EDUCATION

THE UTAH STATE BOARD OF EDUCATION

Report to the Education Interim Committee

Early Interactive Reading Software Program Report

January 2021

Teresa Hartvigsen
Digital Teaching and Learning Program Specialist
teresa.hartvigsen@schools.utah.gov

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

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

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

ADA Compliant: 1/7/2021

Early Interactive Reading Software Program Report

STATUTORY REQUIREMENT

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

EXECUTIVE SUMMARY

The Early Interactive Reading Software Program encourages literacy growth and achievement in students in grades K-3. The program addresses early reading through the use of computer-based literacy software which provides individualized instruction designed to supplement students' classroom learning. During the 2019-2020 school year, these software programs were used in 139 local education agencies (LEAs) and by approximately 150,169 students. The schools use the software to build literacy skills for all students in kindergarten and first grade, as well as for intervention with students in second and third grade. The independent evaluation for the 2019-2020 school year is attached.

Utah's Early Interactive Reading Software Program

2019-2020 Program Evaluation Findings

Submitted to the Utah State Board of Education
October 2020

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

ADA Compliant: 1/7/2021

Table of Contents

EXECUTIVE SUMMARY 1

INTRODUCTION 4

PROGRAM IMPLEMENTATION FINDINGS 6

WHAT DID USAGE LOOK LIKE FOR EISP PARTICIPANTS? 7

PROGRAM IMPACTS ON LITERACY ACHIEVEMENT 9

DID THE PROGRAM HAVE AN OVERALL EFFECT ACROSS ALL VENDORS? 10
WHAT IMPACTS DOES EISP HAVE ON LITERACY SKILLS AS MEASURED BY THE ACADIENCE READING? 11
HOW DID PROGRAM USAGE EFFECT PROGRAM IMPACTS? 12

DISCUSSION, LIMITATIONS AND RECOMMENDATIONS 18

REFERENCES 20

APPENDIX A. EVALUATION METHODS 22

APPENDIX B: ANALYSES SAMPLES 27

APPENDIX C. DATA PROCESSING & MERGE SUMMARY 29

APPENDIX D: ACADIENCE READING MEASURES 31

APPENDIX E: DETERMINING EFFECT SIZE BENCHMARK 32

APPENDIX F. PROGRAM USE BY VENDOR AND GRADE 33

List of Figures

Figure 1: % Weeks Met Recommended Usage, by grade 9
Figure 2. Kindergarten MOY Mean Scores 13
Figure 3. First Grade MOY Mean Scores 14
Figure 4. Second Grade MOY Mean Scores 15
Figure 5. Third Grade MOY Mean Scores 16

List of Tables

Table 1. 2019-2020 Program Enrollment Overview 5
Table 2. 2019-2020 Program Enrollment by Grade 6
Table 3. Vendor 2019-2020 Minimum Dosage Recommendations 7
Table 4: Average EISP usage by grade from August 5th- March 9th 8
Table 5. Predicted Means of Acadience Composite Scores for Matched Treatment and Control, MRD Group 10
Table 6: Predicted Means of Acadience Composite Scores for Matched Treatment and Control, ITT Group 11
Table 7. Predicted Means of MOY Acadience Reading Literacy Domains for Matched Treatment and Control, MRD Sample 12
Table 8. Mean Score Differences on MOY Acadience Reading Composite Scores by Grade and Subgroup, MRD Sample 17

Acknowledgements

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

We also acknowledge Malia Mcllvenna, research consultant, at USBE for all her efforts preparing and transferring the student data used for the analyses.

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

Executive Summary

Evaluation Purpose

The Early Interactive Software Program (EISP) was designed to increase the literacy skills of all students in K-3 through the use of adaptive computer-based literacy programs. The program provided Utah's Local Education Agencies (LEAs) with an option to select among four adaptive computer-based literacy programs: Imagine Learning, Curriculum Associates (i-Ready), Lexia® (Core5), and Waterford. The Evaluation and Training Institute (ETI), the EISP external evaluator, studied two core aspects of the program: 1) student's use of the program during the first half of the school year ("program implementation"); and, 2) the effects the program had on increasing students' literacy achievement ("program impacts"), as measured by student's middle-of-year test scores, for all software programs combined. The 2019-2020 evaluation methods were adapted in response to the Covid-19 pandemic, which resulted in soft school closures and potentially shortened students' program exposure. The computer-based software programs were designed to be implemented across a full academic year. This report depicts program impacts at mid-year, before students received the full program. The evaluation was not designed to replace a full year-long study of program implementation and its impact on student learning, but the results can be used as a touchstone for gauging mid-year progress.

