Early Intervention Software Program Evaluation

2015-2016 Program Results

Submitted to the Utah State Board of Education
November, 2016

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

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

Table of Contents

Executive Summary

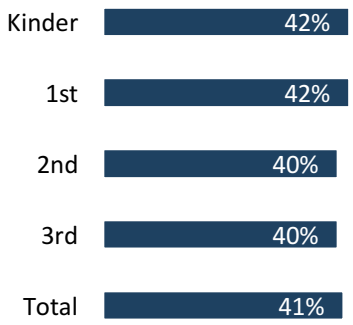
Introduction

In 2012, the Utah State Board of Education (USBE) funded the Early Intervention Software Program ("EISP") to support the growth of K-3 students' literacy. For the 2015-2016 program year, Local Education Agencies (LEAs) in Utah selected from among eight computer-based literacy programs which provide individualized instruction and are designed to supplement students' classroom learning. The Evaluation and Training Institute (ETI) was hired by the USBE to answer the following questions: "*Did students use the program as recommended by the software vendors?*" "*Did the program have an overall effect as measured by DIBELS across all vendors?*," and, "*Were there differences in treatment effects among vendors*?" Answers to these questions and our recommendations are provided in detail in the main report, and the highlights are presented in the executive summary sections below.

Program Impacts

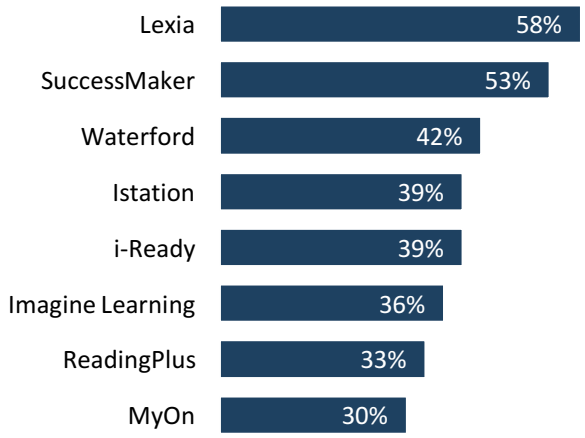
Did students use the program as recommended by software vendors?
Fewer than half of the students used the program at or above a relaxed version of the vendors' recommended weekly use, and less than a quarter of the LEA's met the minimum recommended use when calculated school-wide. These findings are based on a relaxed calculation of program fidelity that was developed to account for competing educational priorities and other factors that exist in schools: students who met 80 percent of the vendors recommended average weekly use were considered to have met program fidelity. Low fidelity of program use undermines the program's effectiveness.

Figure 1. Fidelity of Use by Grade



**Fidelity: Ave Min >=80% of vendors recommended minutes of use per week*

Figure 2. Fidelity of Use by Program



**Fidelity: Ave Min >=80% of vendors recommended minutes of use per week*

Did the program have an overall effect across all vendors?

Students who used the program had higher reading test scores as measured by DIBELS (on average) than students who did not use the program in kindergarten and second grade. In addition, the results were heavily dependent on how much time a student used the programs. In general, the effects of the program increased as program use increased. No statistically significant results above a small effect size (i.e. over .19) were found for first or third grade program students.

Table 1. Program-wide Treatment and Control Group Composite Score Means and Effect Sizes, by Level of Use

Usage Group	Kindergarten			1st Grade			2nd Grade			3rd Grade		
	Tr.	Cntrl	ES	Tr.	Cntrl	ES	Tr.	Cntrl	ES	Tr.	Cntrl	ES
Intent to Treat (lowest use)	N=8,272			N=11,709			N=2,874			N=2,521		
	148	140	.09	188	195	-.05	–	–	–	–	–	–
Relaxed Optimal	N=2,785			N=5,486			N=1,137			N=781		
	154	139	.21	_	_	_	161	154	.09	–	–	–
Optimal (highest use)	N=441			N=1,102			N=159			N=95		
	156	137	.36	–	–	–	161	135	.32	–	–	–

Note: A dash in a cell means that the treatment does not have a significant effect. ITT (lowest use): all students; Relaxed optimal (second highest use): students must meet at least 80% of vendors recommended dosage; Optimal (highest use): students must meet vendors' recs for at least 80% of the weeks used and use it for the minimum weeks recommended.

Were there differences in treatment effects among vendors?
We found differences among vendors, but the findings need to be reviewed with caution: in at least four cases, after all the data were cleaned, merged and split by grade we had very low sample sizes for the usage groups closest to vendors recommended minimum program use (MyOn; Istation, Reading Plus and Waterford). Small sample sizes made it difficult to detect small treatment effects if they were present. Five out of seven programs[1] had a positive impact on end-of-year composite scores as measured by DIBELS in kindergarten (Istation, Waterford, Imagine Learning, Core5, MyOn), with effect sizes ranging from .37 to 1.12. In addition, one program had a positive impact in first grade (Core5), and two programs had an impact in second grade (Imagine Learning; SuccessMaker).

Key Recommendations & Limitations

In the 2015-2016 EISP evaluation, we found that the program is very effective in kindergarten, and there were also signs of its effectiveness in second grade for intervention students, but low program use is a barrier to having conclusive findings about the program's benefits to students. In certain cases, our findings need to be interpreted with caution due to small sample sizes, and this is especially true when reviewing results for specific vendors.

Our two highest priority recommendations are to continue using the program with kindergarten students, and to focus on improving program implementation and fidelity of use across all grades. In addition, the USBE would benefit from improved data tracking at the program level, and to seek an increased understanding of how students are selected for program use. Specific recommendations for improving the program include:

- All vendors should provide monthly usage reports to schools to help them monitor their fidelity, and they should reach out to schools that are falling behind to offer additional support.
- An implementation evaluation should be sponsored by the state that focuses on how students are selected for the program, how schools and vendors monitor program use, challenges to continual program use, and best practices in using the program according to vendor recommendations.

[1] Reading Plus did not serve students in Kindergarten.

EISP Evaluation Report

Report Background and Purpose

The Utah State Board of Education (USBE) contracted with the Evaluation and Training Institute (ETI), an independent, non-profit research and consulting firm, to evaluate the Early Intervention Software Program ("EISP"). The EISP was designed to improve the literacy achievement of Utah students in Grades K-3 through computer-based software programs that adapt to each student's skill level and offer instruction tailored to meet their unique learning needs. During the 2015-2016 school year, eight vendors provided software and training to schools that opted into the program. The eight vendors were (in alphabetical order): Curriculum Associates (i-Ready), Imagine Learning, Istation, Pearson (SuccessMaker), Lexia Reading Core5® (Core5), MyOn, Reading Plus and Waterford.[2]

The evaluation had three objectives:

1. Evaluate how students used the software through a program implementation fidelity analysis;
2. Evaluate how the program effected students' literacy achievement as measured by DIBELS; and,
3. Document the evaluation findings, including recommendations, in a concise report format that would be used by USBE staff, legislators and other stakeholder groups.

The remainder of this report is organized by evaluation objectives, and we have structured the report to be "user friendly" to a wide audience, including researchers, professional educators, policy staff and non-technical reviewers. To help streamline our presentation of the results, each section includes a brief, non-technical overview of the research methods used, and we have used a question and answer format where appropriate to guide the reader. For more technical reviewers, we have included detailed information about the research methods, statistical sampling and analytic models in the appendices.

NOTE FOR READERS ON UPDATED REPORT RESULTS:

The USBE asked ETI to analyze a revised i-Ready program dataset that did not include time a student spent in non-program assessment activities. The findings presented in this report have since been updated to reflect ETIs analyses with the revised i-Ready data. Redoing the analyses with the revised i-Ready data had very little effect on the original findings.

[2] Core5, MyOn, and Reading Plus were added to the EISP in 2015-2016. A majority of schools using the new programs may not have had an opportunity to implement the programs for a full academic year.

