

# Utah's Early Intervention Reading Software Program

## K-3 Program Evaluation Findings

Submitted to the Utah State Board of Education  
September 2018



Evaluation and Training Institute  
100 Corporate Pointe, Suite 387  
Culver City, CA 90230  
[www.eticonsulting.org](http://www.eticonsulting.org)

## Table of Contents

Executive Summary.....	1
------------------------	---

### What are the goals of this report?

● Evaluation Purpose & Evaluation Questions .....	4
---	---

### What do we know about why the program was created and who participated?

● Program Background and Enrollment .....	5
---	---

● Evaluation Methods.....	8
---------------------------	---

### What do we know about how students used the software programs this year?

● Program Implementation.....	13
-------------------------------	----

### What impact did the program have on students' literacy achievement this year?

● Program-Wide Impacts: The big picture .....	15
● Impact of Each Software Program .....	21

### What trends do we see in program enrollment, use and impact across the years?

● Multi-year Findings .....	25
-----------------------------	----

● Discussion, Limitations and Recommendations.....	28
--	----

References .....	31
------------------	----

Appendix A: Analyses Samples .....	32
------------------------------------	----

Appendix B. Data Processing & Merge Summary .....	36
---	----

Appendix C: DIBELS Next Measures .....	39
--	----

Appendix D: Determining Effect Size Benchmark .....	40
---	----

Appendix E. Program Use by Vendor and Grade .....	41
---	----

Appendix F: Multi-Year Dosage Recommendations.....	43
--	----

## List of Figures

Figure 1. DIBELS Indicator & Literacy Skill Measures .....	10
Figure 2: Students who met vendors minimum dosage recommendations .....	14
Figure 3: Students who met the dosage recommendations by grade.....	14
Figure 4. DIBELS Literacy Domain Effect Sizes by Grade, Highest Dosage Sample .....	18
Figure 5. % Change in Benchmark Status from BOY to EOY, Kindergarten .....	19
Figure 6. % Change in Benchmark Status from BOY to EOY, 1 <sup>st</sup> Grade.....	19
Figure 7. % Change in Benchmark Status from BOY to EOY, 2nd Grade.....	20
Figure 8. % Change in Benchmark Status from BOY to EOY, 3rd Grade .....	21
Figure 9. Impact of Individual Vendors on DIBELS Composite Scores, Effect Sizes by Grade ..	24
Figure 10. Grade Levels with Significant Effect Sizes by Vendor and Program Year .....	27

## List of Tables

Table 1. 2017-2018 Program Enrollment Overview .....	5
Table 2. 2017-2018 Program Enrollment by Vendor and Grade .....	6
Table 3. Vendor 2017-2018 Minimum Dosage Recommendations .....	7
Table 4. Predicted Means of DIBELS Composite Scores for Matched Treatment and Control, Program-wide, Highest Dosage Sample.....	16
Table 5. Predicted Means of EOY DIBELS Literacy Domains for Matched Treatment and Control, Highest Dosage Sample.....	17
Table 6. Mean Score Differences on EOY DIBELS Composite Scores by Grade and .....	21
Table 7. Predicted Means of EOY DIBELS Composite for Matched Treatment and Control, by Vendor, OLS Regression Model .....	23
Table 8. Est. Program Enrollment from 2013/2014 – 2017/2018 .....	25
Table 9. Multi-year Trends in Program Use.....	26
Table 10. Trends in program-wide impacts, effect sizes by dosage sample .....	26

# Acknowledgements

The Evaluation and Training Institute (ETI) thanks Sarah Young (Coordinator, Digital Teaching and Learning) from the Utah State Board of Education (USBE) for her ongoing collaboration and direction throughout this evaluation project.

We also acknowledge Kristen Campbell and Malia McIlvenna at USBE for helping us understand the ins and outs of the data and appreciate 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 Robert Rubin from Istation, Nari Carter from Imagine Learning, Haya Shamir from Waterford, Sean Mulder from SuccessMaker, Sarah Franzén from Lexia, Dave McMullen from MyOn and Sarah Hoerr and Kevin Sheridan from ReadingPlus. Each of these individuals provided necessary data from their products that were used to complete the evaluation project.

# 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
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) provides Utah's Local Education Agencies (LEAs) with the opportunity to select from among seven adaptive computer-based literacy software programs. The program's goal is to increase the literacy skills of all students in K-1 and struggling readers in Grades 2-3. As the EISP external evaluator, the Evaluation and Training Institute (ETI) studied three aspects of the EISP: 1) students use of the program during the school year ("program implementation"); 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); and 3) trends in program implementation and impacts across multiple years of program implementation.

## Program Implementation Findings

Program vendors provided recommendations on program dosage for students to achieve the benefits in literacy skill development from their participation in the software programs. The implementation study was designed to determine the extent to which students met each vendors' recommendations for average weekly use and total weeks of use. A majority of students (72-83%) using five of the seven software programs met the requirements for total weeks of use, which ranged from 15-28 weeks, and is 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 seven vendors, there were three in which more than half of the students met the recommendations for weekly minutes of use, on average.

## Program-wide Impacts Findings

The program had a positive impact on students' literacy skill development in kindergarten and first grade, regardless of their program dosage, and in 3<sup>rd</sup> grade for students with the highest program dosage. There were no statistically significant positive effects for students in second grade. In general, the effectiveness of the program increased in strength as dosage increased from the lowest to highest dosage. The program was most effective for students in kindergarten who had the highest program dosage ( $ES=.16$ ), which is also higher than the average effect size seen in similar intervention programs. In addition, K-1 students with the highest program dosage ended the year above benchmark (mean composite scores of 157 and 209 respectively) for their grade. Students who scored above benchmark had a 90-99% likelihood of achieving subsequent early literacy goals (Dynamic Measurement Group, 2016).

## Vendor Impacts Findings

In addition to examining program-wide impacts, we studied the impacts of individual program vendors on students' literacy achievement. In kindergarten students who met a minimum program dosage threshold had higher literacy achievement, as measured by their mean literacy composite scores, compared to a group of non-program students with similar characteristics. To measure the strength of these effects, we looked at the average effect sizes produced by similar education intervention programs. In kindergarten, the effects were stronger than those found in similar intervention programs for all five software vendors included in this analyses<sup>1</sup>. In addition, two vendors had positive impacts on students in first grade and one vendor had a positive impact in second grade; however, these effects were smaller than the average effect size benchmark.

## Multi-Year Evaluation Findings


We studied the trends in program enrollment, students' program use, and the program impacts on student achievement over the past few years of program implementation. Program enrollment and program use increased exponentially each year, indicating that LEAs are making incremental improvements in students' usage as the program continues. The trends in program impacts were more complex and varied each year depending on the vendor and students' grade level. We consistently saw strong impacts for students in kindergarten for multiple vendors, but not in Grades 1-3. In addition, when comparing the strength of the program impacts across years using effect sizes, we found that the strength of the effect sizes were decreasing each year. However, we caution readers from drawing the conclusion that the program is less effective now than it was in the beginning of its implementation. For example, we know that schools in Utah are increasing their use of computer-based intervention programs and it's possible that more of our control students are using programs similar to those being measured. In addition, through our 2016-2017 qualitative study of program implementation, we now understand that students need to be monitored by teachers to ensure that they are progressing through the curriculum appropriately and that time in the program may not tell the complete story.

## Discussion & Recommendations

The 2017-2018 program had a positive effect in kindergarten and first grade (looking at the program as a whole), and had mixed effects on students in first through second grade depending on the specific vendor. 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 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, and we recommend combining students across multiple

---

<sup>1</sup> ReadingPlus was used in only in the upper-early grade levels and MyOn did not have enough kindergarten students to be included in this analysis.



program years as one approach for increasing the sample sizes of specific vendors. Combining cohorts of students would allow us to measure the program impacts for all vendors and grades and who met the same program dosage criteria. We also 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, studying the connection between students' progression through the program content and time spent on the software would help us determine if students are learning during their time in the software. This information could also be used to study the relationship between the amount of program content covered and students' literacy achievement.



# 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' individualized learning needs. The programs were supplied by multiple vendors and were implemented in schools, grades K-3. The Evaluation and Training Institute (ETI) conducted its annual evaluation of the EISP, which focused on how the reading software programs were used and the impact they had on students' literacy achievement. The evaluation included results for the combined impact of all the software programs taken together ("program-wide" impacts) and a comparison of the relative effects on literacy achievement for each of the software providers ("individual vendor impacts").

This report includes findings from the 2017-2018 academic year, the EISP's fifth year of implementation, as well as an overview of cumulative program findings from previous program years. These findings are intended to help the Utah State Board of Education (USBE) and Local Education Agencies (LEAs) understand how the program is 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. Did the program effects differ based on student or school characteristics?
4. Were there differences in treatment effects among vendors?
5. What are the trends in implementation and literacy achievement across the years?

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.

In this report we include a description of the EISP and 2017-2018 student enrollment, a summary of our research methods, findings related to each research question and the two study objectives (program implementation and program impacts), and trends in findings across the program years. Finally, we discuss the key findings and the study limitations.

