-18%	-3	-3%
Core5	Non-Program	133	14%	255	28%	364	57%	619	40%
	Program	148	16%	302	33%	400	61%	702	44%
	T-C Difference	15	2%	47	5%	36	4%	83	4%
Imagine-Learning	Non-Program	178	16%	312	28%	399	57%	711	39%
	Program	231	20%	298	26%	403	56%	701	38%
	T-C Difference	53	4%	-14	-2%	4	-1%	-10	-1%
Istation	Non-Program	7	22%	5	16%	5	36%	10	22%
	Program	5	15%	6	18%	5	25%	11	21%
	T-C Difference	-2	-7%	1	2%	0	-11%	1	-1%

Table 11. Treatment and Control Comparison of Change in DIBELS Next Benchmark Levels for At-risk Second Grade Students, by Vendor

2nd		Well Below to Below Bench		Well Below to Bench		Below to Bench		WB or Below to Bench	
Waterford	Non-Program	7	14%	1	2%	8	38%	9	13%
	Program	15	29%	1	2%	13	65%	14	20%
	T-C Difference	8	15%	0	0%	5	27%	5	7%
Success-Maker	Non-Program	10	27%	6	16%	8	36%	14	24%
	Program	8	22%	3	8%	10	43%	13	22%
	T-C Difference	-2	-5%	-3	-8%	2	7%	-1	-2%
MyON‡	Non-Program	19	20%	9	10%	21	44%	30	21%
	Program	19	20%	9	9%	25	54%	34	24%
	T-C Difference	0	0%	0	-1%	4	10%	4	3%
Core5	Non-Program	166	17%	83	8%	174	39%	257	18%
	Program	185	19%	68	7%	202	44%	270	19%
	T-C Difference	19	2%	-15	-1%	28	5%	13	1%
Imagine-Learning	Non-Program	129	18%	68	9%	122	39%	190	18%
	Program	111	15%	95	13%	134	41%	229	22%
	T-C Difference	-18	-3%	27	4%	12	2%	39	4%
Istation‡	Non-Program	11	16%	2	3%	5	28%	7	8%
	Program	9	13%	2	3%	5	31%	7	8%
	T-C Difference	-2	-3%	0	0%	0	3%	0	0%

‡ **Indicates ITT Group**

Table 12. Treatment and Control Comparison of Change in DIBELS Next Benchmark Levels for At-risk Third Grade Students, by Vendor

3rd		Well Below to Below Bench		Well Below to Bench		Below to Bench		WB or Below to Bench	
Success-Maker	Non-Program	9	20%	9	20%	5	36%	14	24%
	Program	11	24%	5	11%	6	43%	11	19%
	T-C Difference	2	4%	-4	-9%	1	7%	-3	-5%
MyON‡	Non-Program	33	15%	41	18%	57	65%	98	32%
	Program	32	14%	44	20%	57	64%	101	32%
	T-C Difference	-1	-1%	3	2%	0	-1%	3	0%
Core5	Non-Program	159	15%	145	14%	196	52%	341	24%
	Program	174	17%	144	14%	220	57%	364	26%
	T-C Difference	15	2%	-1	0%	24	5%	23	2%
Imagine-Learning	Non-Program	74	15%	73	14%	80	47%	153	23%
	Program	83	16%	80	16%	103	61%	183	27%
	T-C Difference	9	1%	7	2%	23	14%	30	4%
Istation‡	Non-Program	21	25%	8	9%	14	42%	22	19%
	Program	17	19%	12	14%	7	23%	19	16%
	T-C Difference	-4	-6%	4	5%	-7	-19%	-3	-3%

‡ Indicates ITT Group

How did hours of use effect student outcomes?

The unstandardized regression coefficients depicted in **Tables 13-16** represent the relationship between hours of use and literacy outcomes: the coefficient represents a unit change in composite score for every additional hour of use. It should be noted that this comparison does not include control students, so even when a statistically significant relationship between hours of use and literacy scores were found, the control students could also have improved at the same rate (however, for obvious reasons, control students who did not use the program cannot be included). This analysis allowed us to see the relative effects each vendor had within their sample of program students.