Program Implementation Results

In this section we provide an overview of the 2015-2016 program enrollment numbers and measure students' program use against the software vendors' recommendations. Program enrollment numbers are based on all students with more than five minutes of use, while the program fidelity sample excludes students who used multiple software vendors from within the same school[3].

Program Enrollment

In 2015-2016, 388 schools and approximately 68,891 students used one of eight software programs. The most frequently used software program was Imagine Learning (184 schools), followed by Core5 (73), and i-Ready (55). Istation was used by four schools and is the vendor with the smallest number of students.

Table 2. Program Enrollment

Program	LEAs	Schools	Students
Istation	2	4	898
Waterford	20	55	7,609
i-Ready	17	55	12,015
Imagine Learning	41	184	23,798
SuccessMaker	10	23	3,679
Core5	17	73	17,346
Reading Plus	5	14	1,095
MyOn	8	16	2,451

Note: Schools could use multiple programs for different grades. There were 388 unique schools participating in EISP in 2015-2016.

Table 3 presents the distribution of student enrollment by grade and within each program. Overall, most of the programs (5 out of 8) had the highest number of participants in the first grade.

Table 3. Program Enrollment by Vendor and Grade

Program	Kinder	1st	2nd	3rd
Istation	N=181	272	257	188
	20%	30%	29%	21%
Waterford	3,119	2,882	1,351	257
	41%	38%	18%	3%
i-Ready	2,301	3,234	3,361	3,119
	19%	27%	28%	26%

[3] Imagine Learning cloud version program users were not included in the fidelity sample due to a system error when tracking student usage. Student counts were calculated after cleaning the data for duplicates and removing students with use below a certain threshold (e.g. weeks of use with five or fewer minutes of use).

Program	Kinder	1st	2nd	3rd
Imagine Learning	7,061 30%	9,197 39%	4,750 20%	2,790 12%
SuccessMaker	681 19%	1,078 29%	988 27%	932 25%
Core5	3,258 19%	4,011 23%	5,516 32%	4,561 26%
Reading Plus	--	21 2%	324 30%	750 68%
MyOn	104 4%	500 20%	924 38%	923 38%

Program Use

Recommended Dosage

Each vendor provided recommendations for using the software programs in order for it to have an impact on student achievement. **Table 4** displays vendors' recommended average use (in minutes) and the suggested minimum number of weeks by grade. The USBE informed LEAs about the software vendors' recommended minimum average weekly use and minimum weeks of use for the program to have optimal benefits to learning. In addition, both recommendations were used in the evaluation to create samples of students who had met relaxed weekly use and used the program for a recommended number of weeks (or more). Recommended weekly use ranged from 45 minutes to 80 minutes of use per week, and suggested weeks of use ranged from 12 to 28.

Table 4. Vendor Minimum Dosage Recommendations

Program	Kindergarten ALL Student	First Grade ALL students	Second Grade Intervention Students	Third Grade Intervention Students	Suggested Minimum Instructional Weeks
Imagine Learning	45 min/week	60 min/week	60 min/week	60 min/week	20 weeks
i-Ready	45 min/week	45 min/week	45 min/week	45 min/week	20 weeks
Istation	60 min/week	60 min/week	60 min/week	60 min/week	12 weeks
Core5 Reading	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
MyOn	45 min/week	45 min/week	45 min/week	45 min/week	20 weeks
Reading Plus	45 min/week	45 min/week	45 min/week	45 min/week	15 weeks
Successmaker	45 min/week	45 min/week	60 min/week	60 min/week	15 weeks
Waterford	60 min/week	80 min/week	80 min/week	80 min/week	28 weeks

*Core5 bases its usage recommendations on student performance, and students who score below grade level are assigned usage recommendations that are greater than those for students who score at or above grade level.

Fidelity of Minimum Recommended Use

How do we calculate program fidelity?

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

How well did students comply with vendors' dosage recommendations?

Overall, 31% of participating students met the fidelity of use requirements (based on average minutes of use per week) and also used the software for at least the minimum number of weeks suggested by vendors. The highest percentage of students who used the program for the suggested number of weeks and met fidelity were in the first grade (35%).

Table 5. Program-wide Usage Summary by Grade

Grade	Met Fidelity	Met Suggested Weeks of Use Recs	Fidelity and Weeks of Use Recs
K	N=6,640	8,736	5,014
	42%	54%	32%
1	8,488	13,144	7,057
	42%	65%	35%
2	6,510	8,989	5,088
	40%	55%	31%
3	5,001	5,143	2,994
	40%	41%	24%
Total	26,639	35,898	20,153
	41%	55%	31%

*Met Fidelity: Ave Min >=80% of vendors recommended minutes of use

[4] In previous EISP evaluations we created an estimate of average minutes of use by using the student's program start and end dates, while adjusting for school breaks and state testing.

Table 6 presents fidelity of use information for each program. SuccessMaker, Core5, and Istation had the highest overall usage, with 47%, 41%, and 36% of students using the programs as intended, respectively. MyOn had the lowest program use overall. A more detailed summary by program and grade can be found in **Appendix D**.

Table 6. Usage Summary: by Program

Program	Met Fidelity	Met Suggested Weeks of Use Recs	Met Fidelity and Weeks of Use Recs
Istation	N=346	713	320
	39%	79%	36%
Waterford	3,155	3,843	2,417
	42%	51%	32%
i-Ready	4,641	5,137	2,972
	39%	43%	25%
Imagine Learning	7,520	12,573	6,176
	36%	61%	30%
SuccessMaker	1,788	2,292	1,590
	53%	68%	47%
Core5	10,077	9,275	7,055
	58%	54%	41%
Reading Plus	303	486	230
	33%	53%	25%
MyOn	741	522	289
	30%	21%	12%

*Met Fidelity: Ave Min >=80% of vendors recommended minutes of use

How do we report and calculate school level fidelity?

A school or district met fidelity of use if at least **80%** of their students' average minutes of use were greater than or equal to **80%** of the vendor's average minutes of use recommendations. A separate fidelity of use report is provided to the state with specific information on usage at the school level, and to identify the schools in danger of losing funding if they do not increase their students' fidelity within two years (2015-2016 to 2016-2017). Preliminary results show less than a quarter of schools met program fidelity.

Literacy Achievement Results

We evaluated the EISP's effectiveness by comparing the literacy achievement of groups of students who used the program to groups of students who did not use the program. We measured literacy achievement using the DIBELS Next test, which was administered in schools throughout the state in grades K-3. The DIBELS Next measures are used throughout Utah, and are strong predictors of future reading achievement (see **Appendix B** for more information about the DIBELS Next). Our evaluation results are presented in two sections: 1) Program-wide impacts, and 2) Individual vendor impacts. The program-wide analyses measure the impact of the EISP across all eight software programs, providing stakeholders with a big-picture view of how the program performs. In the individual vendor impacts section, we explore the relative impacts each program vendor had on literacy achievement. Important details about our methods and how the statistical analyses were performed are included below and in each section.

How did we create our analytic samples?

We collected data from fifteen different sources to create our master dataset for the EISP analyses. A summary of our cleaning procedures and analyses samples can be found in **Appendix A and C.** The data sources included: eight program vendors, who provided us with usage information for each student who used their programs; DIBELS data from two online data entry reporting systems (V-port and AMPLIFY) and four districts; and student information system (SIS) demographic data provided by the USBE. We cleaned and reviewed each data file before creating the master dataset, which we then used to create the matched treatment and control group samples. In second and third grade, the program was designed to target intervention students only, and our second and third grade samples included participants who were below grade level at the beginning of the year. Our samples consisted of a program-wide and individual vendor treatment and control group matched samples for different levels of use.