Program Enrollment and Implementation Findings

During the 2019-2020 school year, EISP was implemented in 139 Local Education Agencies (LEAs) and by 150,169 students throughout the state of Utah. Core5 was used by the most students (95,639), followed by Imagine Learning (38,966), i-Ready (9,411), and Waterford (6,152). State-wide program implementation set the stage for large numbers of students to receive program benefits, however, it was important for students to meet minimum usage requirements (set by program vendors) in order for the program to impact students' literacy achievement.

To that end, program vendors provided LEAs with recommendations on how many minutes per week students should have used the program, on average, as well as the total number of weeks the program should be used. The implementation study was designed to determine the extent to which students used the program, despite the lack of continuous in-person attendance during the academic year. We studied program implementation in two ways, making adjustments to account for Covid-19. First, we examined the average program usage from the beginning of the program (August 5th) through the end of in-person instruction (March 9th). Across all the grades in the EISP, students used the program for an average of 49 minutes per week, 1,068 average total minutes and for an average of 19 weeks. Students in first and second grade had the highest average usage across all categories, using the program an average of 50-52 minutes per week and for 20-22 weeks. Second, we used each vendor's weekly minute recommendations to evaluate how many students consistently met or exceeded the vendor's recommended minutes. We found that less than half of the students, regardless of grade level, were able to meet vendors' recommendations. That is, less than half used the program for more than 51% of the weeks. This trend highlights inconsistencies in program use from week-to-week, which may limit the programs' ability to generate strong impacts on students' literacy achievement.

Program-wide Impacts Findings

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 then compared how different program usage levels impacted students (Intent to Treat; "ITT") as well as the impact for students who used the software as recommended at least 51 percent of the weeks they were in the program (Met Recommended Dosage; "MRD"). Finally, we completed our analyses with an examination of program effects for specific groups of students.

Overall Impact

We found statistically significant treatment effects in all grade levels (K-3) among students who used the program as recommended by vendors for 51 percent or more weeks the program was used (MRD sample). Predicted mean score differences between treatment and control students ranged from a low of 4.23 points (second grade) to a high of 15.17 points (kindergarten). Effect sizes (calculated using Cohen's d; ES) were used to describe the magnitude of the program impact and were interpreted as meaningful if they reached a minimum threshold of .26[1]. Kindergarten had the highest effect size (ES: .33), but all other grade levels yielded effect sizes below the .26 threshold.

Impact on Literacy Skills

We examined the program's benefits on specific literacy skill development by comparing Acadience reading mean scores between treatment and control students in the MRD sample. This analysis gave stakeholders a view into how the software changed students' test scores in specific skill areas. Across all grade levels and literacy measures, program students had higher mean scores than their control group counterparts, although these differences were small for most literacy measures (from 0 to 5 points for 10/11 subscales).

Program Usage and Program Impacts

To determine how dosage affected outcomes, we studied the differences in middle-of-year mean scores among the different levels of program use. We split the Intent to Treat (ITT) treatment sample into the following five usage groups based on the percentage of weeks the students met vendors weekly recommended minutes: 0%; 1-25%, 26-50%, 51-75%, and 76% or more weeks. Across all grades, as program usage increased, students achieved higher mean reading composite scores at the middle-of-year.

[1] More information on the effect size benchmark is presented in Appendix E.

Student Characteristics and Program Impacts

We were interested in examining how the program may benefit students in specific subgroups, which led us to conduct a separate subgroup analysis. We examined how the program impacted students identified as English Language Learners (ELL), low-income, special education designation (SPED) status, and those who attended a Title 1 school. Across all grades, differential treatment effects were most pronounced in kindergarten, especially for ELL (27 points higher), SPED (22 points) and low-income (21 points) students. First and third grade had mean score differences in the double digits as well, with a low of 11 points (ELL third grade) and a high of 15 points (first grade SPED and low-income). The only two groups without statistically significant findings were ELL students in first grade and SPED students in second grade.