We found a positive relationship between additional hours of program use and increased literacy scores for most program vendors in kindergarten and first grade, which, as noted in other analyses, further emphasizes the importance of program dosage for producing learning benefits. **Tables 13-14** reveal that for every additional hour of use, the end-of-year composite score increased by an average of .38 – 1.32 points in kindergarten for five of six programs, and .43 -.77 points in first grade for four of five programs. As we can see in **Tables 15-16**, one vendor had a significant and positive effect in second grade (.32 points), while two vendors had significant, positive

effects in third grade (.3-.5 points). In general, the same vendors produced statistically significant effects on the individual literacy domains.

Table 13. Effects of Hours of Program Use on Literacy Scores in Kindergarten

	N	EOY Composite	MOY FSF	EOY LNF	EOY PSF	EOY NWFCLS
Istation	161	NS	NS	NS	NS	NS
Waterford	1,879	.527***	.172***	.148***	.117***	.172***
Imagine Learning	6,080	.568***	.080***	.160***	.080***	.218***
SuccessMaker	249	.979**	.357***	.354**	NS	.363*
Core5	6,542	.378***	-.025*	.107***	.053***	.195***
MyON	151	1.32*	NS	NS	NS	1.05**

*p ≤ .05. ** p ≤ .01. *** p ≤ .001.
Note. Model covariates were sex, Hispanic, special education, school Title I status, ELL status, and the appropriate BOY or MOY Composite or subscale score. NS (not significant) in a cell means the program did not have a significant effect. Table depicts the Unstandardized Beta Coefficients.

Table 14. Effects of Hours of Program Use on Literacy Scores in 1st Grade

	N	EOY Composite	EOY NWFCLS	EOY NWF-WWR	EOY DORF
Istation	181	-.987*	-.382*	NS	NS
Waterford	2,131	NS	NS	.052***	.063*
Imagine Learning	8,568	.434***	.051*	NS	.111***
SuccessMaker	629	.767**	.238*	NS	NS
Core5	8,455	.463***	.195***	.041***	.106***
MyON	502	.610*	NS	NS	NS

*p ≤ .05. ** p ≤ .01. *** p ≤ .001.
Note. Model covariates were sex, Hispanic, special education, school Title I status, ELL status, and the appropriate BOY or MOY Composite or subscale score. NS (not significant) in a cell means the program did not have a significant effect. Table depicts the Unstandardized Beta Coefficients

Table 15. Effects of Hours of Program Use on Literacy Scores in 2nd Grade

	N	EOY Composite	EOY DORF
Istation	88	NS	NS
Waterford	213	NS	NS
Imagine Learning	2,138	NS	NS
SuccessMaker	132	NS	NS
Core5	2,409	.315***	.071**
MyON	142	NS	NS

*p ≤ .05. ** p ≤ .01. *** p ≤ .001.
Note. Model covariates were sex, Hispanic, special education, school Title I status, ELL status, and the appropriate BOY or MOY Composite or subscale score. NS (not significant) in a cell means the program did not have a significant effect. Table depicts the Unstandardized Beta Coefficients.

Table 16. Effects of Hours of Program Use on Literacy Scores in 3rd Grade

	N	EOY Composite	EOY DORF
Istation	119	NS	NS
Imagine Learning	1,777	NS	NS
SuccessMaker	170	NS	NS
Core5	2,539	.299**	.044*
Reading Plus	86	NS	NS
MyON	328	.507*	NS

*p ≤ .05. ** p ≤ .01. *** p ≤ .001.
Note. Model covariates were sex, Hispanic, special education, school Title I status, ELL status, and the appropriate BOY or MOY Composite or subscale score. NS (not significant) in a cell means the program did not have a significant effect. Table depicts the Unstandardized Beta Coefficients.

Summary, 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 summarize the key findings for both goals in this section, discuss the limitations involved in interpreting our findings, and, finally, present a brief set of recommendations to help improve the program.

Program Implementation

Program implementation results show that schools are doing a better job implementing the EISP than in previous years. More students met the recommendations for total weeks of use compared to those who met the average weekly use recommendations for all but one vendor. Across programs, 42 percent of students met the average use and 67 percent met the total weeks recommendations. When we relax the dosage requirements by 80 percent to allow schools some flexibility in program implementation, 60 percent of students met the average use and 77% met the weeks of use recommendations. Even with these strong results, there is room for improvement. MyOn, ReadingPlus and Istation each had lower percentages of students who met their use recommendations compared to the other vendors. These are also the three vendors with the fewest students using their programs, making it even more important for these vendors to increase their students' usage if we are to have a sufficient sample to measure outcomes based on vendors minimum recommended use. We recommend that vendors and school staff continue to work toward improving their usage through active involvement in the implementation process.