We used Coarsened Exact Matching (CEM, Iacus et al., 2008) to match groups of students using the program ("treatment group") to groups of students who did not use the program ("control group"). The students were matched on data from the beginning of the school year, and across several important characteristics (grade, achievement level, gender, race, poverty status, and other criteria). CEM minimized differences between the two groups prior to enrollment in the program *(see **Appendix D** for more information on CEM and how it was used to match students)*.

ETI created three usage groups to study the effects of increased program use on student test scores. Each program vendor provided schools with a recommendation for how much time the student should use the program before benefits are observed. This minimum use recommendation is an important predictor of literacy achievement, and we wanted to determine how student use characteristics effect their outcomes. In addition to the minimum minutes per week recommended by vendors, we also determined that the number of weeks a student used the program was also an important predictor of later test scores. We operationally defined the combination of weekly use and weeks of

use as “program dosage”. We created three independent program-wide samples to determine the effects of program dosage on students’ achievement:

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

What statistics do we provide in our results?

The data were analyzed using STATA (v. 14) and SPSS (v. 22), and the statistical models used were consistent across vendors. Where appropriate, we provided mean scores for our treatment and control groups, which are meaningful when comparing treatment and control groups from the same sample. We also provided effect sizes (ES, based on Cohen’s Delta, or “d”; see **Appendix E** for more information on how it was calculated) to help readers understand the magnitude of treatment effects. Effect sizes enabled us to provide a standardized scale to compare results based on different samples. Cohen (1998) categorizes effect sizes as small (0.2), medium (0.5), and large (0.8). ETI reported effect sizes below .2 (small) when they were statistically significant, but we also noted that the real-world effects most likely were not substantive and would not improve learning outcomes.

Program-Wide Impacts

How did we study the program-wide impacts?
We studied the program-wide impacts by comparing a sample of treatment group students drawn from all vendors to a matched sample of control students. A two-level random intercept statistical model with school as the level-2 predictor was used to predict student outcomes. We determined that using a two-level regression model (also known as a "hierarchical linear regression model", or HLM) allowed us to study the differences in treatment and control group student outcomes, while controlling for other student-level predictors, and, at the school-level, controlling for Title 1 status. In general, non-significant predictors were removed from statistical models to increase the variance we could explain with the significant predictors of achievement.

What were the program-wide treatment effects?
Dosage is the most important determinate in program-wide treatment effects. As seen in **Figures 3 - 4**, program-wide effects on DIBELS Next end-of-year (EOY) composite scores increase with dosage, and the more a student uses the program the better his/her EOY outcomes. These treatment effects are not seen across all grades, however, and the results show substantive effects in kindergarten and second grade:

- In kindergarten the treatment effects more than double when you move from ITT (lowest dosage) to ROPT (second highest dosage), and quadruple when you go from ITT to OPTI (the highest dosage) usage groups.
- In second grade students in the OPTI group (the highest dosage) have over a three-fold increase in the treatment effect size when compared to the lowest dosage group (ITT).

Students with the highest average program dosage in kindergarten and second grade had the highest treatment effect sizes overall as measured by their average DIBELS Next Composite score (.36 and .32, respectively).In addition, kindergarten students with the lowest average program dosage (ITT) and second grade students with the middle highest average program dosage (ROPT sample) had treatment effects , but the effects were too low to be categorized as a "small treatment effect" (.09, respectively). Students in Title 1 schools had results that were similar to non-Title 1 schools (see **Appendix E** for Title 1 school results).

Figure 3. Kindergarten: Means of EOY Composite for Matched Treatment and Control, by Usage Group

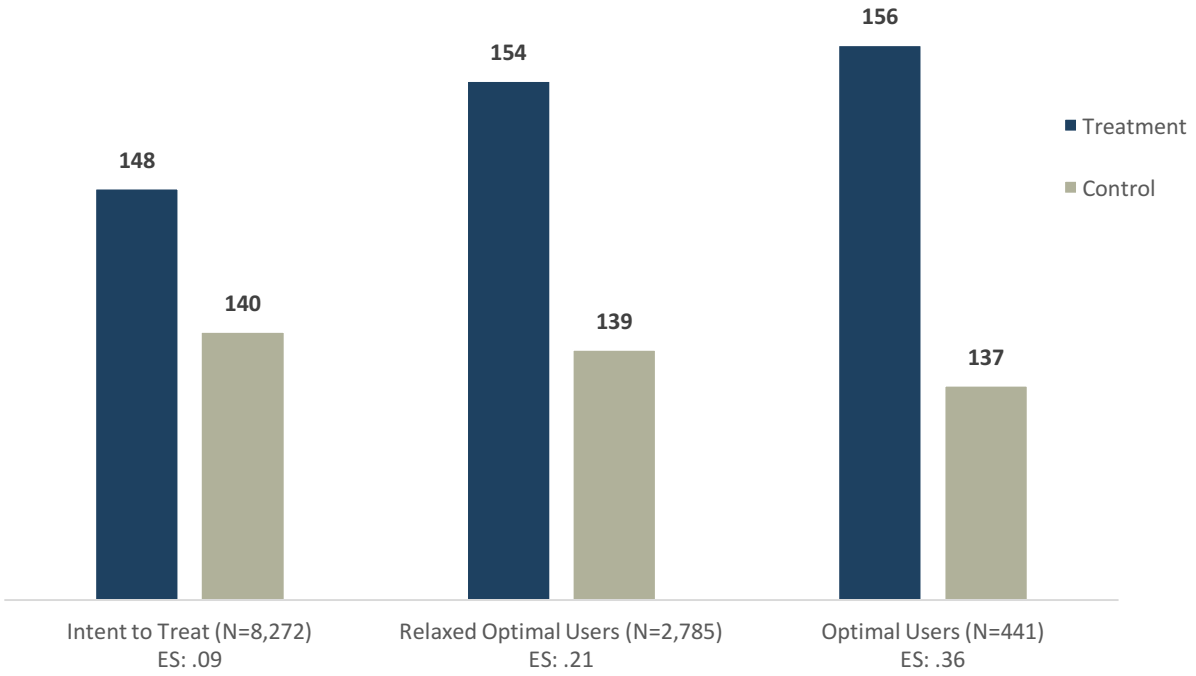
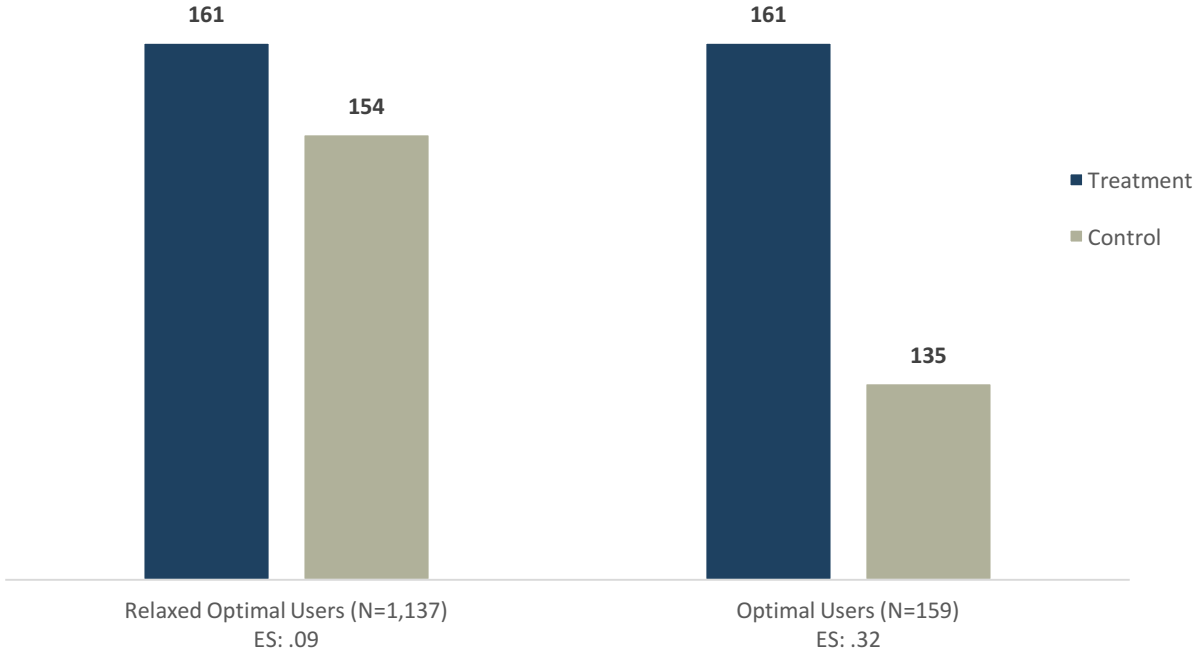