Discussion & Recommendations

Despite the limitations caused by Covid-19 and the soft closure in March, we identified positive student literacy achievement outcomes, specifically for students who met the vendor recommendations for weekly minutes of use for at least 51 percent of the weeks. Our findings underscore the importance of consistent program use from week-to-week.

We highlighted how consistent program use improves program outcomes through a descriptive analysis of program usage categories. Students' average test scores increased the closer they came to meeting vendors minimum recommended weekly program use. In the highest usage group (76-100% Weeks Met Recommendations; "WMR") the MOY mean score was "above benchmark" in all four grades. K-3 students in the second highest group (51-75% WMR) also had middle of year (MOY) mean scores that were "above benchmark." Students with scores "above benchmark", have a 90-99% likelihood of achieving subsequent reading outcomes (Dynamic Measurement Group, Inc., 2018).

Across all grades, we identified that only 24-39% of the students used the program in the highest two usage categories (51% WMR or higher). Most students in EISP fell into the lowest three categories, not meeting the weekly minute recommendations for more than half of the time that they were in the program. We therefore recommend that the state encourage consistent weekly use and continue to hold LEAs accountable for meeting vendors' dosage recommendations so that students may achieve higher outcome scores. We also recommend that future evaluations continue to explore the ways in which dosage at different levels impacts student outcomes, but this is only possible for the highest usage groups if enough students reach these usage categories.

Introduction

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 through the third grade. To participate in the Early Interactive Software Program (EISP), Local Education Agencies (LEAs) submitted applications to the USBE requesting funding for the use of specific reading software programs prior to the start of each school year. Four software vendors provided software and training to schools through the EISP in 2019-2020. The four vendors were (in alphabetical order): Curriculum Associates ("i-Ready"), Imagine Learning, Lexia® ("Core5®"), and Waterford.

The Evaluation and Training Institute (ETI) contracted with the Utah State Board of Education (USBE) to study how the reading software programs were used by schools and the impact they had on students' literacy development. The evaluation included results for the combined impact of all the software programs used in Utah schools.

In 2019-2020, the soft closure of Utah schools due to the Covid-19 pandemic caused two major setbacks to the evaluation: 1) students did not participate in an end-of-year reading assessment, which we typically use to measure program outcomes; and 2) students potentially stopped using the program earlier than in typical years, which limited their program exposure. As a result, ETI worked with the USBE to adapt the evaluation design and report format. This year's annual report focused on studying mid-year progress for all vendors combined using middle-of-year assessment data. The computer-based software programs were designed to be implemented across a full academic year. This report depicts program impacts at mid-year, before students received the full program. The evaluation was not designed to replace a full year-long study of program implementation and its impact on student learning, but the results can be used as a touchstone for gauging mid-year progress.

The following research questions were used to guide our program-wide evaluation:

1. How did students use the software program?
2. Did the program have an overall affect across all vendors?
3. What interactions between student characteristics and school type affect program impacts?
4. How did program usage effect program impacts?

In the remainder of this report, we include a description of the EISP and 2019-2020 program enrollment findings related to each research question and the two study objectives (program implementation and program impacts). A detailed summary of our research methods is included in **Appendix A**. We summarize the key findings and study limitations in the final sections of this report.

In 2019-2020, the four EISP software vendors were used in 139 LEAs and 573 schools and by 150,169 students. Due to a change in the legislation, EISP was offered to all students in K-3rd Grade, regardless of their beginning-of-year reading level[2]. As depicted in **Table 1**, Core5 was the most frequently used program (313 schools, 95, 639 students), while Waterford was used with the fewest students among the four vendors (6,152 students).

Table 1. 2019-2020 Program Enrollment Overview

Program	**LEAs**	**Schools**	**Students (K-3)**
Core5	53	313	95,639
Imagine Learning	43	168	38,966
i-Ready	18	39	9,411
Waterford	25	53	6,152
Total	139	573	150,169

Note. Some LEAs/schools used multiple vendors. Totals represent unique cases of LEAs and schools. Data source: software vendor data.

[2] In prior years, EISP was intended as an intervention for second and third grade students reading below grade level.