Please see our companion report, "**Best Practices for Improving Early Intervention Software Programs in Utah Schools**," to review detailed recommendations for program improvement.

Program Impacts on Literacy Achievement

We studied how the program impacted student literacy achievement in the aggregate across all seven programs (program-wide) and for each program vendor (vendor-specific).

Program-Wide. Our program-wide analyses underscore the importance of appropriate program use, and the overall program effects were dependent on how much a student used the program (which we call "dosage"). As dosage increased, program-wide literacy achievement increased. Students in the high dosage (more use) groups had literacy achievement outcomes two to three times larger than the students in the lower dosage groups (depending on the grade). We found statistically significant effects in favor of the

treatment group across all grade levels for the highest dosage group and the treatment effect sizes revealed that the differences in outcomes between program and non-program students were stronger than those found in similar intervention programs, except for in first grade. The strongest effect was found in kindergarten (ES=.2), which in practice is greater than the average effect size (.13) of similar interventions. When we analyzed the relationship between individual differences and outcomes, certain groups benefitted more than others. For example, students who were low-income, female, classified as special education, from Title I schools or English Language Learners (ELL) all scored higher on the literacy composite measure then their counterparts, suggesting that the program had a more pronounced positive effect on these populations.

Vendor-Specific. Most vendors had a positive effect on literacy outcomes in kindergarten, but only a few vendors had positive impacts in grades one through three. For example, four of six vendors produced statistically stronger literacy scores in kindergarten when compared to a matched control group, versus two vendors in first grade, two vendors in second, and two vendors in third grade. The magnitude of the effects was also generally stronger for these analyses: eight vendors produced effect sizes greater than .13 (the average effect size of similar programs) across all grade levels.

Evaluation Limitations

This evaluation is based on a complex amalgamation of secondary data sets, provided by multiple stakeholders (the state, DIBELS Next vendors, and program vendors), and there are limitations to our findings based on the type of research design, the data used and the ability to have adequate power to detect small effect sizes in our samples. Because of these limitations, the reader must exercise caution when interpreting the findings.

Quasi-Experimental Research Designs. We utilized a quasi-experimental research design (QED), a common design for studies in which naturally occurring groups of program and non-program students exist. Because these groups were not assigned at random, they could have had pre-existing differences that affect literacy outcomes, such as parental education, extracurricular supports and resources, among other things that influence learning achievement. To combat the potential effects of pre-existing group differences, we used a statistical matching process called Coarsened Exact Matching (CEM) to attempt to balance treatment and control groups on several covariates (important predictors of literacy outcomes). The matching process created balanced groups at the beginning of the year, however, there may have been factors that we could not measure that affected student learning during the year. Our results must be seen as probable outcomes, but there may be other variables influencing them.

Secondary Data. “Secondary data” were data collected by outside sources and transferred to the evaluators. The secondary data in this study were from program vendors, the state, and DIBELS Next databases, and some limitations arose from its use. First, a majority of our DIBELS Next data were collected and stored through the

DMG and AMPLIFY systems. These systems offer efficient transfer of DIBELS Next test scores, but they are limited and not all LEA's use them, and therefore we only had scores for a subgroup of program students. Other factors that affected the sizes of our samples included: students who used more than one software program, duplicate IDs, incomplete DIBELS scores, and other missing or incorrect data (such as student IDs) among other factors.

Statistical Power to Determine Program Effects. Statistical "power," or the probability that a statistical test will reject a false null hypothesis, is an important consideration when conducting analysis. In general, the smaller a sample size, the less likely one can find a statistically significant effect. In certain analyses, for specific vendors, this was a limiting factor in our evaluation. In addition (and related to small sample size limitations), due to a combination of low enrollment and low overall percentages of students who used the program as intended, for some vendors we could not isolate students based on a threshold for minimum usage (Istation, MyOn and ReadingPlus). This is a limitation to their findings because we know that the program's positive impacts on students are more pronounced when students use the software as recommended, and, had these vendors had either higher enrollment numbers or greater percentages of students who used the software as intended, we may have shown better results for select vendors.