Figure 4. Second Grade: Means of EOY Composite for Matched Treatment and Control, by Usage Group



The program-wide treatment effects are smaller for individual DIBELS Next scales. For example, we found small treatment effects for Letter Naming Fluency (LNF) and Nonsense Word Fluency (correct letter sounds; NWF-CLS; results are shown in **Table 7**). Program students perform better than comparison students across all kindergarten literacy subscales, with effect sizes ranging from .08-.22. Treatment effects are present in first and second grade for certain subscales, but they are below a small effect size (below .2). We did not find any treatment effects for third grade intervention students.

Table 7. Predicted Means of EOY DIBELS Scales for Matched Treatment and Control, Program-Wide, ROPT sample

		Kindergarten			1st Grade			2nd Grade		
		Tr.	Cntrl	ES	Tr.	Cntrl	ES	Tr.	Cntrl	ES
Composite Score		N=2,785			N=5,486			1,137		
		154	139	.21	–	–	–	161	154	.09
First Sound Fluency (FSF)	K	37	35	.08						
Letter Naming Fluency (LNF)	K-1	54	49	.22						
Phoneme Segmentation Fluency (PSF)	K-1	53	50	.12						
Nonsense Word Fluency-CLS	K-2	47	41	.18	86	83	.06			
Nonsense Word Fluency-WWR	K-2	7.7	6.9	.08	–	–	–	–	–	–
Oral Reading Fluency	1-6	–	–	–	68	66	.08	56	53	.16
DAZE	3-6	–	–	–	–	–	–	–	–	–

Note: Program students are matched to comparison students using CEM for each all vendors and then matched for program usage. Predicted means are then reported based on the coefficients from a multilevel regression model. Level 1 covariates are sex, Hispanic, special education, and BOY Composite score. School Title I status is modeled as a Level 2 variable. A dash in the cell indicates that the program did not significantly effect the score for that subscale and grade, or that fewer than 10 observations were available. N=9,998 each for program and comparison students. There were no significant effects in third grade.

Individual Program Impacts

How did we study individual vendor impacts?

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

1. We conducted a usage effects analysis, and measured the relationship between students' program use and DIBELS composite scores for an ITT treatment and matched control group sample; and,
2. We conducted a between group mean score analysis for treatment and control group students in each vendors' ROPT sample (and used the ITT sample when ROPT samples were too small to detect program effects for certain programs and grades).

Usage Effects Analysis

How did we study the effects of program usage on literacy scores?
We regressed weeks of use on outcome scores for each vendor's matched sample of treatment and control students using the following Ordinary Least Squares equation:

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

The unstandardized regression coefficient (*b* in the above equation) represents the relationship between weeks of program use and learning outcomes: the coefficient represents a unit change in Composite score for every additional week of use. It should be noted that this comparison does not include control students, so even when a statistically significant relationship between weeks of use and literacy score is found, the control students could also be improving at the same rate (however, for obvious reasons, control students who did not use the program cannot be included). This analysis allows us to see the relative effects each vendor had within their sample of program students.

How do weeks of use effect student outcomes?
Table 8 shows the unstandardized regression coefficients when we regressed weeks using the program on EOY Composite scores. For every additional week of use, the end-of-year composite score increased by an average of .22 – 3.2 points in kindergarten, and .44 - 1.29 points in first grade for six of the programs. Three programs had significant and positive effects in second grade (.22 - 1.01 points), while only one program had significant, positive effects in third grade (2.26 points).

Table 8. Weeks of Use OLS Regression Coefficients

	Kindergarten	1st Grade	2nd Grade	3rd Grade
Istation	1.89	–	–	2.26
Waterford	.79	–	–	–
i-Ready	-.66	.45	.22	-1
Imagine Learning	.65	.62	–	–
SuccessMaker	–	–	1.01	–
Core5	.22	.78	.86	–
Reading Plus*	N/A	N/A	–	–
MyOn	3.2	–	–	–

Note: Model covariates are gender, Hispanic, special education, school Title I status, and BOY Composite score. A dash in a cell means that the treatment is not a significant effect for the model.
*The Reading Plus program targeted 2nd and 3rd grade students.

Treatment and Control Group Comparison of Literacy Composite Scores

How did we study differences between treatment and control group outcomes among vendors?

Similar to our program-wide approach, we created a matched control group for each program vendor using CEM (**see Appendix D**). We created eight matched samples, one for each vendor, which allowed us to have tightly matched control groups for each program vendor. We studied the differences between vendor program students and non-program students using an ordinary least squares (OLS) regression analysis to predict the EOY composite scores by group, using the same regression equation presented in the above section "Usage Effects Analysis." We controlled for a student's BOY literacy achievement, gender, ethnicity, poverty status, and special education status in our regression model. The OLS regression model was used because we did not have an adequate sample size (N) for each vendor to conduct a two-level analysis. Similarly, our vendor-specific samples were not large enough to study the program effects of students who met vendors' exact usage recommendations (e.g. optimal usage), and we studied a subset of students who met a relaxed version of vendors' recommendations instead. We used the ITT sample (all students, regardless of use) when we had a low ROPT sample, and we wanted to see if any effects could be found with a larger sample of students.

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

Table 8 presents the OLS regression results for each program and grade. A majority of programs had a positive impact on students in kindergarten (**Table 9**), following the same trend as depicted in other analyses. In first grade, only one program had a significant, small positive effect, while two programs produced positive results in second grade. There were no positive impacts for third grade students.

Table 9. Predicted Means of EOY Composite for Matched Treatment and Control, by Vendor

	Kindergarten			1st Grade			2nd Grade			3rd Grade		
	Tr.	Cntrl	ES	Tr.	Cntrl	ES	Tr.	Cntrl	ES	Tr.	Cntrl	ES
Istation	170	128	1.12	–	–	–	–	–	–	–	–	–
Waterford	149	134	.42	–	–	–	–	–	–	–	–	–
i-Ready	–	–	–	–	–	–	–	–	–	–	–	–
Imagine Learning	155	134	.52	–	–	–	162	143	.33	–	–	–
SuccessMaker	–	–	–	–	–	–	197	167	.52	–	–	–
Core5	159	143	.43	214	207	.11	–	–	–	–	–	–
Reading Plus		N/A			N/A		–	–	–	–	–	–
MyOn	–	–	.37**	–	–	–	–	–	–	–	–	–

Note: Model covariates are gender, Hispanic, special education, school Title I status, and BOY Composite score. A dash in a cell means that the treatment is not a significant effect for the model.

**MyOn had a very small sample size in kindergarten, and there were a large number of cases missing BOY and EOY DIBELS data. ETI used multiple imputation (SPSS, v. 23; details about imputation algorithms can be found at: https://www.ibm.com/support/knowledgecenter/en/SSLVMB) to generate possible values for missing values (for MyOn only), thus creating several "complete" sets of data missing values. The effect size given is the average for the imputed data.