Table 2 presents 2019-2020 program enrollment by vendor and grade level. Student participation by grade varied by program. Imagine Learning and Core5 had a fairly even distribution of students across Grades K-3, while Waterford was used more frequently in earlier grades, and i-Ready was used more frequently in the upper-early grades.

Table 2. 2019-2020 Program Enrollment by Grade

Program	Kinder	1st	2nd	3rd
Core5	21,916	25,270	24,736	23,776
Imagine Learning	9,925	10,995	9,953	8,093
i-Ready	1,694	2,514	2,672	2,530
Waterford	2,878	2,503	766	5
Total	36,413	41,282	38,127	34,404

Note. Data source: software vendor data in K-3.

Program Implementation

It is important for evaluators to study program implementation prior to measuring the program impacts on student learning. With increased understanding of how a program was implemented, more meaningful conclusions can be made about the program impacts. Students must use the program long enough to have an impact on their literacy skill development. As a result, the most important aspect of EISP implementation was dosage, which is how much of the program a student received during the school year.

Each vendor provided recommendations for using the software program in order for it to have an impact on students' literacy achievement (**Table 3**). These recommendations included both a range of minutes per week, and a range of total weeks in the program. Recommended weekly use ranged from 20 minutes to 80 minutes of use per week and suggested total weeks of use ranged from 18 to 30 weeks.

Table 3. Vendor 2019-2020 Minimum Dosage Recommendations

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

Note. Core5 usage recommendations are automatically adjusted based on student need so that students who were working below grade level were assigned usage recommendations that were greater than those who worked at or above grade level.

In the following section, we explored the differences in usage across grade levels in order to better understand how the program was implemented, despite the lack of continuous in-person attendance during the academic year. We received usage data from all four program vendors through the end of in-person instruction (week of March 9th). Based on this factor, we present the average program usage from the beginning of the program (August 5th) through the end of in-person instruction (March 9th). We also wanted to understand the percentage of students who consistently met or exceeded the vendor's recommended minutes based on the weeks they were in the program. We used the vendors' recommended minutes per week to conduct this analysis.

What did usage look like for EISP participants?

In Table 4, we present the average number of minutes and weeks that students used the program through the end of in-person instruction. Across all the grades in the EISP, students used the program for an average of 49 minutes per week, 1,068 average total minutes and for an average of 19 weeks. Students in first and second grade had the highest average usage across all usage categories, using the program an average of 50-52 minutes per week and for 20-22 weeks. We also see that third-grade students had the lowest averages across all categories- average weekly minutes (40), average total minutes (741) and average weeks of use (16). A more detailed summary of student use, by vendor, is included in **Appendix F**.

Table 4: Average EISP usage by grade from August 5th- March 9th

Grade	N	Ave Weekly Min.	Ave Total Min.	Ave Wks. of Use
K	36,413	45	923	19
1	41,282	52	1177	22
2	38,127	50	1079	20
3	34,404	40	741	16
Total	**150,169**	**49**	**1,068**	**19**

We also wanted to understand how closely students followed vendors' recommendations for weekly minutes of use during the weeks they used the program. For each week that the child was in the program, we compared their weekly usage to the weekly minutes recommendations provided by each vendor.[3] We calculated the total percentage of weeks that the weekly minute recommendations were met and created the following five descriptive categories: met the recommendations for 0% of the weeks used, 1-25% of the weeks used, 26-50% of the weeks used, 51-75% of the weeks use, and 76% or more of the weeks used. Throughout the report, we define these categories as "WMR", weeks met recommendations.