# Program Background & Enrollment

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 EISP, LEAs (districts and charter schools) submitted applications to the USBE requesting funding for the use of specific reading software programs prior to the start of each school year.

Seven software vendors provided software and training to schools through the EISP in 2017-2018. The seven vendors were (in alphabetical order): Imagine Learning, Istation, Pearson ("SuccessMaker"), Lexia® Core5® Reading (Core5), MyOn, Reading Plus and Waterford. These software programs were used in 79 LEAs and 403 schools and by approximately 100,951 students. Core5 was the most frequently used program (188 schools, 50,000+ students), while Istation was used the least (7 schools; 1,238 students).

**Tables 1-2** present the 2017-2018 enrollment of LEAs and students who used each vendor. While the EISP was intended for second and third grade students reading below grade level (referred to as "intervention" throughout the report), some educators implemented the program with their entire class, and in these instances, students reading at grade level ("non-intervention") also had access to the software programs. Our report focused on intervention students in Grades 2-3, however, we have provided enrollment information for both types of students so readers may understand how the program was implemented in practice and as intended.

Table 1. 2017-2018 Program Enrollment Overview

Program	LEAs	Schools	Students	
			All K-3	All K-1 & 2-3 Intervention
Istation	5	7	1,238	926
Waterford	23	52	6,398	5,712
Imagine Learning	45	168	33,035	23,997
SuccessMaker	8	19	2,015	1,220
Core5	39	188	52,807	32,136
Reading Plus	2	14	1,246	174
MyOn	8	33	4,211	1,512

*Note.* Count of LEAs/schools are not unique due to instances where multiple programs were used within a LEA/school. Data source: pre-merged data in K-1 and vendor data merged to DIBELS in Grades 2-3.

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 the other grades (**Table 2**).

Table 2. 2017-2018 Program Enrollment by Vendor and Grade

Program	Kinder	1st	2 <sup>nd</sup>		3 <sup>rd</sup>	
			All	Intervention	All	Intervention
Istation	350	356	334	125	198	95
Waterford	2,731	2,588	868	283	211	110
Imagine Learning	8,357	11,013	7,880	2,446	5,785	2,181
SuccessMaker	192	586	581	185	656	257
Core5	11,337	13,441	14,341	3,518	13,688	3,840
ReadingPlus	N/A	N/A	218	7	1,028	167
MyOn	123	582	1,655	367	1,851	440
Total	23,090	28,566	25,877	6,931	23,417	7,090

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

## Usage Recommendations

Each vendor provided recommendations for using the software program in order for it to have an impact on students' literacy achievement (**Table 3**). Recommended weekly use ranged from 20 minutes to 80 minutes of use per week, and suggested weeks of use ranged from 15 to 28 weeks. For LEAs to continue to receive program funding, the state requires 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<sup>2</sup>.

<sup>2</sup> ETI submitted a separate report to the USBE on school level fidelity.

Table 3. Vendor 2017-2018 Minimum Dosage Recommendations

Program	Kindergarten ALL Students	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	45-60 min/week	28 weeks
Imagine Learning	40 min/week	45 min/week	45 min/week	45 min/week	18 weeks
SuccessMaker	45 min/week	45 min/week	45 min/week	45 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	45-60 min/week	45-60 min/week	45-60 min/week	45-60 min/week	20 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 for students who worked at or above grade level.

# 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 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 DIBELS data. In Grades 2-3, our focus was on struggling readers, and we needed valid SSIDs in the vendor and DIBELS data as well as beginning-of-year DIBELS scores to identify the students reading below grade level.

### **Impact Study 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 A](#) for descriptive statistics of the students included in our samples*). Our analytic 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. Students needed to have accurate state student Ids (SSIDs) and complete DIBELS 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 relied on comparing program students' achievement outcomes to non-program students' outcomes (known as “control students”), so that we could analyze what impact the program had on learning achievement. Control students were drawn from schools across the state of Utah who did not participate in the EISP. 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.*** 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 students dosage characteristics affected their outcomes. We operationally defined the combination of weekly use and weeks of use as “program dosage”. We created three matched 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:

- The **Highest Dosage** sample was comprised of students who met the vendors recommended use (in minutes) 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. In past reports this sample was referred to as the optimal (OPTI) sample.
- The **Medium Dosage** group use sample was comprised of students who used the program greater than or equal to 80% of vendors’ recommended use<sup>3</sup>. Students in this sample had the second highest program dosage. In past reports this sample was referred to as the relaxed optimal (ROPT) sample.
- The **Lowest Dosage** sample includes all students who used the program for any amount of time, and shows how effective the program was irrespective of use. In past reports this sample was referred to as the intent to treat (ITT) sample.

***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; “Medium Dosage”). Although we lowered our minimum dosage threshold, there were certain instances, for certain vendors and grades, in which the sample size was still too small for us to detect small program effects<sup>4</sup>. For these instances, we used the Lowest Dosage sample (all students, regardless of use) and reported any findings which were statistically significant. 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.

---

<sup>3</sup> “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.

<sup>4</sup> We identified all instances in which we had an insufficient sample size for using the nomenclature, IS (e.g. insufficient sample).

## 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: 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; and student information system (SIS) demographic data provided by the Utah State Board of Education (USBE). See [Appendix B](#) for details on how we created our master dataset.

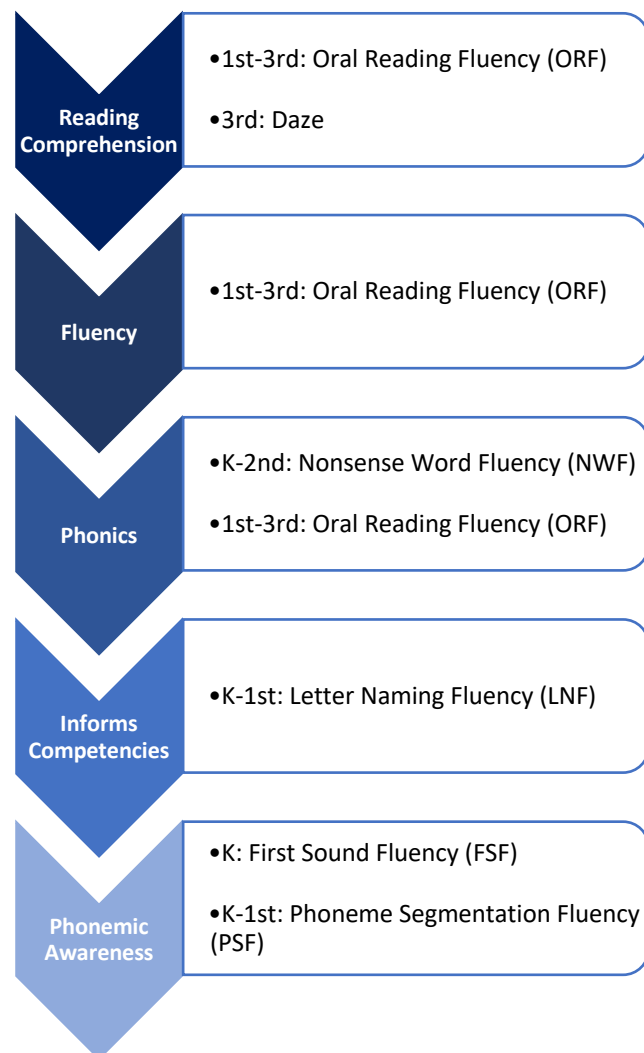
## Which instruments did we use to measure literacy achievement?

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 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 C](#) for additional detail on the DIBELS Next 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.

Figure 1. DIBELS Indicator & Literacy Skill Measures





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

Our study relied on three types of statistical analyses, all based on comparing the program samples to matched groups of their peers, which included: hierarchical linear models, independent t-test mean score comparisons, and benchmark outcome visual analyses.

*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 used regression analyses to study how program participation impacted students with specific characteristics, such as English Language Learners, special education students, economic disadvantaged students, non-white students, and students from Title 1 schools. We included students who met our criteria for the highest program dosage in this analysis sample.

*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. Changes in students’ reading proficiency benchmark levels were reported for the highest dosage 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?*

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.



## *How did we study the multi-year trends in program implementation and program impacts?*

There were several changes made to the evaluation design and methods throughout the duration of the evaluation and we focused on the past three implementation cycles (2015-2016, 2016-2017, and 2017-2018) to report findings in which our analyses methods were consistent across years. We reported the effect sizes for three levels of program dosage to study the entire program over time, and visually identified the grade levels in which vendors had an impact on students' literacy achievement. To study the trends in program implementation, we reported students' average weeks of use, total minutes of use, and weeks of use. Program usage descriptives reported prior to 2015-2016 were estimated from students' total minutes and the program start-and-end dates, while usage reported after 2015-2016 was calculated from actual weeks of use<sup>5</sup>.

## *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<sup>6</sup>, 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). 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.<sup>7</sup> More information on how we selected our ES benchmark is provided in [Appendix D](#).

---

<sup>5</sup> Beginning in 2015-2016, we received weekly program use data from vendors and calculated more accurate descriptives.

<sup>6</sup> Effect sizes are calculated by taking the difference in the two groups means divided by the average of their pooled standard deviations.