Impacts on Literacy Domains

The DIBELS Next composite score is a strong predictor of future reading achievement, but the instrument also captures scores for specific literacy skills or "domains" that combine to form the bigger construct of literacy.

How did we study the individual program impacts on specific domains of literacy?

Some vendors had small sample sizes with respect to certain domains, and we could not use OLS regression models to study the differences in treatment and control students. Instead, we conducted a between groups mean score test (a "t-test"), and calculated effect sizes (where significant differences were found) using Cohens D. We used the following formula to test mean score differences:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{S_1}^2}{N_1} + \frac{{S_2}^2}{N_2}}}$$

How did individual vendors have an impact on specific domains of literacy?

Kindergarten

Table 10 presents effect sizes for each vendor. Almost all the vendors who served kindergarten students (5 out of 7) produced positive effects for the DIBELS subscales measuring phonemic awareness: First Sound Fluency (FSF) and Phoneme Segmentation Fluency (PSF). Letter Naming Fluency (LNF), which is an indicator of risk, produced positive effects among five of the seven vendors. Two programs produced significant, positive effects for two Nonsense Word Fluency (NWF) subscales, which assesses alphabetic principles and basic phonics, while two other programs produced positive effects for one of the two NWF subscales. The effect sizes ranged from small to medium for measures of alphabetic principles and phonics (.15 - .51) and phonemic awareness (.09 - .65).

Table 10. Kindergarten DIBELS Subscale Effect Sizes, T-test

Scale	Imagine Learning	i-Ready (N=244)	Istation* (N=115)	Core5 (N=724)	MyOn* (N=49)	SuccessMaker (N=158)	Waterford (N=793)
FSF	.31	-	.65	.09	-	.22	.15
LNF	.38	-	.29	.30	.36*	-	.25
PSF	.43	-	.95	.28	.50*	.41	.30
NWF-CLS	.30	-	-	.31	.51*	-	.24
NWF-WWR	.23	-	-	.15	-	-	-

Note: We used an "*" to identify effect sizes generated from the ITT group. All other effect sizes represent the ROPT group. A dash in a cell means that the treatment is not a significant effect for the model. The Reading Plus program was used by students in upper grades and was not included in this table.

First Grade

In first grade, four out of seven programs produced small effects in Nonsense Word Fluency, while none of the programs had an impact on Oral Reading Fluency.

Table 11. First Grade DIBELS Subscale Effect Sizes, T-test

Scale	Imagine Learning (N=1998)	i-Ready (N=437)	Istation (N=106)	Core5 (N=1758)	MyOn* (N=260)	SuccessMaker (N=559)	Waterford (N=615)
NWF-CLS	.09	-	.45	.22	-	-	-
NWF-WWR	.10	-	.34	.16	-	-	.10
DORF Fluency	-	-	-	-	-	-	-

Note: We used an "*" to identify effect sizes generated from the ITT group. All other effect sizes represent the ROPT group. A dash in a cell means that the treatment is not a significant effect for the model. The Reading Plus program was used by students in upper grades and was not included in this table.

Second Grade

In second grade two of the eight programs produced significant, positive effects in Oral Reading Fluency, which is a measure of reading comprehension. Effect sizes were considered small, at .18 - .29.

Table 12. 2nd Grade DIBELS Subscale Effect Sizes, T-test

Scale	Imagine Learning (N=321)	SuccessMaker (N=83)
DORF Fluency	.18	.29

Note: ROPT sample

Third Grade

In third grade, only one program had a statistically significant impact on a DIBELS literacy subscale. Treatment students did better than control students in Oral Reading Fluency, with a medium effect size of .53.

Table 13. 3rd Grade DIBELS Subscale Effect Sizes, T-test

Scale	Reading Plus* (N=58)
DORF Fluency	.53
DAZE	-

Note: ITT sample.

Summary, Limitations and Recommendations

ETI evaluated two facets of the EISP: program implementation and its impacts on student learning. Like all research, our evaluation has limitations, and they are important to understand when reviewing the results. In this section we give a summary of the evaluation results, followed by a description of the limitations of our research design and our recommendations to improve the program and suggestions for future evaluations.

Program Implementation

Not enough students are using the program as the vendors intended. Only slightly over half of the program students are meeting a relaxed calculation of vendors' recommended minimum weekly usage (i.e. 80% of recommended average weekly use), a fact that makes it difficult to evaluate the program's effectiveness in improving literacy achievement. When program usage was analyzed by school, however, the results show even lower levels of fidelity of implementation: we found that less than a quarter of the schools met a relaxed version of the recommended minimum program use. We observed these results even after we adjusted ("relaxed") calculating minimum average weekly use to account for competing educational priorities in schools.

Program Impacts on Literacy Achievement

We studied program impacts through two lenses: 1) program-wide (analyzing all vendors together) and each individual vendor. The program-wide results showed that the EISP had a small to medium effect on overall reading proficiency in kindergarten and second grade, but not for first or third grade program students. The strongest effects were for a small group of high dosage users in kindergarten (N=441; ES: .36) and in second grade (N=159; ES: .32). In other words, when children are using the program as intended, the program has positive effects across half of the grades.

Six out of seven programs had an impact on more than two literacy subscales in kindergarten, four program vendors had an impact on a measure of alphabetic principles and knowledge (NWF) in first grade, two vendors had an impact on Oral Reading Fluency in second grade, and one program had an impact in third grade. These results need to be interpreted with caution, however, because when we constrained our analytic samples for each vendor to students who used the program at higher levels (and then by grade), in some cases we had very small sample sizes that made it hard to detect small treatment effects. In addition, our findings show how well a sample of students perform when they use the program close to vendors' recommendations. The program impacts may be more pronounced when students use it as intended; however, it was not possible to conduct this analyses for individual vendors due to a combination of low overall student fidelity and other factors that reduced the samples.

Evaluation Limitations

Two limitations related to this evaluation are associated with quasi-experimental research designs, and our reliance on secondary data that we used to match program students to state student demographic information and DIBELS Next test scores.

Quasi-Experimental Research Designs. It was not possible to randomly assign students to program and non-program groups, which is a research method that minimizes pre-existing differences between program (treatment) and non-program (control) students. We used a quasi-experimental research design (QED), which is indicated when naturally occurring groups of program and non-program students exist, and there is not an opportunity to randomize to either group. These naturally occurring groups could have had pre-existing differences that are related to the results, such as extracurricular support, parental factors, and other things that influence learning achievement. We attempted to control for pre-existing group differences using a statistical matching process (Coarsened Exact Matching), where non-program students are matched to program students at the beginning of the school year using student achievement and social-demographic data. The matching process created balanced groups at the beginning of the year, however, there may have been variables that we could not measure that affected student learning. Without random-assignment to groups, there are many variables we could not control for, so the results must be seen as probable outcomes, but there may be other variables influencing them.

Secondary Data. "Secondary data" are data collected by outside sources and transferred to the evaluators. The secondary data in this study were from program vendors, the state, and DIBELS Next databases, and some limitations arose from its use. First, a majority of our DIBELS Next data were collected and stored through the V-Port and AMPLIFY systems. These systems offer efficient transfer of DIBELS Next test scores, but they are limited and not all LEA's use them, and therefore we only had scores for a subgroup of program students. In some cases, we collected additional DIBELS data for programs with small populations of students (we did this to have sufficient sample sizes for these vendors to be included in our analyses); however, we did not request additional DIBELS data for the programs with larger populations. Other factors that affected the sizes of our samples included: students who used more than one software program, duplicate IDs, and incomplete DIBELS scores, other missing or incorrect data (such as student IDs) among other factors (see Appendix C for more information).