Overall Conclusions & Recommendations

The 2016-2017 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, outcome measure and analyses method. When reviewing our current evaluation results with those from previous years, it is easy to recommend that the program be continued for kindergarten students. It is more difficult to fully endorse the program's use with students in 1st through 3rd grade as samples sizes, samples used for analysis and other factors varied across the evaluation. However, with select vendors, there were indicators that students in these upper-early grades benefitted from the program, so we are recommending that more data be collected and results reviewed for future cohorts. Additionally, schools are doing a better job implementing the program according to vendors' recommendations. This last issue is particularly important, because our results this year show a direct and positive relationship between higher levels of program use leading to stronger student outcomes. We believe that if schools could continue improving at program implementation, student benefits would also improve. To help the state and schools improve program implementation, we have created a companion report, "Best Practices for Improving Early Intervention Software Programs in Utah Schools", with a list of best practices for successful use of the program. The best practices findings were based on a separate study of schools, and serves as a companion to the empirical findings presented in this report.

References

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

Evaluation and Training Institute. (2016, September). Early Intervention Software Program Evaluation: 2015-2016 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 DIBELS 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). *DIBELS Next Assessment Manual.* Dynamic Measurement Group Inc. http://wenatchee.innersync.com/assessment/documents/dibelsnext_assessment manual.pdf

StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP

Appendix A: Methods and Analyses Samples

In this Appendix, we present a detailed description of the methods used in our analyses within the main report. We include information on our matching process for creating the comparison groups and present descriptive information on each of the samples in our analyses.

Implementation findings – we used descriptive statistics to show how program participants used the programs, using a sample that included K-1 and 2-3 intervention students. 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 only, and did not remove students with inaccurate SSIDs, students who used multiple software providers, or students with incomplete DIBELS data. The EISP targets intervention students in second and third grade, and of necessity, we needed valid SSIDs in the vendor and DIBELS data as well as beginning-of-year DIBELS scores to identify the intervention students in our sample.

We studied the program impacts across all programs (program-wide) and for each software program provider (individual program impacts). Students needed to have accurate state student Ids (SSIDs) and complete DIBELS data (outcome data) to be a viable case for our sample. We scrubbed the data to exclude students who may have used multiple software programs.

Program-wide analyses – as our largest sample, we used a two-level regression model ("hierarchical linear regression model", or HLM) to compare treatment students to control students on DIBELS Next composite scores and literacy subscales. For the program-wide analyses, we created three separate matched treatment and control groups based on levels of program dosage. In order of lowest to highest program dosage, our final program-wide samples included: 70,782 treatment students ("intent to treat"); 45,084 treatment students (relaxed optimal use: ROPT) and 19,458 treatment students (optimal use).

Individual program impacts – our sample size varied by vendor, and we used an Ordinary Least Squares (OLS) regression model to compare treatment students to control students on DIBELS Next composite scores. We created a new matched treatment and control group sample for each program vendor and usage group for which we had a sufficient sample size.

Coarsened Exact Matching (CEM) Method - we used Coarsened Exact Matching (CEM) to statistically match each treatment child with a control child who is most similar to them. If no matches could be made, children were removed from the sample. Using CEM, we are able to construct a comparison group of control children who resemble the treatment sample as closely as possible on specific observable characteristics, such as grade, gender, race/ethnicity, and performance on pre-test measures. In the following tables, we present the characteristics of the treatment group for each matched sample

used in our analyses. As a result of our CEM procedure, the matched controls have the same descriptive information as those in the below tables.

Students in the ITT sample had the lowest program use, students in the ROPT sample had the second highest use, and the Optimal Use sample had the highest use. **Tables A1 – A4** present the characteristics of the treatment group for each matched sample used in our analyses. As a result of our CEM procedure, our matched controls were the same.

Program-wide Sample

Table A1. Program-Wide Sample by Grade, Intent to treat

	N	Female	Caucasian	Hispanic	Multiple races	Asian	Pacific Islander	African American	American Indian	SPED	Low-income	ELL	BOY Comp
K	15062	N=7225 48%	11242 75%	2634 17%	392 3%	215 1%	257 2%	206 1%	116 1%	1275 8%	5335 35%	1591 11%	35
1st	20466	9939 49%	15478 76%	3360 16%	606 3%	304 1%	321 2%	282 1%	115 1%	2080 10%	7633 37%	2046 10%	127
2nd	5122	2512 49%	3236 63%	1448 28%	113 2%	68 1%	87 2%	101 2%	69 1%	1169 23%	2877 56%	1140 22%	72
3rd	5026	2386 47%	3014 60%	1549 31%	112 2%	76 2%	96 2%	124 2%	55 1%	1318 26%	2839 56%	1248 25%	122