<sup>7</sup> 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).

# 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 is 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. 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 E](#).

## *Did students use the software as intended?*

**Key Finding:** The percentage of students who met vendors' weeks of use and average use recommendations increased from last year within each grade:

- Students who met the average minutes recs increased by 10% in kindergarten; 17% in 1<sup>st</sup> grade; 14% in 2<sup>nd</sup> grade; and 10% in 3<sup>rd</sup> grade.
- Students who met the weeks of use recs increased by 17% in kindergarten; 12% in 1<sup>st</sup> grade; 11% in 2<sup>nd</sup> grade; and 10% in 3<sup>rd</sup> grade.

As shown in **Figure 2**, a majority of students used their respective software programs for the minimum weeks<sup>8</sup> recommended: 72-83% of students among 5/7 vendors. This finding indicates that LEAs are 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.<sup>9</sup> Among the seven vendors, there were three vendors in which at least half of their students used the software for the recommended minutes per week, on average (Core5, Imagine Learning, and SuccessMaker). Students using ReadingPlus and MyOn had the lowest percentage of students who met the average minutes recommendations.

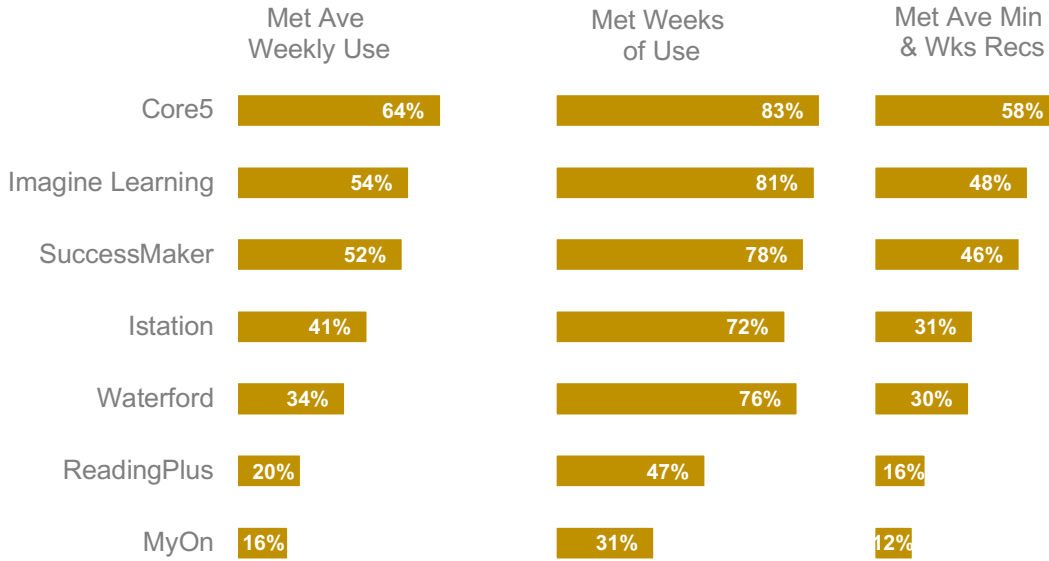
The percentage of students who met vendors' recommendations for both average minutes and total weeks is presented in the last column of Figure 2. These students used the programs as intended on both aspects of dosage: weekly minutes and total weeks. Over half of the students who used Core5 met both recommendations, and almost half of Imagine Learning and SuccessMaker students reached this goal.

---

<sup>8</sup> Vendor recommendations for total weeks of use ranged from 15-28 weeks.

<sup>9</sup> Vendor recommendations for average minutes per week ranged from 45-80 minutes. Core5 had lower recommendations for non-intervention students: 20 minutes per week.

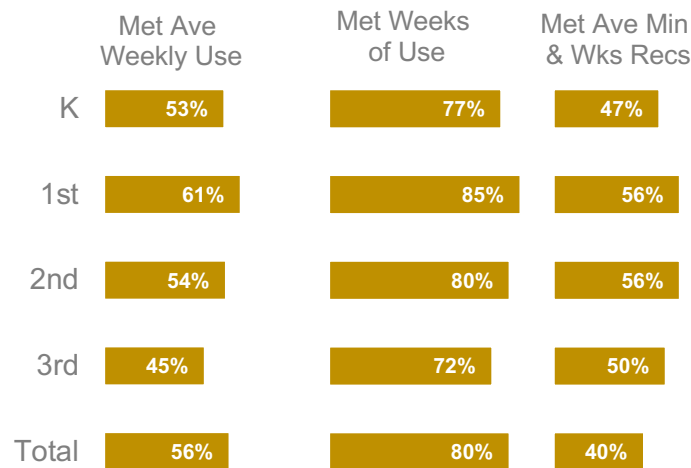
Figure 2: Students who met vendors minimum dosage recommendations



N: Istation (926); Waterford (5,712); IL (23,997); SM (1,220); Core5 (32,136); RP (174); MyOn (1,512)

**Figure 3** provides an overview of program use within each grade. Forty-five to 61 percent of students met the average minutes recommendations across grades, while 72 to 85% met the minimum weeks requirements. Fewer students met the average weekly use recommendations in third grade (45%) among all the grades; however, kindergarten students had the fewest (47%) to meet both the minutes and weeks recommendations.

Figure 3: Students who met the dosage recommendations by grade



N: K (23,090); 1st (28,566); 2nd (6,931); 3rd (7,090)

# 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 seven 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.

## Program-Wide Impacts

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 **Table 4**, the statistically significant program-wide effects on DIBELS Next end-of-year (EOY) composite scores increase with dosage, and the more a student used the program the better his/her EOY outcomes.


- In kindergarten, the treatment effects tripled when you move from the lowest dosage to the highest dosage sample.
- In first grade, students in the highest dosage sample had slightly more than four-fold the effects size when compared to the lowest dosage sample.
- In second grade, there were no statistically significant treatment effects.
- In third grade, only the highest dosage sample produced a statistically significant effect.

### **Who is included in each dosage sample?**

- **Highest Dosage:** students met vendors' recommendations for at least 80% of the weeks it was used, and used it for the total weeks recommended by vendors.
- **Medium Dosage:** students met at least 80% of vendors recommended dosage
- **Lowest Dosage:** students of all usage.

Effect sizes (ES) describe 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 define this threshold as any effect size equal or greater to .13, which is the average effect size seen in similar intervention programs (Lipsey et. al, 2012). Students with the highest program dosage in kindergarten, first and third grade had the highest treatment effect sizes overall, as measured by their average DIBELS Next Composite scores (ES: .16, .09 and .1, respectively). The .16 effect size in kindergarten is meaningful when compared against the average effect size of .13 produced by similar intervention programs.

Table 4. Predicted Means of DIBELS Composite Scores for Matched Treatment and Control, Program-wide, Highest to Lowest Dosage Samples

	Kindergarten			1 <sup>st</sup> Grade			2 <sup>nd</sup> Grade Intervention			3 <sup>rd</sup> Grade Intervention		
	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES	Tr.	C	ES
<b>Highest Dosage</b>	N=12,152			N=17,250			N=3,542			N=2,772		
	157	144	<b>.16*</b>	209	198	.09*	170	166	NS	278	268	.1*
	23,150			26,682			7,188			6,174		
	149	140	.09*	191	184	.06*	158	159	NS	260	257	NS
<b>Lowest Dosage</b>	31,362			28,252			8,750			9,162		
	145	140	.05*	187	184	.02*	154	156	NS	254	257	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 .13, the average for similar intervention programs, are highlighted in bold.  
\*p ≤ .05.

In addition to examining the program effects on composite measures of literacy, we examined the program's benefits on specific literacy skill development in **Table 5**. 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 6 points). The largest difference in mean scores was observed for developing kindergarten students alphabetic principles and basic phonics skills (NWF: CLS), with program students scoring 6 points higher, on average, than the control group. Program participation had less of an impact in the upper-early grade levels. Program students did slightly better than non-program students on measures of basic phonics (NWF) and reading comprehension (ORF) in first grade, fluency in second grade, and fluency and reading comprehension in third grade.