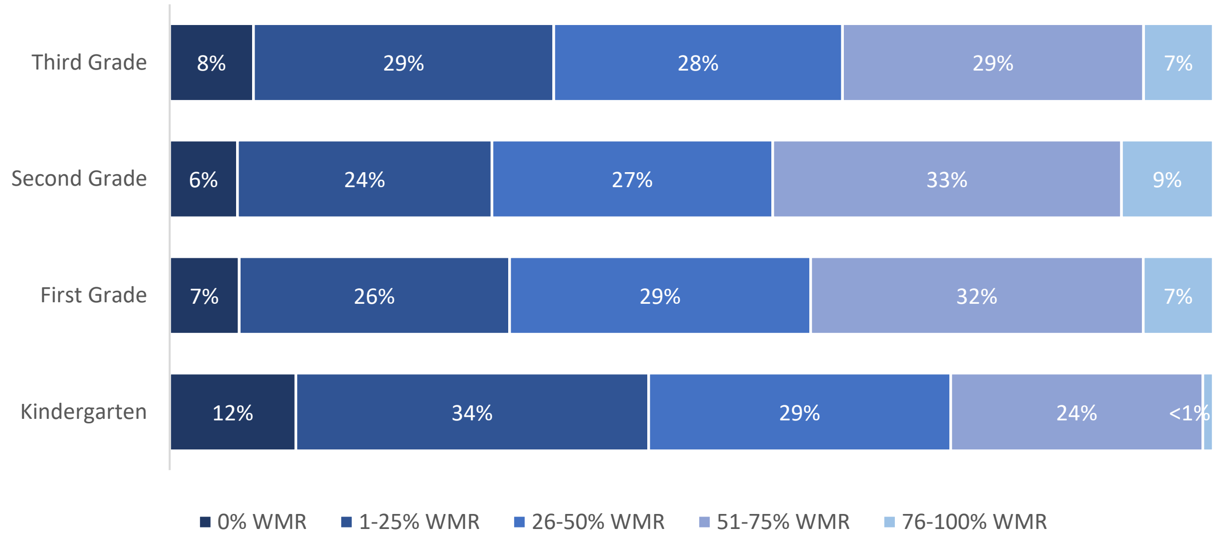
Figure 1 presents the distribution of students who met the weekly recommended minutes during program use. It shows that less than half of the students met vendors' recommendations for more than 51% of the weeks they used the program, irrespective of grade. This trend highlights inconsistencies in program use from week-to-week, which may need to be addressed with schools by program vendors.

First grade through third grade had similar distributions of students within each category of use, ranging from over a third (36%) to just less than half (42%) meeting the recommended 51%+ weeks of usage. As illustrated in Figure 1, second grade slightly outpaced the others.

Kindergarten had the highest percentage of students in the lowest dosage group (0% Weeks Met Recommendations "WMR"; 12% of sample) and the lowest percentage of students in the highest dosage group (76-100% WMR group; 0% of sample). Thirty-four percent of kindergarten students fell into the 1-25% WMR group and twenty-nine percent fell into the 26-50% WMR group, suggesting that kindergarten students may need more guidance in order to consistently meet or exceed the weekly minute recommendations.

[3] Weekly minute recommendations vary by vendor.

Figure 1: % Weeks Met Recommended Usage, by grade



N: 131,384: K (30,549); 1st (36,455); 2nd (33,741); 3rd (30,139)
Data source: Vendor data merged with Acadience Reading in Grades K-3.

Program Impacts on Literacy Achievement

We studied how the program impacted literacy achievement by comparing groups of students who used the program to groups of students who did not. This section includes findings on the impact of the EISP across all four software programs, providing a global view of how the program performed as it was used across the state. We have included a detailed methods section for technical reviewers in **Appendix A**.

Program Impacts

We began the program-wide analyses studying the program impacts of students who used the software based on vendors' minutes recommendations (MRD sample) and also for students who participated with different levels of program use (ITT sample). This analysis helped show the relationship between program effects and program use (or dosage) as well as program effects for MOY literacy composite scores for each grade. We further explored the connection between dosage and program outcomes through a descriptive analysis of five dosage categories and their MOY mean scores. 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?

We examined program impacts for students based on the following two analytic samples: 1.) students who met the program vendors' minutes recommendations for at least 51 percent of the weeks the program was used (MRD sample); and 2.) students who used the program, irrespective of their usage, which we identified as our Intent to Treat (ITT) sample.

The ITT analyses showed how the program affected all students throughout the state (in our sample), and the MRD analyses showed how a higher usage threshold was related to effects.

Tables 5 and 6 present the predicted mean scores and effect sizes of the matched treatment and control sample for the ITT and MRD analytic samples. As shown in **Table 5**, there were statistically significant treatment effects in all grade levels (K-3), with predicted mean score differences ranging from a low of 4.23 points (second grade) to a high of 15.17 points (kindergarten). 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 defined this threshold as any effect size equal or greater to .26, which is the average effect size seen in similar intervention programs (Lipsey et. al, 2012). Kindergarten had the highest effect size (ES: .33), which was above the ES threshold. All other grade levels had effect sizes below the .26 threshold.