Table A2. Program-Wide Sample by Grade, Relaxed Optimal Use (ROPT)

	N	female	Caucasian	Hispanic	Multiple races	Asian	Pacific Islander	African American	American Indian	SPED	Low-income	ELL	BOY Comp
K	7869	N=3800 48%	5966 76%	1283 16%	218 3%	115 1%	122 2%	102 1%	63 1%	665 8%	2786 35%	737 9%	35
1st	12331	5986 49%	9580 78%	1815 15%	352 3%	178 1%	187 2%	152 1%	67 1%	1196 10%	4438 36%	1094 9%	131
2nd	2780	1355 49%	1761 63%	782 28%	57 2%	34 1%	52 2%	56 2%	38 1%	589 21%	1549 56%	610 22%	73
3rd	2477	1179 48%	1445 58%	807 33%	58 2%	41 2%	51 2%	55 2%	20 1%	592 24%	1410 57%	672 27%	123

Table A3. Program-Wide Sample by Grade, Optimal Use

	N	female	Caucasian	Hispanic	Multiple races	Asian	Pacific Islander	African American	American Indian	SPED	Low Income	ELL	Boy Comp
K	3563	N=1663 47%	2719 76%	583 6%	94 3%	46 1%	50 1%	39 1%	32 1%	256 7%	1266 36%	331 9%	37
1st	4619	2216 48%	3880 84%	518 11%	110 2%	29 1%	39 1%	25 1%	18 0%	366 8%	1491 32%	285 6%	135
2nd	886	435 49%	603 68%	240 27%	9 1%	4 0%	10 1%	10 1%	10 1%	163 18%	482 54%	191 22%	80
3rd	661	324 49%	406 61%	219 33%	13 2%	3 0%	9 1%	7 1%	4 1%	126 19%	374 57%	166 25%	133

Vendor-specific Sample

Table A4. Vendor-specific Sample by Grade, Relaxed Optimal Use and ITT Samples

	Grade	N	Female	Caucasian	Hispanic	Multiple	Asian	Pacific Islander	Black	American Indian	SPED	Low-income	ELL	BOY Comp
Waterford	K	1242	599 48%	995 80%	184 15%	14 1%	8 1%	11 1%	13 1%	17 1%	101 8%	430 35%	29 2%	36
	1	1183	584 49%	1005 85%	124 10%	27 2%	4 0%	6 1%	8 1%	9 1%	125 11%	482 41%	32 3%	126
	2	71	31 44%	48 68%	14 20%	1 1%	1 1%	1 1%	1 1%	5 7%	24 34%	46 65%	9 13%	69
	Total	2496	1214 49%	2048 82%	322 13%	42 2%	13 1%	18 1%	22 1%	31 1%	250 10%	958 38%	70 3%	79
Imagine Learning	K	3081	1470 48%	2409 78%	476 15%	83 3%	32 1%	39 1%	21 1%	21 1%	298 10%	1128 37%	306 10%	32
	1	4552	2182 48%	3526 77%	700 15%	125 3%	54 1%	76 2%	43 1%	28 1%	476 10%	1714 38%	443 10%	123
	2	1049	522 50%	717 68%	273 26%	16 2%	3 0%	18 2%	10 1%	12 1%	232 22%	547 52%	204 19%	73
	3	680	294 43%	457 67%	178 26%	18 3%	4 1%	12 2%	10 1%	1 0%	197 29%	329 48%	154 23%	118
	Total	9362	4468 48%	7109 76%	1627 17%	242 3%	93 1%	145 2%	84 1%	62 1%	1203 13%	3718 40%	1107 12%	87
Core5	K	3305	1616 49%	2384 72%	597 18%	103 3%	70 2%	68 2%	59 2%	24 1%	229 7%	1141 35%	397 12%	39
	1	5113	2501 49%	3945 77%	760 15%	153 3%	89 2%	81 2%	61 1%	24 0%	415 8%	1660 32%	498 10%	136
	2	1429	680 48%	866 61%	448 31%	30 2%	14 1%	26 2%	32 2%	13 1%	271 19%	836 59%	364 25%	73
	3	1422	706 50%	739 52%	560 39%	20 1%	15 1%	39 3%	33 2%	16 1%	304 21%	906 64%	439 31%	122
	Total	11269	5503 49%	7934 70%	2365 21%	306 3%	188 2%	214 2%	185 2%	77 1%	1219 11%	4543 40%	1698 15%	98