Table 5. Predicted Means of EOY DIBELS Literacy Domains for Matched Treatment and Control, Highest Dosage Sample

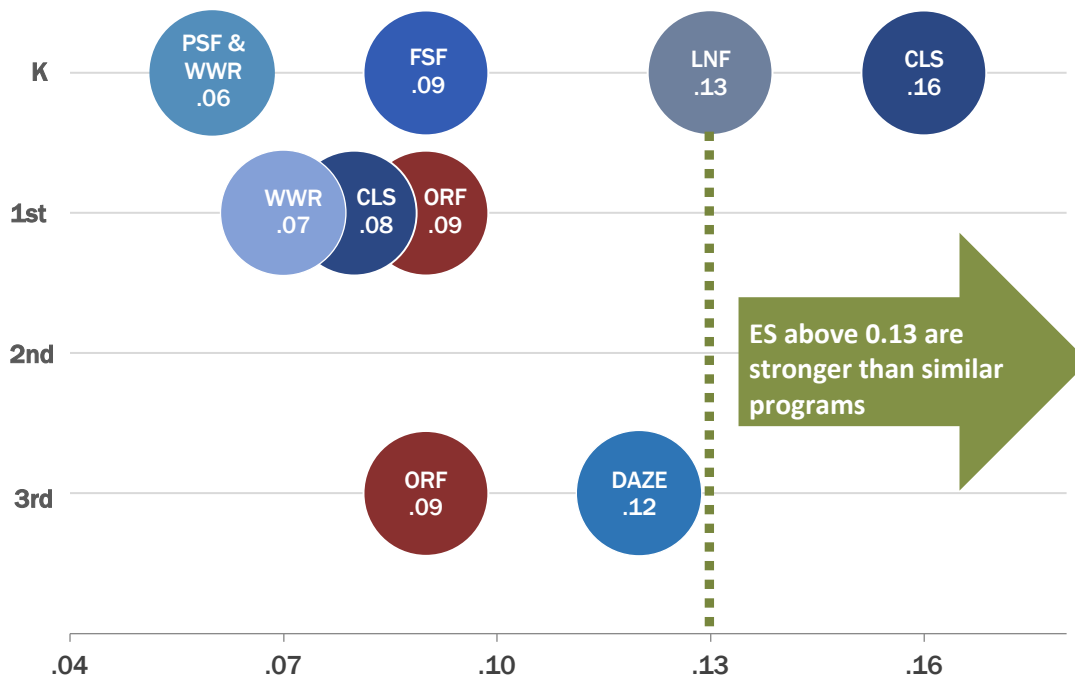
DIBELS Scale	Kindergarten			1 <sup>st</sup> Grade			2 <sup>nd</sup> Grade			3 <sup>rd</sup> Grade		
	N=			N=			N=3,542			N=2,722		
	12,106 – 12,152			17,200-17,246								
	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.
First Sound Fluency (FSF)	40**	37	3		N/A			N/A			N/A	
Letter Naming Fluency (LNF)	54**	50	4		N/A			N/A			N/A	
Phoneme Segmentation Fluency (PSF)	53**	51	2		N/A			N/A			N/A	
Nonsense Word Fluency-CLS	50**	44	6	89**	85	4		N/A			N/A	
Nonsense Word Fluency-WWR	9**	8	1	28**	26	2		N/A			N/A	
Oral Reading Fluency (ORF)		N/A		70**	68	2		NS		76**	74	2
DAZE		N/A			N/A			N/A		14**	13	1

Note. NS (not significant) in a cell means the program did not have a statistically significant effect. N/A: measure not administered in grade.

\*p ≤ .05. \*\*p ≤ .01.

In **Figure 4**, we present the effect sizes for each statistically significant DIBELS literacy domain in which the treatment group had higher mean scores compared to the matched control group to aid in interpreting the practical significance of the findings. Effect sizes increase in strength from the left to the right of the figure and are plotted by grade. As expected, we see significant treatment impacts for grade levels in which the DIBELS composite reading scores were also significant (kindergarten, first and third grade). There were no statistically significant treatment effects for either the composite or for specific literacy domains in second grade. Two subscales in kindergarten produced effects greater than or equal to similar intervention programs: Letter Naming Fluency (LNF) and Nonsense-Word Fluency: Correct letter sounds (NWF: CLS;). Letter naming fluency measures students' ability to recognize letters and Nonsense-Word Fluency measures students' understanding of alphabetic principles and blending.

Figure 4. DIBELS Literacy Domain Effect Sizes by Grade, Highest Dosage Sample



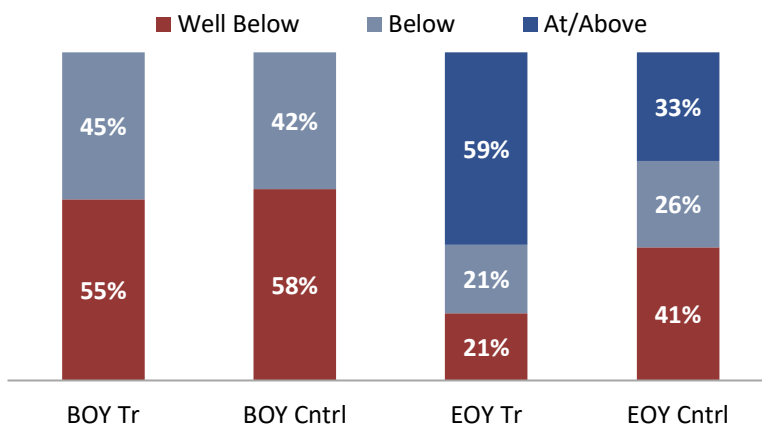
\*Note: All data points displayed in figure were statistically significant at  $p \leq .05$ .

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

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 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 depict the percent of students who started the year Well Below Benchmark or Below Benchmark for their grade, and follow their change in reading status in comparison to their non-program counterparts (**Figures 5-8**). 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, followed by a small difference in third grade. We describe the findings for each grade level in more detail in the following paragraphs.

**Kindergarten:** In kindergarten, 360 EISP students and 360 comparison students in the matched Highest Dosage sample began the school year below grade level based on their beginning of year reading DIBELS scores. Of these, 59 percent in the treatment group ended the year reading at grade level, compared to 33 percent of comparison students (a difference of 26 percent).

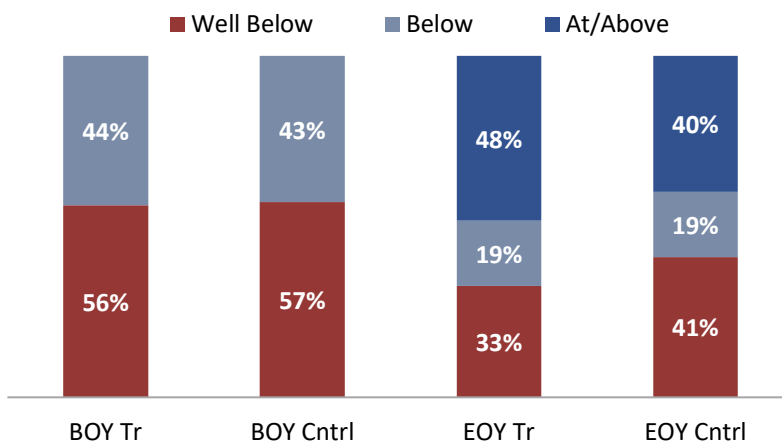
Figure 5. % Change in Benchmark Status from BOY to EOY, Kindergarten



Data source: Students reading below benchmark at BOY, matched kindergarten Highest Dosage sample. N: 720

**First Grade:** Among the program students in first grade who started the year reading below grade level, 48 percent (1,182/2,450) were reading at grade level by year end (**Figure 6**). In comparison, 40 percent of the non-program students (963/2,417) moved from reading below grade level to reading at grade level from beginning (BOY) to the end of the school year (EOY). The difference in growth between the treatment and comparison group was 8 percent.

Figure 6. % Change in Benchmark Status from BOY to EOY, 1<sup>st</sup> Grade

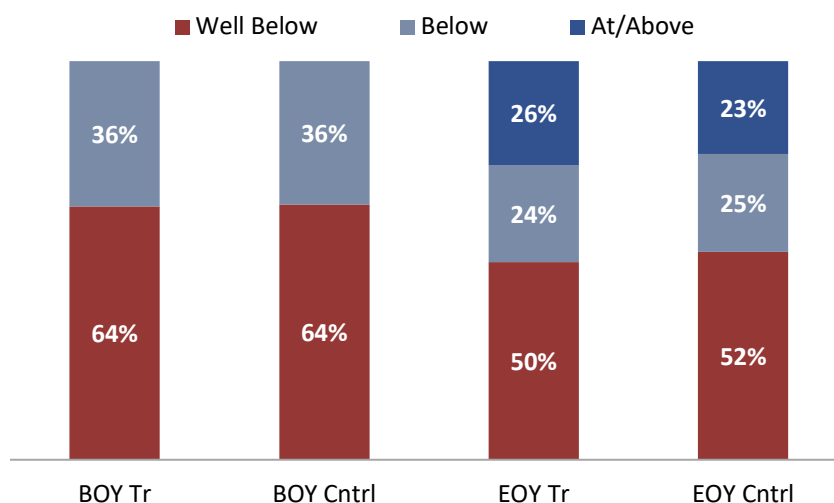


Data source: Students reading below benchmark at BOY, matched 1<sup>st</sup> Grade Highest Dosage sample. N: 4,863



**Second Grade:** As shown in **Figure 7**, the difference in growth between program and non-program struggling readers in second grade was negligible: 3% more program students were reading at grade level than their non-program peers. Moreover, within the sample of struggling 2<sup>nd</sup> grade readers, only a small percentage reached At/Above Benchmark status (26% of treatment students vs. 23% of control students). Approximately half of the students in both groups fell within the Well Below benchmark category at EOY, indicating that there is a 10-20% likelihood of these students achieving subsequent reading goals without intensive support outside of core curriculum (Dynamic Measurement Group, 2016).