The analytic samples developed through the merging of limited secondary data files, could have affected the literacy achievement results, but would not have affected the program fidelity results (because fidelity did not depend on merging multiple data sets). These results are based on a sample of students that were successfully merged across several data sets and who had complete data, so the results are tightly linked to our samples and their generalizability is limited.

Recommendations

The recommendations we provide below are based on our understanding of the findings and tempered by the evaluation limitations:

Program Use Recommendations

- We have found the program to be effective in kindergarten, and to a lesser extent, in second grade, but our ability to determine the program's effectiveness was hampered by low use. We recommend that the state find ways to work with the program vendors to increase program use to levels consistent with vendors recommended use.
- We recommend that program vendors provide monthly usage reports to schools and notify school staff when they are falling behind.
- Vendors should continue providing training to LEAs at the beginning of the school year that clearly articulates the importance of following vendors' usage guidelines. We recommend that vendors develop a triage strategy to assist LEAs that fall behind on usage (possibly targeted trainings).
- Vendors with smaller numbers of enrolled students need to be even more vigilant in encouraging fidelity of use so that it is possible to determine the effects they have on students' learning. We recommend that vendors with small enrollment numbers work with the evaluators and/or the state throughout the year to review student use patterns.

Program Data Recommendations

- The state should emphasize to LEAs the importance of capturing complete and accurate SSID data for students enrolled in the program.
- We recommend that program vendors provide the evaluators with SSIDs for students who use their software for other state-wide initiatives (such as ELL interventions), so we can control for the effects due to multiple-programs being used by some students.
- This year was the first year vendors provided a breakdown of students' minutes of use for each week of software use. While this process went smoothly for most vendors, some vendors had more difficulty than others. We recommend that vendors work to review and improve their internal usage tracking mechanisms early-on in the process so the end-of-year data file preparation runs smoothly.

Future Evaluation Recommendations

It is clear that schools are struggling to use the software as intended, and understanding the reasons behind a teacher or schools' choice to use the program may be an important factor in determining the effectiveness of the program. Increased understanding of the barriers to implementation and the intricacies of how and when the programs are used and with whom is needed to fully understand which conditions lead to positive impacts. We recommend that the evaluation adopts a pilot study designed to examine the different constraints placed on elementary students and teachers, and monitor how these factors relate to implementation at the ground level. The pilot study will need to have a direct link to teachers' reasons for using the software, and will depend on teacher feedback as an important measure of software use, program efficacy, and how the program meshes with existing curricular requirements.

References

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

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

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

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

Appendix A. Student Program Use

Table 1 presents a comprehensive summary of usage for each vendor and grade. In addition to fidelity of use, the table includes usage frequencies, such as average minutes and number of weeks of use, the percentage of students who met the vendor's minimum weeks of use requirements, and the percent of students who reached at least 80% of the recommended average weekly use requirements and minimum weeks of use ("met fidelity"). The results reported here are based on all the students enrolled in the software prior to cleaning and merging the data for the outcome analyses.

Table 1. Usage Summary: by Grade and Program

	Grade	Ave Use Recs (Min)	Ave Weekly Use (Min)	Weeks of Use Recs	Ave Weeks of Use	% Met Fidelity	% Met Weeks of Use Recs	% Met Fidelity and Weeks of Use Recs[5]
Istation	0	60	41	12	22	34%	85%	32%
	1	60	53	12	26	68%	93%	67%
	2	60	38	12	20	23%	76%	22%
	3	60	39	12	15	22%	61%	14%
	Total		**44**	**12**		**39%**	**79%**	**36%**
Waterford	0	60	48	28	25	53%	58%	41%
	1	80	57	28	26	39%	57%	31%
	2	80	48	28	19	27%	30%	17%
	3	80	51	28	15	25%	17%	7%
	Total		**51**	**28**		**42%**	**51%**	**32%**
i-Ready	0	45	32	20	18	31%	46%	22%
	1	45	38	20	20	42%	55%	32%
	2	45	34	20	18	38%	46%	27%
	3	45	34	20	14	41%	25%	17%
	Total		**35**	**20**		**39%**	**43%**	**25%**
Imagine Learning	0	45	37	20	21	42%	60%	35%
	1	60	44	20	23	35%	71%	30%
	2	60	42	20	20	34%	54%	27%
	3	60	41	20	17	29%	37%	20%
	Total		**41**	**20**		**36%**	**61%**	**30%**
Success-Maker	0	45	38	15	17	55%	63%	42%
	1	45	49	15	22	70%	72%	66%
	2	60	43	15	19	40%	68%	35%
	3	60	46	15	19	45%	67%	40%
	Total		**45**	**15**		**53%**	**68%**	**47%**

[5] Students must have met at least 80% of the vendors' average requirements for minutes per week and total weeks of use.

	Grade	Ave Use Recs (Min)	Ave Weekly Use (Min)	Weeks of Use Recs	Ave Weeks of Use	% Met Fidelity	% Met Weeks of Use Recs	% Met Fidelity and Weeks of Use Recs[5]
Core5*	0		46	20	16	48%	39%	28%
	1		55	20	20	66%	64%	51%
	2	20 - 60	51	20	21	60%	63%	48%
	3		48	20	18	57%	45%	33%
	Total		**50**	**20**		**58%**	**54%**	**41%**
Reading-Plus	1	45-75	40	15	15	42%	42%	26%
	2	45-75	34	15	15	37%	43%	31%
	3	45-75	33	15	16	32%	57%	23%
	Total		**33**	**15**		**33%**	**53%**	**25%**
MyOn	0	45-60	20	20	4	6%	0%	0%
	1	45-60	26	20	11	17%	15%	5%
	2	45-60	31	20	14	28%	25%	14%
	3	45-60	38	20	13	43%	23%	15%
	Total		**32**	**20**		**30%**	**21%**	**12%**

*Core5 bases its usage recommendations on student performance, and students who score below grade level are assigned usage recommendations that are greater than those for students who score at or above grade level

Appendix B: DIBELS Next

The Dynamic Indicators of Basic Early Literacy skills (DIBELS Next) is a statewide assessment used to measure students acquisition of early literacy skills at the beginning, middle, and end of the academic year. The online data entry systems, AMPLIFY and V-port[6], were used by a majority of LEAs throughout the state to capture DIBELS Next data. In order to increase the sample size of some of the less frequently used software programs we also requested and received DIBELS Next data from the following districts: Cache, Canyons, Garfield, and Ogden.

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 1** provides a summary of the DIBELS subscales used in our analyses.

Table 1. 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

[6] 2015-2016 was the first year in which V-port was used by districts in the state to house the DIBELS Next data.

DIBELS Next Scale	Description	Early Literacy Construct	Grade
Daze (DAZE)	Students read a passage with every seventh word replaced by a box containing the correct word and two distractor words. Assesses student's ability to construct meaning from text using word recognition skills, background information and prior knowledge, and familiarity with linguistic properties (e.g., syntax, morphology).	Reading Comprehension	3-6

**DIBELS NEXT Manual: http://wenatchee.innersync.com/assessment/documents/dibelsnext_assessmentmanual.pdf*

Appendix C. Data Processing and Merge Summary

We collected data from fifteen different sources to create our master dataset for the EISP analyses. The data sources included: eight program vendors, who provided us with usage information for each student who used their programs, two online data entry reporting systems (V-port and AMPLIFY) and four districts which provided us with DIBELS data. In addition, the USBE provided us with student information system (SIS) demographic data. We cleaned and reviewed each data file before creating our master dataset, which we then used to create our matched treatment and control group samples for the outcome analyses. In this Appendix we provide a detailed summary of our cleaning process throughout the creation of the final data files in our analyses.

Software Program Data

Our cleaning process for the program vendor data files included making sure all program schools that received licenses were included in the data as long as they used the program in 2015-2016, removing students outside of grades K-3, identifying and processing duplicate IDs within vendors' data, and formatting variables as needed, among other steps. We looked for duplicate IDs within each vendors' data, deleting cases that were the same student with different usage reported, and keeping any unique cases after removing exact replicas. We created new variables to use in our analyses, such as total weeks of use, average minutes of use, and other program fidelity measures.