	Grade	N	Female	Caucasian	Hispanic	Multiple	Asian	Pacific Islander	Black	American Indian	SPED	Low-income	ELL	BOY Comp
Success-Maker	K	126	53 42%	116 92%	5 4%	3 2%	0 0%	1 1%	1 1%	0 0%	17 13%	45 36%	0 0%	34
	1	392	189 48%	326 83%	37 9%	13 3%	3 1%	6 2%	5 1%	2 1%	28 7%	88 22%	5 1%	133
	2	59	36 61%	48 81%	9 15%	0 0%	1 2%	1 2%	0 0%	0 0%	8 14%	18 31%	3 5%	88
	3	59	31 53%	47 80%	11 19%	1 2%	0 0%	0 0%	0 0%	0 0%	12 20%	21 36%	8 14%	133
	Total	636	309 49%	537 84%	62 10%	17 3%	4 1%	8 1%	6 1%	2 0%	65 10%	172 27%	16 3%	109
Istation‡	K	161	69 43%	106 66%	54 34%	1 1%	0 0%	0 0%	0 0%	N/A	19 12%	99 61%	26 16%	25
	1	179	79 44%	118 66%	60 34%	1 1%	0 0%	0 0%	0 0%	N/A	26 15%	109 61%	31 17%	113
	2	88	40 45%	47 53%	38 43%	1 1%	1 1%	1 1%	0 0%	N/A	20 23%	66 75%	24 27%	62
	3	118	57 48%	79 67%	33 28%	4 3%	0 0%	1 1%	1 1%	N/A	33 28%	78 66%	17 14%	123
	Total	546	245 45%	350 64%	185 34%	7 1%	1 0%	2 0%	1 0%	N/A	98 18%	352 64%	98 18%	81
MyOn‡	K	149	75 50%	97 65%	43 29%	7 5%	2 1%	0 0%	0 0%	0 0%	7 5%	57 38%	32 21%	34
	1	498	243 49%	373 75%	87 17%	19 4%	12 2%	3 1%	3 1%	1 0%	46 9%	166 33%	40 8%	126
	2	142	74 52%	120 85%	17 12%	3 2%	0 0%	0 0%	1 1%	1 1%	31 22%	45 32%	3 2%	74
	3	311	159 51%	233 75%	70 23%	4 1%	1 0%	1 0%	1 0%	1 0%	73 23%	140 45%	55 18%	126
	Total	1100	551 50%	823 75%	217 20%	33 3%	15 1%	4 0%	5 0%	3 0%	157 14%	408 37%	130 12%	107

‡ **Indicates ITT Group**

Note: Model covariates were gender, Hispanic, special education, school Title I status, and BOY Composite score. The Relaxed Optimal (ROPT) use sample was used for the following vendors: Waterford, Imagine Learning, Lexia, and Success-Maker. The Intent to Treat sample was used for Istation and MyOn.

Appendix B. Program Use Descriptives

Table B1 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 B1. Program Use by Vendor and Grade

	Grade	N	Ave Wkly Min of Use	Ave Total Minutes	Ave Wks. of Use	% Met Wks. Recs	% Met Ave Use Recs	Wks. Met 80% Ave Use Recs	Met 80% Ave Min. Recs		Met 80% Ave Min. & 80% Wks. Recs		Met 80% Min. & Wks. Recs	
Istation	K	176	32	595	17	19%	2%	3	6	3%	71	40%	4	2%
	1	199	62	1665	27	53%	58%	17	172	86%	150	75%	134	67%
	2	95	43	928	19	17%	26%	8	35	37%	38	40%	21	22%
	3	143	48	956	19	24%	21%	9	74	52%	56	39%	41	29%
	Total	613	47	1078	21	31%	28%	10	287	47%	315	51%	200	33%
Waterford	K	3159	57	1673	28	64%	42%	18	2144	68%	2625	83%	2007	64%
	1	2998	71	2098	29	69%	30%	17	1883	63%	2601	87%	1754	59%
	2	258	56	1438	24	40%	15%	11	94	36%	173	67%	83	32%
	Total	6415	64	1862	28	65%	35%	17	4121	64%	5399	84%	3844	60%
Imagine Learning	K	7529	43	1000	22	67%	41%	13	4409	59%	5935	79%	3848	51%
	1	9478	52	1394	25	79%	32%	14	5204	55%	8104	86%	4855	51%
	2	2382	51	1252	23	68%	34%	12	1223	51%	1899	80%	1131	47%
	3	1951	46	1030	20	61%	23%	9	824	42%	1369	70%	719	37%
	Total	21340	48	1206	23	72%	35%	13	11660	55%	17307	81%	10553	49%