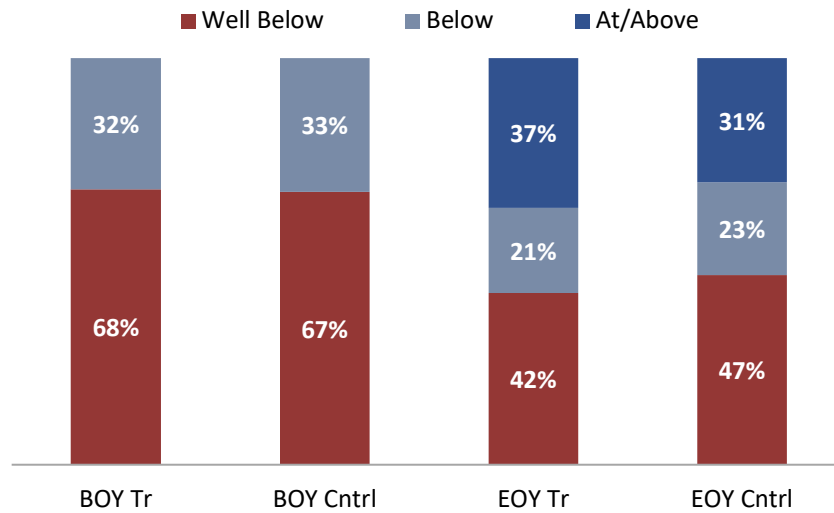
Figure 7. % Change in Benchmark Status from BOY to EOY, 2nd Grade



Data source: Students reading below benchmark at BOY, matched 2nd Grade Highest Dosage sample. N: 3,372

**Third Grade:** In **Figure 8** we can see that slightly more third grade program students were reading at grade level compared to non-program students by the end of the school year (a 6% difference). Thirty-seven percent of program students and 31 percent of non-program students in the matched Highest Dosage sample identified as Below or Well Below Benchmark at the beginning of the school year reached At/Above benchmark status by year end.

Figure 8. % Change in Benchmark Status from BOY to EOY, 3rd Grade



Data source: Students reading below benchmark at BOY, matched 3rd Grade Highest Dosage sample. N: 2,626

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

**Table 6** shows the mean score differences in DIBELS composite scores at program exit for certain subgroups of program students. Program students who were identified as low-income, special education (SPED), and English Language Learners had lower predicted means scores than their higher income, general education, and English speaking program counterparts in specific grades. These differential treatment effects were the most pronounced for special education students in Grades 1 and 3: in 1st grade they scored 34 points lower and in 3<sup>rd</sup> grade they scored 47 points lower than general education treatment students.

Table 6. Mean Score Differences on EOY DIBELS Composite Scores by Grade and Subgroup, Highest Dosage Sample

	Kindergarten	1 <sup>st</sup> Grade	2 <sup>nd</sup> Grade	3 <sup>rd</sup> Grade
Low-income	-2	-13	-8	-6
Special Education (SPED)	-13	-34	-23	-47
Title I Schools	8	NS	NS	NS
ELL	NS	-20	-10	NS
Non-white	-2	NS	NS	16

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

Kindergarten (N=12,024); 1<sup>st</sup> Grade (N=17,250); 2<sup>nd</sup> Grade (N=3,542); 3<sup>rd</sup> Grade (N=2,722)

All mean differences displayed in table were statistically significant at  $p \leq .05$ .

# Individual Vendor Impacts

The vendor-specific analyses were designed to help program stakeholders understand the effectiveness of the individual programs and make informed decisions. With this in mind, we have done our best to conduct comprehensive analyses in which readers 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 a mean comparison between each program and a matched control group that shows program effects on overall literacy scores.

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

**Table 7** presents the predicted means and mean score differences of program and non-program students in the matched medium dosage sample for each vendor and grade. Vendors with sample sizes that may be too small to detect small program effects were identified with “IS”, insufficient sample, and findings that were not statistically significant were identified as “NS”, not significant. Five vendors had a positive impact on students in kindergarten (Istation, Waterford, Imagine Learning, SuccessMaker, Core5), followed by two vendors in first grade (Imagine Learning; Core5), and one vendor in second grade (Imagine Learning). There were no statistically significant findings in third grade for the vendor specific analyses. In kindergarten and first grade, the average predicted DIBELS 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 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) at end-of-year.

Table 7. Predicted Means of EOY DIBELS Composite for Matched Treatment and Control, by Vendor, OLS Regression Model

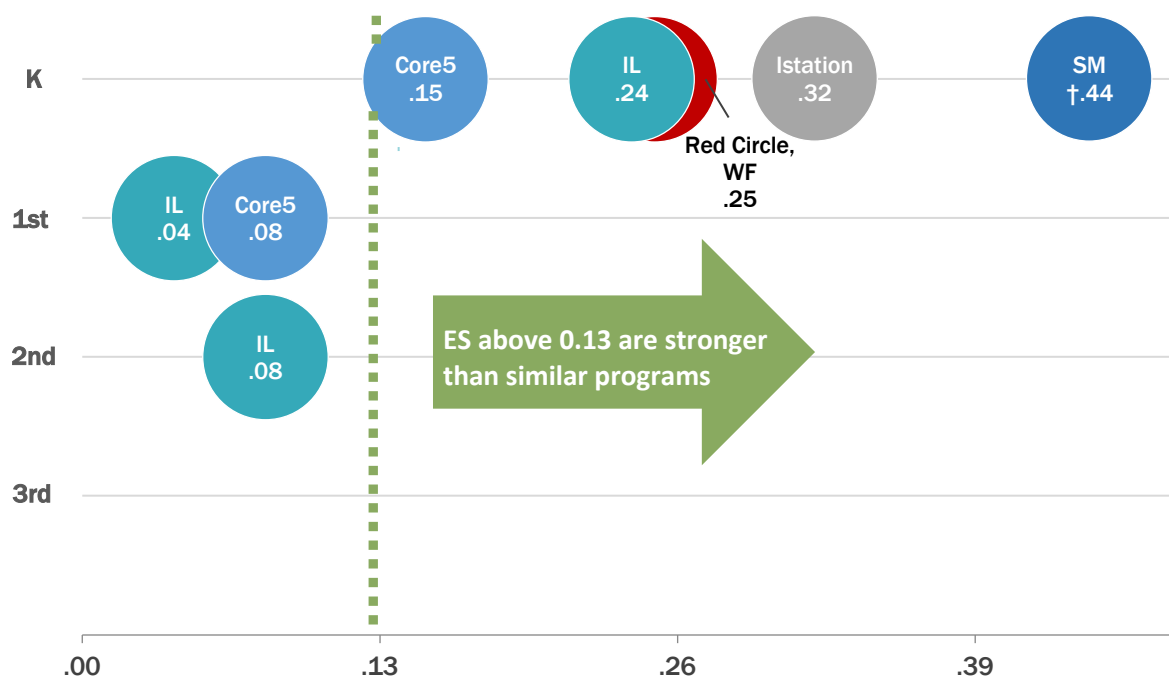
	K			1 <sup>st</sup>			2 <sup>nd</sup>			3 <sup>rd</sup>		
	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.
Istation	N=244			N=436			N=136			IS		
	160	148	12		NS			NS				
WF	2,734			1,816			204			IS		
	146	137	9		NS			NS				
IL	8,110			13,546			2,940			1,806		
	146	137	9	184	182	2	161	156	5		NS	
SM	322†			696			178			272		
	141	131	9		NS			NS			NS	
Core5	12,454			16,268			4,196			4,114		
	151	145	6	200	195	5		NS			NS	
RP	N/A			N/A			IS			IS		
MyOn	IS			NS			104			250		
								NS			NS	

*Note.* 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. † Lowest dosage sample reported for SM in kindergarten.

\*p ≤ .05

Like the program-wide analyses, we present effect sizes for the individual analyses to identify the strength of the treatment effects in relationship to similar intervention programs. Effect sizes increased in strength from the left to the right of the **Figure 9** and are plotted by grade. Effect sizes to the right of the dotted line are stronger than the average effect sizes produced by similar intervention programs and are therefore more meaningful based on this frame of reference. As displayed in **Figure 9**, all vendors that were used with kindergarten students produced effect sizes greater than the effect size benchmark, including: SuccessMaker (ES.44), Istation (.32), Waterford (ES: .25), Imagine Learning (ES: .24), and Core5 (ES: .15). Vendors in Grades 1-3 had small effect sizes, none of which were greater than the effect size benchmark.

Figure 9. Impact of Individual Vendors on DIBELS Composite Scores, Effect Sizes by Grade



Note. IS for medium dosage group: Istation 3rd grade (n=16); WF 3rd grade (n=8); MyOn 1st grade (n=16); RP in 2nd/3rd grade (n=0-50). † Lowest dosage sample used for SM in kindergarten. IS for lowest dosage group: WF 3rd grade (n=34); RP 2nd grade (n=6). All data points displayed in figure were statistically significant at  $p \leq .05$ .

## Multi-year Findings

In this section we identify the key trends in program enrollment, student program use, and its impacts on student achievement across the past few years of program implementation.

### *What are the multi-year trends in program enrollment?*

**Table 8** depicts program enrollment of Local Education Agencies, schools and students over the past four years of the EISP. It is clear that program enrollment continues to increase exponentially with approximately 64,000 more students enrolled from 2014/2015 to 2017/2018.