We considered student records with five minutes or less of use reported for a week to be an error. We counted these weeks as having zero usage and updated the total minutes to reflect this change. After we cleaned and processed each program's data, we combined all eight vendors' data (N=69,305)[7]. We then identified and removed duplicate IDs between vendors[8] and any IDs that did not comply with the state student ID (SSID) format (N=65,130).

Note: The implementation findings were based on the program data prior to merging data files, and students with incorrect IDs were not excluded.

SIS Data

We reviewed the SIS data provided by the USBE to ensure that all LEAs who were listed as 2015-2016 participants were included in the data. We created additional variables as needed and processed duplicate IDs.

DIBELS Data

After we received the DIBELS data we made sure the requested variables were included in each file. We formatted the district DIBELS file to correspond to the format of the online versions to create a combined file. Our combined DIBELS file consisted of

[7] Number of cases after missing IDs were deleted.
[8] These IDs were also deleted from our pool of potential control students.

181,736 cases. 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 with 177,368 cases.

Master Merged Data File

We merged the SIS data from the USBE into our master DIBELS file (177,368) and were left with 168,322 cases[9]. Next, we merged our master vendor data (65,130) into the DIBELS and SIS data. Our final merged file consisted of 53,535 treatment students and 114,787 potential comparison students.

Table 1. Overview of Data Cleaning Process by Program

	Vendor Data (unique cases)	Vendor Data: Across Vendor Dups Removed		Dups & Invalid SSID Format Removed		Vendor Data Merged to DIBELS & SIS Data	
Program	N	N	% of Total	N	% of Total	N	% of Total
Istation	898	862	96%	862	96%	844	94%
Waterford	7,639	7,313	96%	6,998	92%	5,185	68%
i-Ready	13,503	13,103	97%	12,430	92%	8,710	65%
Imagine Learning	23,798	22,973	97%	22,967	97%	18,914	79%
Success-Maker	3,679	3,126	85%	3,096	84%	2,957	80%
Core5	17,363	16,323	94%	16,252	94%	15,234	88%
Reading-Plus	1,095	856	78%	819	75%	428	39%
MyOn	2,451	2,189	89%	1,961	80%	1,263	52%
Total	70,426	66,745	95%	65,385	93%	53,535	76%

Cleaning the Master Merged File

EISP has two separate purposes in lower and upper grade levels:

- K-1: targets all students; and
- 2-3: used as an intervention.

This year, we used the BOY benchmark levels to exclude students who started the year at or above grade level in our outcome analyses. While this reduced our sample size for these grades, our goal was to determine the effectiveness of the programs on the students the original legislation intended to serve.

In addition, a large number of English Language Learners (ELL) students throughout the state of Utah use Imagine Learning as part of a separate state-wide initiative, and participating students were not tracked using SSIDs. In order to prevent cross-contamination of our treatment or control group sample, we removed all students

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

identified as ELL using the USBE demographic data. The Imagine Learning program reported a system error in capturing usage data for its cloud version program users, and these students were also removed from our sample of Imagine Learning program users. Finally, cases were removed with missing DIBELS composite scores (at BOY or EOY). Our final matched file, after all the processing was complete, consisted of approximately 25, 436 treatment students.

Table 2. Overview of Cases Lost by Program

	Cases Lost after Cleaning IDs (format/dups)		Cases Lost after Merge		Est. lost after Deleting Cases[10]		Est. Total Cases Lost	
Program	N	%	N	%	N	%	N	%
Istation	36	4%	18	2%	248	29%	302	34%
Waterford	641	8%	1,813	26%	604	12%	3,058	40%
i-Ready	1,073	8%	3,720	30%	1,362	16%	6,155	46%
Imagine Learning	831	3%	4,053	18%	4,156	22%	9,040	38%
Success-Maker	583	16%	139	4%	241	8%	963	26%
Core5	1,111	6%	1,018	6%	2,169	14%	4,298	25%
Reading-Plus	276	25%	391[11]	48%	55	13%	722	66%
MyOn	490	20%	698	36%	124	10%	1,312	54%
Total	5,041	7%	11,850	18%	8,959	17%	25,850	37%

[10] We removed IL cloud users; ELL students; duplicate cases across AMP/V-port/District DIBELS data, etc.

[11] Almost half of students using RP come from Uintah District, which did not use DIBELS to measure student outcomes.

Appendix D: Methods and Sample

EISP was designed as an intervention for second and third grade students, and we only included students with beginning of year scores that were below grade level in our outcome 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 scrubbed the data to exclude students who may have used multiple software programs. This included excluding students who were identified as English Language Learners (ELL), due to a separate state-wide initiative that offers ELL students one of the participating vendor's programs.

For the program-wide analyses, we created three separate matched treatment and control groups based on levels of program dosage. In order of lowest to highest program dosage, our final program-wide samples included (approximately): 25,500 treatment students ("intent to treat"); 10,000 treatment students (relaxed optimal use: ROPT) and 2,000 treatment students (optimal use). We created a new matched treatment and control group sample for each program vendor and usage group for which we had a sufficient sample size.

Our analyses methods varied based on the sample size for the different types of analyses:

- **Implementation findings** – this sample included all students who used the programs in K-3, prior to cleaning invalid SSIDs, ELL students, identifying second and third grade intervention students, and merging files. We used descriptive statistics to show how the program was used by program participants.
- **Program-wide analyses** – as our largest sample, we used a two-level regression model ("hierarchical linear regression model", or HLM) to compare treatment students to control students on DIBELS Next composite scores and literacy subscales.
- **Individual program impacts** – our sample size varied by vendor, and we used an Ordinary Least Squares (OLS) regression model to compare treatment students to control students on DIBELS Next composite scores and t-tests for DIBELS literacy subscales.

Coarsened Exact Matching (CEM) Method

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

Tables 1 – 5 present the characteristics of the treatment group for each matched sample used in our analyses. As a result of our CEM procedure, our matched controls are the same. *Note, tables below are estimates of our sample.

Program-wide Sample

Table 1. Program-Wide Sample by Grade, Intent to treat

Grade	N	Female		Hispanic		African American		Caucasian		SPED		Low-Income		BOY Composite
K	8272	3977	48%	788	10%	78	1%	7016	85%	704	9%	3295	40%	36
1	11709	5787	49%	876	7%	71	1%	10,277	88%	1058	9%	3339	29%	125
2	2874	1292	45%	326	11%	46	2%	2384	83%	677	24%	1060	37%	76
3	2521	1139	45%	304	12%	18	1%	2093	83%	731	29%	881	35%	127

Note: Data presented are for the treatment group. The treatment and control group are equivalent.

Table 2. Program-Wide Sample by Grade, Relaxed Optimal Use

Grade	N	Female		Hispanic		African American		Caucasian		SPED		Low-Income		BOY Composite
K	2785	1314	47%	281	10%	28	1%	2356	85%	260	9%	1318	47%	36
1	5486	2713	49%	418	8%	36	1%	4816	88%	512	9%	2203	40%	127
2	1137	493	43%	129	11%	15	1%	932	82%	289	25%	516	45%	73
3	781	343	44%	120	15%	8	1%	619	79%	238	30%	315	40%	124

Note: Data presented is for the treatment group. The treatment and control group are equivalent.

Table 3. Program-Wide Sample by Grade, Optimal Use

Grade	N	Female		Hispanic		African American		Caucasian		SPED		Low-Income		BOY Composite
K	441	205	46%	44	10%	4	1%	372	84%	48	11%	252	57%	35
1	1102	518	47%	68	6%	9	1%	978	89%	121	11%	526	48%	123
2	159	64	40%	15	9%	0	0%	136	86%	45	28%	81	51%	61
3	95	41	43%	13	14%	0	0%	78	82%	36	38%	53	56%	98

Note: Data presented are for the treatment group. The treatment and control group are equivalent.