	Grade	N	Ave Wkly Min of Use	Ave Total Minutes	Ave Wks. of Use	% Met Wks. Recs	% Met Ave Use Recs	Wks. Met 80% Ave Use Recs	Met 80% Ave Min. Recs		Met 80% Ave Min. & 80% Wks. Recs		Met 80% Min. & Wks. Recs	
SuccessMaker	K	283	45	915	19	81%	45%	13	164	58%	241	85%	153	54%
	1	768	48	1018	19	63%	54%	14	566	74%	521	68%	493	64%
	2	234	47	936	18	72%	29%	9	117	50%	173	74%	108	46%
	3	203	43	812	17	62%	18%	8	68	33%	131	65%	63	31%
	Total	1488	47	957	19	68%	43%	12	915	61%	1066	72%	817	55%
Core5	K	7563	48	1036	20	56%	48%	12	4443	59%	4916	65%	3605	48%
	1	10173	55	1466	25	76%	59%	16	7155	70%	8466	83%	6590	65%
	2	2958	57	1517	25	78%	49%	15	1985	67%	2530	86%	1842	62%
	3	3138	53	1325	23	69%	42%	13	1944	62%	2407	77%	1756	56%
	Total	23832	53	1317	23	69%	52%	14	15527	65%	18319	77%	13793	58%
Reading-Plus	3	184	36	519	13	41%	28%	6	82	45%	96	52%	58	32%
MyOn	K	567	41	469	7	12%	31%	4	225	40%	87	15%	79	14%
	1	1253	44	648	13	29%	35%	6	648	52%	496	40%	318	25%
	2	416	45	786	15	36%	37%	8	216	52%	190	46%	140	34%
	3	565	51	1047	18	48%	48%	11	374	66%	341	60%	278	49%
	Total	2801	45	713	13	30%	37%	7	1463	52%	1114	40%	815	29%

Appendix C. Data Processing & Merge Summary

We reviewed and cleaned data from ten different sources in preparation of completing our analyses, including program usage data from seven software program providers, student literacy achievement data from two DIBELS Next systems (DMG and AMPLIFY), 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 87,857 to 86,722 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.

To create the vendor data used in our outcome analyses, we identified and removed duplicate IDs across vendors[7] (N=5,471) and any IDs that did not comply with the state student ID (SSID) format (N=593). 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. Either way, we did not include these students in order to report the individual impacts for each software provider. For similar reasons, we excluded students who used Imagine

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

Learning through a separate state-wide grant[8] prior to reporting the program impacts. See **Table C1**. in the section, "Impact of Data Cleaning on Vendor Samples", for additional details on how vendors' samples were impacted throughout the data cleaning and merge process.

SIS Data

We were provided SIS data for all students enrolled in EISP LEAs. We reviewed the SIS data provided by the USBE to ensure that all LEAs who were listed as 2016-2017 participants were included in the data. The raw data file consisted of 205,793 cases, of which almost five percent were duplicate records. After cleaning the data of duplicates, our SIS data consisted of 196,085 records.

DIBELS Next Data

Similar to our process of requesting data from vendors, we provided an example file to help ensure consistency between both systems. We started with two separate DIBELS data files, one which was pulled from the DMG database by USBE staff (n=40,651) and the other prepared by AMPLIFY (n=157,650). USBE staff worked to correct invalid SSIDs in the DMG data prior to submitting it to ETI. We cleaned the duplicate IDs within the individual DIBELS data files, deleting 256 duplicate cases[9] from DMG and 98 from AMPLIFY. This left us with 40,387 in the DMG data and 157,456 cases from the AMPLIFY data. We combined the DIBELS data files (n=197,842) and cleaned duplicates cases in the new file (n=3,432). 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 DIBELS file containing 177,368 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 DIBELS file (177,368) and were left with 168,322 cases[10]. Next, we merged our master vendor data into the DIBELS and SIS data. Our final merged file consisted of 53,535 treatment students and 114,787 potential comparison students. After processing the data for missing BOY and EOY test scores, among other steps, our final, pre-matched dataset consisted of 90,268 cases.