Table 8. Est. Program Enrollment from 2013/2014 – 2017/2018

	Includes all K-3 students		All K-1 & 2/3 intervention only		
	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018
LEAs	32	45	72	79	79
Schools	145	218	388	338	403
Students	38,553	36,790	68,891	86,723	100,951

*Note.* Data reported prior to 2015-2016 includes non-intervention students in Grades 2-3. Student counts may contain duplicates and should be seen as estimates.

### *What are the multi-year trends in students' program use?*

**Table 9** presents the change in average usage from 2013-2014 to 2017-2018. Prior to 2015-2016, we estimated students' average weekly use and total weeks of use<sup>10</sup>, and we did not present these usage statistics for those years. In 2015-2016, we received weekly program use data from vendors, which provided us with more accurate usage statistics. As displayed in **Table 9**, LEAs appear to be doing a better job overall with program implementation from 2015-2016 to 2017-2018 in all three areas (average minutes, total weeks, and total minutes).

<sup>10</sup> Averages were calculated from students' total minutes and the program start-and-end dates prior to 2015-2016.

Table 9. Multi-year Trends in Program Use

	2013-2014	2014-2015	2015-2016	2016-2017	2017-2018
Avg. Minutes per Week	--	--	46	52	54
Total Weeks	--	--	20	23	25
Total Minutes	1,118	1,013	1,047	1,292	1,455

*Note.* Three additional programs (MyON, ReadingPlus, Core5) were added to EISP in 2015-2016, and i-Ready was removed as a vendor in 2016-2017.

### *How did the program-wide impacts change from year-to-year?*

Over the past three years of program implementation, the strongest program effects were reported for students with the highest program dosage and for students in kindergarten (**Table 10**). There were no consistent trends found for students in other grade levels across the years. For example, within the highest dosage group, the program was most effective for kindergarten and 2<sup>nd</sup> grade students in 2015-2016, for all K-3 students in 2016-2017, and for all grades except 2<sup>nd</sup> grade in 2017-2018. The effect sizes appear to diminish in strength from 2015-2016 to 2017-2018, highlighting a need for greater understanding about the programs' implementation to maximize the programs impact on students' literacy achievement. These findings may be due to an increase in similar intervention programs used by control students, to differences in how students engage with the program during their time using the software, among other reasons.

Table 10. Trends in program-wide impacts, effect sizes by dosage sample

	Dosage Group	2015-2016	2016-2017	2017-2018
Kindergarten	Highest Dosage	<b>0.36</b>	<b>0.20</b>	<b>0.16</b>
	Medium Dosage	<b>0.21</b>	0.11	0.09
	Lowest Dosage	0.09	0.06	0.05
1 <sup>st</sup> Grade	Highest Dosage	NS	<b>0.13</b>	0.09
	Medium Dosage	NS	0.05	0.06
	Lowest Dosage	-0.05	NS	0.02
2 <sup>nd</sup> Grade	Highest Dosage	<b>0.32</b>	<b>0.18</b>	NS
	Medium Dosage	0.09	0.08	NS
	Lowest Dosage	NS	NS	NS
3 <sup>rd</sup> Grade	Highest Dosage	NS	0.14	0.1
	Medium Dosage	NS	NS	NS
	Lowest Dosage	NS	NS	NS

*Note.* NS: non-significant. ES in bold were greater than or equal to 0.13, the average of similar intervention programs, and should be seen as meaningful impacts on learning.

## How did the vendor-specific impacts change from year-to-year?

**Figure 10** provides a visual representation of the grades in which statistically significant effect sizes were reported for each vendor from 2015-2016 to 2017-2018. Shaded grades depict effect sizes that were greater than or equal to .13, the ES benchmark for similar intervention programs. Vendors with consistent impacts across all three years and with effect sizes greater than or equal to the .13 benchmark included: Waterford, Imagine Learning, and Core5 in kindergarten. Core5 also had an impact on literacy achievement in other grade levels across all three years, but these effect sizes were not higher than the .13 benchmark each year. Imagine Learning also produced an effect size greater than or equal to .13 for second grade for two years (2015-2016 and 2016-2017).

When reviewing these findings, it is important to note that for certain vendors and grades, we could not study treatment effects for a group of students who met the minimum threshold of usage for inclusion in our study<sup>11</sup>. MyOn and ReadingPlus were most affected by limitations involving small sample sizes.

Figure 10. Grade Levels with Significant Effect Sizes by Vendor and Program Year

	2015-2016	2016-2017	2017-2018
Istation	<b>K</b>	NS	<b>K</b>
Waterford	<b>K</b>	<b>K 1 2</b>	<b>K</b>
ImagineLearning	<b>K 2</b>	<b>K 2 3</b>	<b>K 1 2</b>
Successmaker	<b>2</b>	NS	<b>K</b>
Core5	<b>K 1</b>	<b>K 1</b>	<b>K 1</b>
ReadingPlus	NS	NS	NS
MyON	<b>K</b>	<b>K 3</b>	NS

*Note.* All grades presented had positive and significant effect sizes. Outlined grades had effect sizes less than 0.13 and filled in grades had effect sizes that were equal to or greater than 0.13.

<sup>11</sup> Low enrollment, low overall percentages of students who used the program as intended, our focus on struggling readers in 2<sup>nd</sup> and 3<sup>rd</sup> grade, among other factors, all contributed to lower sample sizes for certain vendors and grades.



# Discussion, Limitations and Recommendations

## ***Understanding the Relationship between Program Dosage & Program Impacts***

The evaluation results from this year underscore how important it is to use the program for the recommended time, and that the overall program impacts were dependent on how much a student used the program (which we call “dosage”). These findings were consistent with last year’s results, which underscores the message even more. In Grades K-1, as dosage (use) increased program-wide literacy achievement also increased; however, the trends in Grades 2-3 are mixed, and are harder to understand.

## ***Understanding the Mixed Results in Grades 2-3***

We have found mixed results across yearly cohorts of students in Grades 2-3. For example, this year we did not find any positive program impacts in the second grade, while last year students in the second grade received benefits from being in the program. Similarly, our results for individual vendors lacked consistency in second and third grade, and we saw positive treatment effects for some vendors only in certain years. There are several possible reasons for inconsistencies across our findings in Grades 2-3, but we have identified two possible explanations: additional support for at-risk control students, and a lack of learning engagement even with high dosage program use.

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 DIBELS Next 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.

A second explanation for lackluster and inconsistent program impacts in Grades 2-3 is based on the use of program dosage to make determinations about which students were using the program at beneficial levels; however, dosage does not account for learning engagement and on-task vs. off-task behavior while in the program. From our 2016-2017 qualitative study of program implementation, we learned that program dosage is not the only important aspect of program implementation needed for the development of literacy skills. How students are using the program, such as learning engagement and time-on-task, may be just as important as how long they are using the program (Best Practices for Improving Early Intervention Software Programs in Utah Schools, 2017). The following quote from the 2016-2017 implementation study highlights this point:

*Evaluation and Training Institute*

*“They [students] are meeting their goal, but they're not progressing...Often times, it's because they're closing out the activity, they're just wasting their time. They're spending 60 minutes but they're opening one activity, closing it out. Opening one activity, closing it out. Checking more minutes, so they're not ever completing anything.” (Teacher)*

Through this example we see how important it is for students to use the program appropriately and why meeting the usage recommendations may not be enough to increase the literacy skills of all students. Students in Grades 2-3 are older than their K-1 counterparts and are more desensitized to the novelty of using interactive software program than the younger students. Students in third grade have a bevy of high-tech gadgets and software to interact with, and it may be harder to keep them engaged in a school-based reading program while they are logging minutes of use. In these cases it is easy to see why students may be logged in for the recommended time, but, in effect are “tuned out” and thus not having any positive impacts to their reading skills.

## 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.

*Control Sample Selection.* 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 are using similar intervention programs, which may underestimate the strength of the program impacts. It is also possible that some LEAs are using the same reading interventions with their students using a non-EISP funding source. We requested information from vendors to exclude these LEAs from inclusion in our control sample, but did not receive information on all the LEAs and from all the vendors. For future evaluations, it would be useful for the USBE and vendors to track and share this information with evaluators.

*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 analyses. 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. 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. Low usage was not the only

factor impacting the size of our sample. Other factors that affected the sizes of our samples included: students who used more than one software program, incomplete DIBELS scores, and other missing or incorrect data (such as student IDs).