Table 5. Predicted Means of Acadience Composite Scores for Matched Treatment and Control, MRD Group

Grade		N	Mean	Difference	SD	Effect Size
K (N=20,012)	Program	7,876	167.07	15.17	46.2	**.33**
	Non-Program	12,136	151.90			
1 (N= 20,231)	Program	13,585	216.43	10.98	83.6	.13
	Non-Program	6,646	205.45			
2 (N=20,434)	Program	13,825	264.46	4.23	63	.07
	Non-Program	6,609	260.23			
3 (N=20,385)	Program	10,430	371.35	11.83	70	.17
	Non-Program	9,955	359.52			

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

Table 6 presents the predicted mean scores and effect sizes of the matched ITT sample. As expected, the students in the ITT sample had fewer statistically significant treatment effects compared to the MRD group. Kindergarten was the only grade to produce a positive treatment effect (effect size: .06), which was well below the .26 ES threshold and cannot be considered a meaningful effect on learning. All the other grade levels were not significant (1st grade; 3rd grade) or had a negative effect size (second grade; ES: -.03). The ITT analytic sample includes students who used the program at different dosage levels, not necessarily meeting the vendor recommendations. These findings suggest that the ITT students did not use the program enough to produce strong positive treatment effects.

Table 6: Predicted Means of Acadience Composite Scores for Matched Treatment and Control, ITT Group

Grade		N	Mean	Difference	SD	Effect Size
K (N=43,306)	Program	30,747	147.92	7.65	129.97	0.06
	Non-Program	12,559	140.27			
1 (N= 40,986)	Program	34,289	187.45	2.38	142.98	NS
	Non-Program	6,697	185.07			
2 (N=38,813)	Program	32,105	237.20	-3.26	118.46	-0.03
	Non-Program	6,708	240.46			
3 (N=38,381)	Program	28,344	340.12	-0.37	131.59	NS
	Non-Program	10,037	340.48			

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

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

We examined the program's benefits on specific literacy skill development (**Table 6)** by comparing Acadience Reading mean scores between treatment and control students. This analysis gave stakeholders a view into how the software changed students' test scores in specific skill areas. Program students had higher mean scores than their control group counterparts across all grade levels and literacy measures, although these differences were small for most literacy measures (from 0 to 5 points for 10/11 subscales). The largest difference in mean scores was observed for developing first grade students' alphabetic principles and basic phonics skills (NWF: CLS), with program students scoring 11 points higher, on average, than the control group. In the upper-early grades the reading comprehension subscale, MAZE, was significant.

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

	Kindergarten			1st Grade			2nd Grade			3rd Grade		
Acadience Scale		N= 20,111			N= 20,226-20,231			N= 20,430-20,434			N= 20,381-20,385	
	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.	Tr.	C	Dif.
First Sound Fluency (FSF)	43	39	3		N/A			N/A			N/A	
Letter Naming Fluency (LNF)	43	41	2		N/A			N/A			N/A	
Nonsense Word Fluency-CLS	N/A	N/A	N/A	216	205	11		N/A			N/A	
Nonsense Word Fluency-WWR	4	4	0	23	21	2		N/A			N/A	
Oral Reading Fluency (ORF)		N/A		49	47	2		NS			NS	
Phoneme Segmentation Fluency (PSF)	46	41	5		N/A			N/A			N/A	
MAZE		N/A			N/A			N/A		18	16	2

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

How did program usage effect program impacts?