Vendor-specific Sample

Table 4. Vendor-specific Sample by Grade, ALL Users

Program	Grade	N	Female		Hispanic		African American		Caucasian		SPED		Low-income		BOY Composite
Istation	Kinder	115	61	53%	37	32%	2	2%	73	63%	16	14%	102	89%	25
	1st	167	79	47%	47	28%	1	1%	112	67%	22	13%	147	88%	130
	2nd	53	26	49%	18	34%	0	0%	31	58%	7	13%	43	81%	64
	3rd	31	18	58%	7	23%	0	0%	24	77%	5	16%	27	87%	135
Waterford	Kinder	1,641	796	49%	142	9%	13	1%	1,416	86%	133	8%	1,296	79%	35
	1st	1,711	853	50%	120	7%	10	1%	1,519	89%	168	10%	1011	59%	125
	2nd	296	118	40%	32	11%	2	1%	252	85%	58	20%	149	50%	72
	3rd	99	43	43%	12	12%	1	1%	82	83%	38	38%	47	47%	94
i-Ready	Kinder	880	413	47%	76	9%	7	1%	743	84%	73	8%	234	33%	36
	1st	1206	575	48%	86	7%	9	1%	1,060	88%	96	8%	276	39%	124
	2nd	385	177	46%	41	11%	7	2%	324	84%	89	23%	112	31%	83
	3rd	439	184	42%	50	11%	0	0%	374	85%	120	27%	116	20%	134
Imagine Learning	Kinder	3,868	1809	47%	378	10%	41	1%	3,255	84%	333	9%	1,879	49%	35
	1st	5,908	2886	49%	426	7%	39	1%	5,200	88%	594	10%	2,099	36%	121
	2nd	1,131	493	44%	117	10%	16	1%	957	85%	297	26%	477	42%	74
	3rd	781	359	46%	76	10%	2	0%	678	87%	249	32%	335	43%	119
Success-Maker	Kinder	274	131	48%	11	4%	1	0%	253	92%	37	14%	247	90%	44
	1st	752	356	47%	53	7%	2	0%	673	89%	93	12%	428	57%	124
	2nd	209	107	51%	9	4%	2	1%	194	93%	53	25%	84	40%	77
	3rd	204	105	51%	15	7%	2	1%	181	89%	51	25%	69	34%	134
Core5	Kinder	1,943	953	49%	152	8%	11	1%	1,712	88%	212	11%	1,025	53%	40
	1st	2,945	1478	50%	224	8%	20	1%	2,564	87%	259	9%	1,080	37%	136
	2nd	825	382	46%	128	16%	21	3%	618	75%	194	24%	360	44%	75
	3rd	817	362	44%	125	15%	16	2%	635	78%	260	32%	333	41%	131
Reading Plus	2nd	22	11	50%	1	5%	0	0%	21	95%	3	14%	0	0%	88
	3rd	58	25	43%	13	22%	0	0%	41	71%	7	12%	15	26%	151

Program	Grade	N	Female		Hispanic		African American		Caucasian		SPED		Low-income		BOY Composite
MyON	Kinder	49	25	51%	12	24%	0	0%	35	71%	9	18%	18	37%	30
	1st	260	134	52%	27	10%	0	0%	221	85%	33	13%	171	66%	148
	2nd	74	36	49%	12	16%	0	0%	61	82%	20	27%	49	66%	72
	3rd	130	65	50%	19	15%	0	0%	101	78%	25	19%	81	62%	128

Note: Data presented are for the treatment group. The treatment and control group are equivalent.

Table 5. Vendor-specific Sample by Grade, Relaxed Optimal Use

Program	Grade	N	Female		Hispanic		African American		Caucasian		SPED		Low-income		BOY Composite
Istation	Kinder	42	19	45%	15	36%	0	0%	27	64%	4	10%	19	45%	26
	1st	106	48	45%	32	30%	1	1%	70	66%	14	13%	103	97%	128
	2nd	16	7	44%	5	31%	0	0%	9	56%	2	13%	7	44%	64
	3rd	4	2	50%	1	25%	0	0%	3	75%	0	0%	0	0%	195
Waterford	Kinder	793	368	46%	73	9%	7	1%	686	87%	64	8%	787	99%	32
	1st	615	299	49%	40	7%	2	0%	551	90%	52	8%	453	74%	125
	2nd	93	40	43%	11	12%	1	1%	79	85%	12	13%	70	75%	67
	3rd	11	2	18%	2	18%	0	0%	8	73%	7	64%	8	73%	67
i-Ready	Kinder	244	106	49%	27	9%	1	0%	200	86%	23	8%	80	88%	38
	1st	437	197	48%	39	7%	3	0%	373	89%	38	6%	170	72%	122
	2nd	140	59	48%	14	9%	0	1%	119	85%	44	24%	44	71%	78
	3rd	132	53	42%	14	11%	0	0%	115	85%	45	27%	26	67%	129
Imagine Learning	Kinder	762	370	49%	75	10%	10	1%	642	84%	73	10%	462	61%	33
	1st	1998	984	49%	138	7%	14	1%	1762	88%	201	10%	936	47%	123
	2nd	321	128	40%	23	7%	3	1%	286	89%	110	34%	170	53%	69
	3rd	163	68	42%	19	12%	1	1%	140	86%	63	39%	88	54%	106
Success-Maker	Kinder	158	74	47%	8	5%	1	1%	145	92%	22	14%	134	85%	46
	1st	559	264	47%	27	5%	2	0%	517	92%	69	12%	256	46%	124
	2nd	83	49	59%	3	4%	0	0%	79	95%	13	16%	38	46%	79
	3rd	90	43	48%	1	1%	0	0%	87	97%	22	24%	26	29%	136

Program	Grade	N	Female		Hispanic		African American		Caucasian		SPED		Low-income		BOY Composite
Core5	Kinder	724	348	48%	66	9%	5	1%	623	86%	67	9%	421	58%	40
	1st	1758	910	52%	131	7%	9	1%	1558	89%	138	8%	806	46%	140
	2nd	421	189	45%	62	15%	7	2%	325	77%	87	21%	216	51%	73
	3rd	294	135	46%	65	22%	5	2%	214	73%	80	27%	141	48%	124
MyON	1st	24	10	42%	2	8%	0	0%	22	92%	1	4%	24	100%	166
	2nd	16	6	38%	3	19%	0	0%	12	75%	6	38%	7	44%	86
	3rd	31	13	42%	3	10%	0	0%	25	81%	4	13%	18	58%	137

Note: Data presented are for the treatment group. The treatment and control group are equivalent.

Appendix E. Program-wide Title 1 Results

Table 1. Predicted Means of EOY Composite for Matched Treatment and Control, Program-Wide, Title 1 schools only

	Kindergarten			1st Grade			2nd Grade			3rd Grade		
	Tr.	Cntrl	ES	Tr.	Cntrl	ES	Tr.	Cntrl	ES	Tr.	Cntrl	ES
Intent to Treat	N=3,252			N=3,226			N=1,046			N=881		
Composite	148	139	.10	187	190	-.01	161	157	.05	–	–	–
Relaxed Optimal	N=1,113			N=1,644			N=433			N=265		
	158	142	.22	200	194	.07	163	154	.10	263	255	.08
Optimal	N=234			N=451			N=68			N=50		
	161	144	.30	–	–	–	162	144	.22	–	–	–

Note: Treatment students are matched to control students using CEM. Model covariates are gender, Hispanic, special education, school Title I status, and BOY Composite score. A dash in a cell means that the treatment is not a significant effect for the model.