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

[8] 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.

[9] Deleted duplicate IDs were either incorrect IDs (e.g. same ID but different student name) or had two different outcome scores attached to each ID).

[10] We were provided SIS data for EISP districts only, and this number should not be used to determine the SSID accuracy rate of the DIBELS data.

pool of non-program participants from within the same districts as program participants, 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 ROPT and OPTI samples). However, for our largest sample of program students, the ITT program-wide sample, there were more program students than control students. We had 45,639 treatment students and 44,629 potential control students. This automatically reduced the size of this particular sample. See Appendix A for tables of our final samples.

Impact of Data Cleaning on Vendor Samples

The table below depicts the stages of the cleaning process in terms of how it affected each vendors data. The N's in the first column were reported after the initial cleaning process was complete. We can see from the below table that the samples for MyOn, ReadingPlus and SuccessMaker lost quite a few cases due to students using multiple vendors (e.g. between vendor duplicates), which indicates schools may be using the programs outside of the expectations (e.g. students are not to use more than one program). Additionally, all vendors' samples were affected by cleaning the data to exclude non-intervention students in Grades 2-3, with MyOn and ReadingPlus affected the most (e.g. lost 44% and 46%, respectively). Finally, ReadingPlus lost most of their data due to missingness in the DIBELS data. ReadingPlus is used by one of the only districts to measure student outcomes using an instrument other than the DIBELS.

Table C1. Overview of Data Cleaning Process by Program

	N	Without between vendor duplicates		Without Missing IDs		Without 2nd-3rd Grade non-intervention students		Without Missing DIBELS or SIS	
Istation	867	843	3%	843	0%	591	30%	549	7%
Waterford	7,044	6,756	4%	6,685	1%	6,159	8%	4,256	31%
Imagine Learning	28,593	27,568	4%	27,566	0%	20,878	24%	18,519	11%
Success-Maker	2,042	1,663	19%	1,663	0%	1270	24%	1,180	7%
Core5	38,156	36,514	4%	36,174	1%	22,206	39%	19,928	10%
Reading-Plus	1,762	1,246	29%	1,226	2%	660	46%	86	87%
MyOn	4,926	3,329	32%	3,169	5%	1,765	44%	1,121	36%
Total	83,390	77,919	7%	77,326	1%	53,529	31%	45,639	15%

**Note.* First column "N's" represent count of students after cleaning individual vendors' data and excluding students who used Imagine Learning as part of a separate state initiative and i-Ready students.

Appendix D: DIBELS Next Measures

The Dynamic Indicators of Basic Early Literacy skills (DIBELS Next) is a statewide assessment used to measure students acquisition of early literacy skills at the beginning, middle, and end of the academic year. The online data entry systems, AMPLIFY and DIBELS Measurement Group (DMG)[11], were used by a majority of LEAs throughout the state to capture DIBELS Next data.

According to a technical report produced by the Dynamic Measurement Group (Powell-Smith, et al., 2014), *"The DIBELS 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 DIBELS subscales used in our analyses.

Table D1. DIBELS Next Scales

DIBELS Next Scale	Description	Early Literacy Construct	Grade
Composite Score	DIBELS Composite Score is a combination of multiple DIBELS 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
DIBELS Oral Reading Fluency (DORF)	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

[11] 2016-2017 was the first year in which DMG was used by districts in the state to house the DIBELS Next data.

DIBELS Next Scale	Description	Early Literacy Construct	Grade
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

**DIBELS NEXT Manual: http://wenatchee.innersync.com/assessment/documents/dibelsnext_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 DIBELS Next 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 to similar curriculum or broad instructional programs, defined by Lipsey et al. (2012) as, "*a relatively complete and comprehensive package for instruction in a content area like a curriculum or a more or less free-standing program (e.g., science or math curriculum; reading programs for younger students; broad name brand programs like Reading Recovery; organized multisession tutoring program in a general subject area*" (pg. 35). The average effect size was .13. for these types of instructional programs.