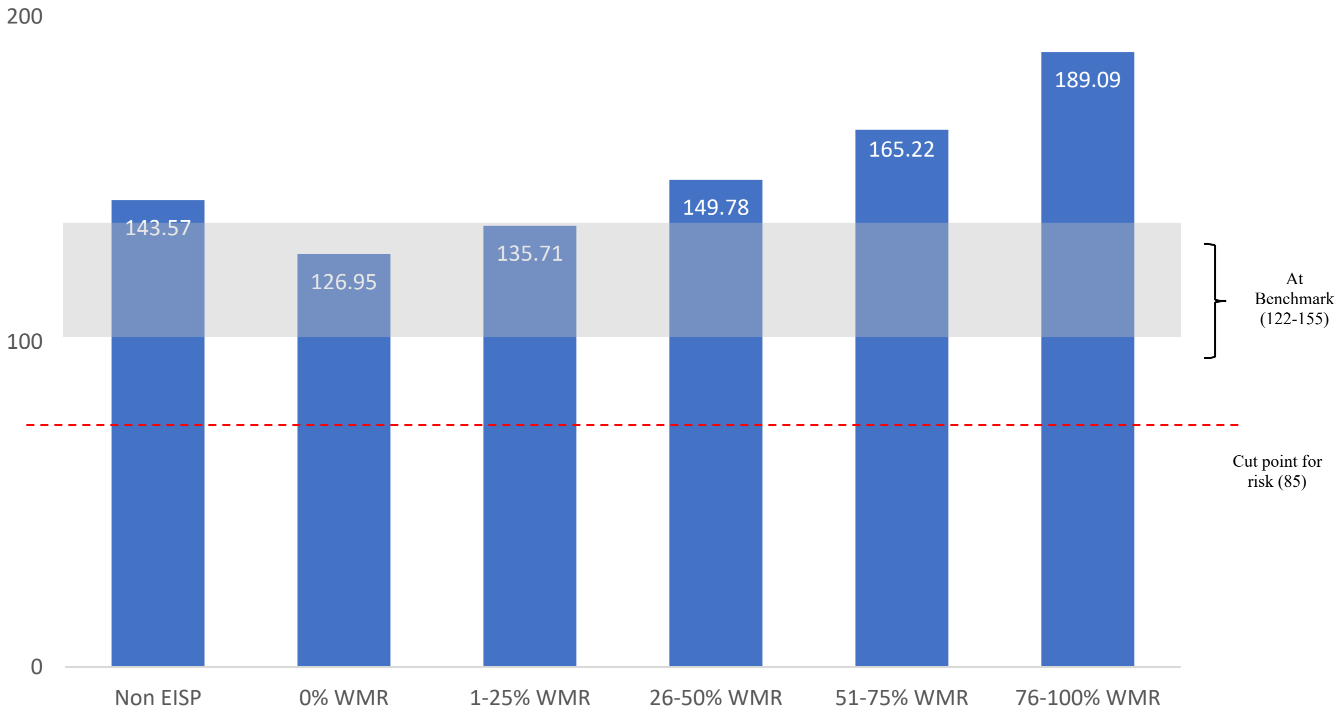
To determine how dosage affected outcomes, we studied the differences in mean scores among the different levels of program use. We split the ITT treatment sample into the following five usage groups based on the percentage of weeks the students met vendors' weekly minutes:

1) 0% weeks met recommendations (WMR)
2) 1-25% WMR
3) 26-50% WMR
4) 51-75% WMR
5) 76% or more WMR

To provide additional context, K-3 students across Utah who were not involved in the Early Software Interactive Program (EISP) were also included in this analysis.

Figure 2 presents the average middle-of-year mean scores for EISP kindergarteners and kindergarteners who were not exposed to an early interactive software program funded by the state ("Non-EISP"). This figure highlights a trend that supports our hypothesis- as program usage increases, students achieve higher mean reading composite scores. Students in the two highest program dosage groups (51-75% and 76-100% WMR) had the highest MOY mean composite scores and were the only groups to achieve an "above benchmark" MOY mean score (156 or greater). Students with scores "above benchmark", have a 90-99% likelihood of achieving subsequent reading outcomes (Dynamic Measurement Group, Inc., 2018). In addition, program students who used the software as intended for at least 26-50% of the weeks, did better than students who did not use EISP at all (non-EISP group).