## Recommendations & Future Research

We have clear evidence of the program’s consistent effectiveness in kindergarten for all vendors who were used in this grade level and, as a result, we recommend that the program continue to be used with kindergarten students. The evidence is less clear in Grades 1-3, and the strength of the programs’ effects vary depending on a combination of the vendor used, students’ dosage, among other factors. Future research is needed to increase our understanding of the conditions which lead to increases in literacy achievement for specific vendors and students. We propose two recommendations for future research:

We recommend expanding the definition of fidelity of program use to include measures beyond minutes and weeks of use (dosage). We understand that program dosage is one aspect of program implementation leading to increases in students’ literacy achievement, but that dosage does not account for learning engagement or time-on-task. Learning engagement results in learning progression, which could be an alternative or adjunct measure of program implementation for studying program use and impacts on reading. Stated simply, beyond being logged into the program software, students must also progress through the lesson content for learning to occur. If students are not progressing, it is important to understand the scope of the problem and the reasons why students may not be progressing as expected (e.g. “Are certain types of students unable to move forward in some programs, or are students intentionally wasting their time?”). Future research may help identify the percentage of students who are progressing through the program content based on the time they are spending in the software and explore the link between lesson progression and literacy outcomes. With an increased understanding of this aspect of program implementation, we can make targeted recommendations to improve the program’s efficacy and help students realize the full program benefits.

It is challenging for us to endorse the program’s use for specific program vendors as sample sizes, samples used for analysis and other factors varied across the evaluation. It is possible the program may work well for students using certain vendors, but we need more students to determine this. Our second recommendation is designed to address this issue, and we propose combining data across years for vendors with small sample sizes. By increasing our sample size, we would also increase our ability to detect small program effects.

# References

- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2<sup>nd</sup> ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dynamic Measurement Group, Inc. (2016, September). *DIBELS Next Benchmark Goals and Composite Score*. <https://dibels.org/papers/DIBELSNextBenchmarkGoals.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 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\\_assessmentmanual.pdf](http://wenatchee.innersync.com/assessment/documents/dibelsnext_assessmentmanual.pdf)
- Evaluation and Training Institute

# Appendix A: Analyses Samples

**Tables A1 – A4** 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 were the same.

## Program-wide Analyses Samples

Table A1. Program-Wide Sample by Grade, Lowest Dosage

	N	Female	Caucasian	Hispanic	Asian	African American	Other	SPED	Low-income	ELL	BOY Comp
K	15,681	7,652 49%	12,574 80%	2,056 13%	138 1%	138 1%	775 5%	1,265 8%	4,234 27%	929 6%	37
1st	14,126	6,903 49%	11,532 82%	1,798 13%	94 1%	103 1%	599 4%	1,202 9%	4,237 30%	838 6%	122
2nd	4,375	2,192 50%	3,213 73%	918 21%	17 0%	45 1%	182 5%	971 22%	2,003 46%	547 13%	72
3rd	4,581	2,192 48%	3,313 72%	1,047 23%	18 0%	35 1%	168 4%	1,128 25%	2,179 48%	701 15%	125

Table A2. Program-Wide Sample by Grade, Medium Dosage sample

	N	Female	Caucasian	Hispanic	Asian	African American	Other	SPED	Low-income	ELL	BOY Comp
--	---	--------	-----------	----------	-------	------------------	-------	------	------------	-----	----------

K	11,575	5,624 49%	9,107 79%	1,658 14%	96 1%	91 1%	623 5%	829 7%	3,504 30%	786 7%	38
1st	13,342	6,549 49%	11,059 83%	1,583 12%	81 1%	84 1%	535 4%	1,055 8%	3,910 29%	765 6%	123
2nd	3,594	1,773 49%	2,651 74%	744 21%	14 0%	37 1%	148 4%	742 21%	1,636 46%	451 13%	74
3rd	3,087	1,464 47%	2,156 70%	769 25%	14 0%	25 1%	123 4%	703 23%	1,500 49%	536 17%	126

Table A3. Program-Wide Sample by Grade, Highest Dosage sample

	N	Female	Caucasian	Hispanic	Asian	African American	Other	SPED	Low- income	ELL	BOY Comp
K	6,076	2,909 48%	4,774 79%	866 14%	52 1%	45 1%	339 6%	388 6%	1,737 29%	451 7%	41
1st	8,626	4,119 48%	7,287 84%	897 10%	47 1%	35 0%	360 4%	604 7%	2,360 27%	429 5%	131
2nd	1,685	821 49%	1,265 75%	342 20%	4 0%	10 1%	64 4%	341 20%	742 44%	206 12%	79
3rd	1,308	643 49%	909 69%	330 25%	7 1%	7 1%	55 4%	282 22%	661 51%	246 19%	134

## Individual Vendor Impacts Analyses Samples

Table A4. Vendor-specific Matched Sample by Grade

	Grade	N	Female	Caucasian	Hispanic	Other	Ave Minutes	Ave Wks.	SPED	Low- income	ELL	BOY Comp
Waterford	K	1,379	679 49%	1,128 82%	183 13%	68 5%	66	32	112 8%	491 36%	57 4%	36
	1	908	428 47%	797 88%	71 8%	40 4%	78	32	97 11%	337 37%	28 3%	118
	2	102	46 45%	87 85%	9 9%	6 6%	81	33	23 23%	59 58%	4 4%	65
	3	IS										
Imagine Learning	K	4,101	1,985 48%	3,126 76%	669 16%	306 8%	51	28	319 8%	1,369 33%	375 9%	35
	1	6,773	3,319 49%	5,579 82%	875 13%	319 5%	57	30	589 9%	2,244 33%	455 7%	120
	2	1,470	695 47%	1,095 74%	306 21%	69 5%	58	29	336 23%	676 46%	189 13%	72
	3	903	415 46%	616 68%	252 28%	35 4%	54	28	249 28%	461 51%	187 21%	119
Core5	K	6,329	3,059 48%	4,877 77%	964 15%	488 8%	63	28	410 6%	1,791 28%	539 9%	41
	1	8,134	3,983 49%	6,642 82%	1,004 12%	488 6%	68	30	575 7%	2,189 27%	507 6%	128
	2	2,098	1,059 50%	1,450 69%	512 24%	136 6%	71	31	402 19%	996 47%	318 15%	76
	3	2,057	992 48%	1,354 66%	561 27%	142 7%	65	30	422 21%	1,063 52%	406 20%	126

	Grade	N	Female	Caucasian	Hispanic	Other	Ave Minutes	Ave Wks.	SPED	Low- income	ELL	BOY Comp
SuccessMaker	K <sup>+</sup>	161	73 45%	142 88%	13 8%	6 4%	43	23	11 7%	44 27%	0 0%	35
	1	348	177 51%	275 79%	44 13%	29 8%	53	24	27 8%	102 29%	29 8%	120
	2	89	37 42%	70 79%	14 16%	5 6%	54	23	18 20%	43 48%	9 10%	86
	3	136	66 49%	107 79%	24 18%	5 4%	51	27	36 26%	63 46%	12 9%	123
Istation	K	122	57 47%	92 75%	25 20%	5 5%	66	30	10 8%	52 43%	12 10%	36
	1	218	103 47%	190 87%	18 8%	10 5%	85	31	12 6%	43 20%	15 7%	133
	2	68	31 46%	39 57%	22 32%	7 10%	128	31	9 13%	40 59%	21 31%	71
	3	IS										
MyOn	K	N/A	-	-	-	-	-	-	-	-	-	-
	1	IS										
	2	IS										
	3	125	57 46%	108 86%	15 12%	2 2%	56	28	32 26%	38 30%	4 2%	140

*Note.* IS: insufficient sample in cell indicates vendor had insufficient sample size to report findings for the medium dosage threshold.

<sup>+</sup> Indicates Lowest Dosage (ITT) Sample was used.



## Appendix B. Data Processing & Merge Summary

We reviewed and cleaned data from nine different sources in preparation of completing our analyses, including program usage data from seven software program providers, combined 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, correcting invalid SSIDs where possible, 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 101,613 to 100,951 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<sup>12</sup> (approximately 6,000 cases) and any IDs that did not comply with the state student ID (SSID) format (N=482 after fixing IDs). 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. For similar reasons, we excluded students who used Imagine Learning through a separate state-

---

<sup>12</sup> These IDs were also deleted from our pool of potential control students.  
*Evaluation and Training Institute*

wide grant<sup>13</sup> prior to reporting the program impacts. See **Table B1.** 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 in Grades K-3. We reviewed the SIS data provided by the USBE to ensure that all LEAs who were listed as 2017-2018 participants were included in the data. The raw data file consisted of 214,578 cases, of which almost five percent were duplicate records. After cleaning the data of duplicates, our SIS data consisted of 203,747 records.

## DIBELS Next Data

In 2018-2019, the USBE prepared and transferred a DIBELS Next data file (n=193,501). 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 193,348 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 and were left with 190,041 cases. Next, we merged our master vendor data into the DIBELS and SIS data, removed non-intervention students in Grades 2-3, and missing data (e.g. beginning and end-of-year composite scores). After completing these steps, the data consisted of 175,039 cases. Lastly, we identified (where possible) schools or students using one of the seven program vendors through non-EISP funding and removed these cases from our pool of potential controls<sup>14</sup>. After processing the data, our final, pre-matched dataset consisted of 98,104 cases, of which, 55,065 were treatment and 43,005 were potential controls.

## Impact of Data Cleaning on Vendor Samples

The table below depicts the impact of the different stages of the cleaning process within 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 and ReadingPlus lost a lot of cases due to duplicate IDs, which indicates schools may be using the programs outside of the expectations as 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 74% and 85%, respectively).

---

<sup>13</sup> 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.

<sup>14</sup> We removed students from non-EISP funded schools who were using an EISP program based on information provided by vendors.

Table B1. Overview of Data Cleaning Process by Program

	N	With Valid IDs	Duplicate Cases of IDs Removed*	Merged	Primary Cases of Dups Removed**	With Complete DIBELS/SIS Data	IL Contamination Removed	Intervention Only
Istation	1,238	1,237	1,177	1,093	1,091	1,041	1,038	764
Waterford	6,399	6,370	5,645	4,826	4,819	4,352	4,250	3,928
Imagine Learning	33,035	33,033	31,868	30,886	30,647	28,759	28,756	20,949
Success-Maker	2,015	2,015	2,002	1,963	1,907	1,799	1,699	1,033
Core5	52,807	52,412	51,302	50,118	49,319	46,776	45,236	27,754
Reading-Plus	1,246	1,213	1,211	1,076	583	511	496	74
MyOn	4,211	4,189	4,185	3,820	2,411	2,217	2,197	563
Total	100,951	100,469	97,390	93,782	90,777	85,455	83,672	55,065

*Note.* First column “N’s” represent count of students after cleaning individual vendors’ data.

\*Still contains primary cases that were duplicates. Removed after merge.

\*\*After removing primary cases of duplicate IDs.

## 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 Medium Dosage and Highest Dosage samples). However, for our largest sample of program students, the lowest dosage program-wide sample, there were more program students than control students. We had 55,065 treatment students and 43,005 potential control students. This automatically reduced the size of this particular sample. In addition, certain vendors and grades lacked a sufficient number of cases to detect small program effects to be included in our medium dosage matched sample (e.g. students who met 80% of vendors dosage recommendations), and we created a matched sample for all students, regardless of use (lowest dosage sample) in these instances.

## Appendix C: 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. 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 C1** provides a summary of the DIBELS subscales used in our analyses.

Table C1. 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
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](http://wenatchee.innersync.com/assessment/documents/dibelsnext_assessmentmanual.pdf)

## Appendix D: 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

# Appendix E. Program Use by Vendor and Grade

**Table E1** 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 E1. 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	# Wks. Met 80% Ave Min. Recs	Met 80% Ave Min. Recs		Met 80% Wks. Recs		Met 80% Min./ 80% Wks. Recs	
Istation	K	349	48	1297	26	72%	16%	12	154	44%	317	91%	148	42%
	1	356	69	1917	28	81%	62%	19	306	86%	305	86%	257	72%
	2	125	71	1898	26	65%	74%	19	112	90%	100	80%	92	74%
	3	95	43	964	21	46%	9%	6	11	12%	56	59%	10	11%
	Total	926	66	1783	26	72%	41%	15	583	63%	778	84%	507	55%
Waterford	K	2728	60	1801	29	77%	44%	20	2082	76%	2454	90%	1961	72%
	1	2584	69	2116	30	79%	25%	17	1523	59%	2317	90%	1431	55%
	2	283	61	1720	26	56%	21%	14	137	48%	191	67%	127	45%
	3	110	49	959	18	28%	45%	11	68	62%	42	38%	33	30%
	Total	5712	64	1923	29	76%	34%	18	3816	67%	5005	88%	3553	62%
Imagine Learning	K	8357	43	1055	23	78%	50%	15	5910	71%	7024	84%	5304	63%
	1	11011	51	1394	26	87%	58%	18	8627	78%	10002	91%	8137	74%
	2	2446	51	1362	25	81%	58%	17	1875	77%	2072	85%	1731	71%
	3	2181	43	991	21	66%	38%	12	1281	59%	1635	75%	1126	52%

		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	# Wks. Met 80% Ave Min. Recs	Met 80% Ave Min. Recs	Met 80% Wks. Recs	Met 80% Min./ 80% Wks. Recs			
	<b>Total</b>	<b>23997</b>	<b>47</b>	<b>1236</b>	<b>25</b>	<b>81%</b>	<b>54%</b>	<b>16</b>	<b>17695</b>	<b>74%</b>	<b>20735</b>	<b>86%</b>	<b>16300</b>	<b>68%</b>
Success-Maker	K	192	43	935	22	89%	44%	13	129	67%	177	92%	121	63%
	1	586	50	1021	20	74%	57%	14	497	85%	460	78%	427	73%
	2	185	44	878	19	77%	45%	12	116	63%	153	83%	109	59%
	3	257	45	1050	22	79%	53%	15	180	70%	212	82%	171	67%
	<b>Total</b>	<b>1220</b>	<b>47</b>	<b>992</b>	<b>20</b>	<b>78%</b>	<b>52%</b>	<b>14</b>	<b>922</b>	<b>76%</b>	<b>1002</b>	<b>82%</b>	<b>828</b>	<b>68%</b>
Core5	K	11282	54	1350	24	76%	60%	16	8062	71%	9317	83%	7310	65%
	1	13224	62	1782	28	89%	72%	21	10995	83%	12227	92%	10425	79%
	2	3518	62	1788	28	87%	57%	19	2684	76%	3224	92%	2544	72%
	3	3840	55	1513	26	80%	52%	17	2707	70%	3314	86%	2547	66%
	<b>Total</b>	<b>32136</b>	<b>58</b>	<b>1593</b>	<b>26</b>	<b>83%</b>	<b>64%</b>	<b>18</b>	<b>24648</b>	<b>77%</b>	<b>28230</b>	<b>88%</b>	<b>22959</b>	<b>71%</b>
ReadingPlus	2	IS												
	3	167	33	606	15	49%	20%	7	59	35%	101	60%	50	30%
	<b>Total</b>	<b>174</b>	<b>33</b>	<b>582</b>	<b>15</b>	<b>47%</b>	<b>20%</b>	<b>6</b>	<b>59</b>	<b>34%</b>	<b>101</b>	<b>58%</b>	<b>50</b>	<b>29%</b>
MyOn	K	123	22	178	8	2%	4%	2	11	9%	9	7%	2	2%
	1	582	22	235	9	10%	5%	3	62	11%	143	25%	26	4%
	2	367	33	665	17	44%	23%	9	116	32%	215	59%	102	28%
	3	440	37	822	20	56%	26%	12	192	44%	297	68%	164	37%
	<b>Total</b>	<b>1512</b>	<b>29</b>	<b>505</b>	<b>14</b>	<b>31%</b>	<b>16%</b>	<b>7</b>	<b>381</b>	<b>25%</b>	<b>664</b>	<b>44%</b>	<b>294</b>	<b>19%</b>

*Note.* Data source: vendor usage data in K-1 before excluding invalid SSIDs, duplicates, missing SIS/outcome data, contamination with other programs, etc. 2<sup>nd</sup>/3<sup>rd</sup> grade students merged with DIBELS to exclude non-intervention students (students reading at grade level at beginning of year).

# Appendix F: Multi-Year Dosage Recommendations

	Year	Kindergarten	1st Grade	2nd Grade Intervention	3rd Grade Intervention	Minimum Weeks
<b>Core5</b>	2015-2016	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
	2016-2017	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
	2017-2018	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
<b>Imagine Learning</b>	2015-2016	45 min/week	60 min/week	60 min/week	60 min/week	20 weeks
	2016-2017	45 min/week	60 min/week	60 min/week	60 min/week	20 weeks
	2017-2018	40 min/week	45 min/week	45 min/week	45 min/week	18 weeks
<b>Istation</b>	2015-2016	60 min/week	60 min/week	60 min/week	60 min/week	12 weeks
	2016-2017	60 min/week	60 min/week	60 min/week	60 min/week	28 weeks
	2017-2018	60 min/week	60 min/week	60 min/week	60 min/week	24 weeks
<b>MyOn</b>	2015-2016	45 min/week	45 min/week	45 min/week	45 min/week	20 weeks
	2016-2017	45-60 min/week	45-60 min/week	45-60 min/week	45-60 min/week	20 weeks
	2017-2018	45-60 min/week	45-60 min/week	45-60 min/week	45-60 min/week	20 weeks
<b>Reading Plus</b>	2015-2016	45 min/week	45 min/week	45 min/week	45 min/week	15 weeks
	2016-2017	45-75 min/week	45-75 min/week	45-75 min/week	45-75 min/week	15 weeks
	2017-2018	45 min/week	45 min/week	45 min/week	45 min/week	15 weeks
<b>Success-maker</b>	2015-2016	45 min/week	45 min/week	60 min/week	60 min/week	15 weeks
	2016-2017	45 min/week	45 min/week	60 min/week	60 min/week	15 weeks
	2017-2018	45 min/week	45 min/week	45 min/week	45 min/week	15 weeks
<b>Waterford</b>	2015-2016	60 min/week	80 min/week	80 min/week	80 min/week	28 weeks
	2016-2017	60 min/week	80 min/week	80 min/week	80 min/week	28 weeks
	2017-2018	60 min/week	80 min/week	80 min/week	45-60 min/week	28 weeks

\*Note: 20 minutes for On-Target students; and up to 60 minutes for High-Risk students





Evaluation and Training Institute  
100 Corporate Pointe, Suite 387  
Culver City, CA 90230  
[www.eticonsulting.org](http://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](mailto:jhobbs@eticonsulting.org)