Figure 2. Kindergarten MOY Mean Scores

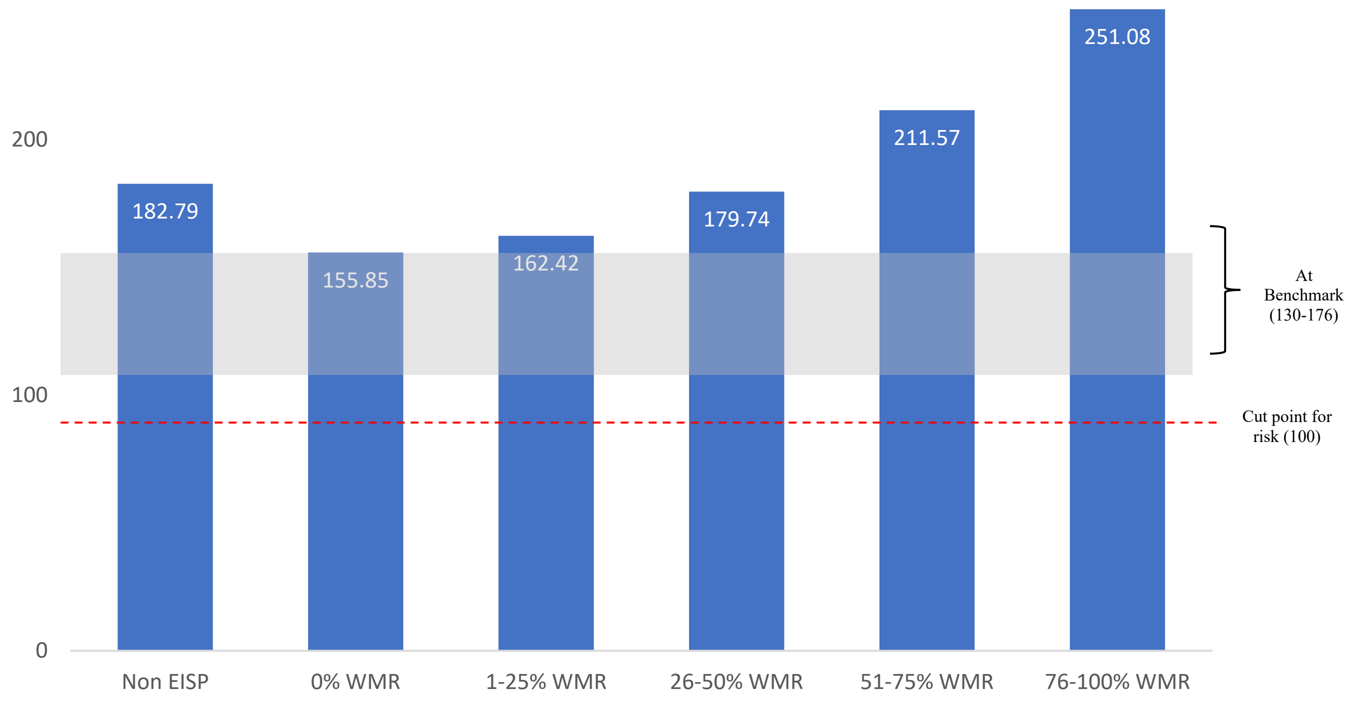


*Note: Kindergarten sample size - Non- EISP n= 12,468; 0% WMR n= 3,698; 1-25% WMR n=10,345; 26-50% WMR n=8,848; 51-75% WMR n=7,377; 76-100% WMR n=514. Students scoring **At Benchmark** (122-155) or **Above Benchmark** goal (156 or greater) have the odds in their favor (approximately 80% to 90% overall) of achieving later important reading outcomes.[4] Students scoring below the **cut point for risk** (85 and below) are unlikely (approximately 10%–20% overall) to achieve subsequent goals without receiving additional, targeted instructional support.*

[4] It is hard to predict the odds of students scoring above the cut point for risk in the **Below Benchmark (85-121)** range. These students are likely (40%-60% overall) to need strategic support that is adjusted to meet individual needs in order to achieve early literacy goals.

Figure 3 presents the average middle-of-year mean scores for first grade program students and Non- EISP first grade students. Students in the highest usage group (76-100%) had the highest MOY mean score (251.08), which falls in the "above benchmark" score range. This score was 39.51 points higher than the mean score of the second highest usage group (students who used the program 51-75% WMR). Interestingly, students in the lowest usage group, (0% WMR) had the lowest MOY mean score, even compared to students who were not in the EISP program. Though the MOY mean score for this group met the benchmark goal of 122, this low mean score may signify that the program must be used consistently, in order to receive the highest program benefits.

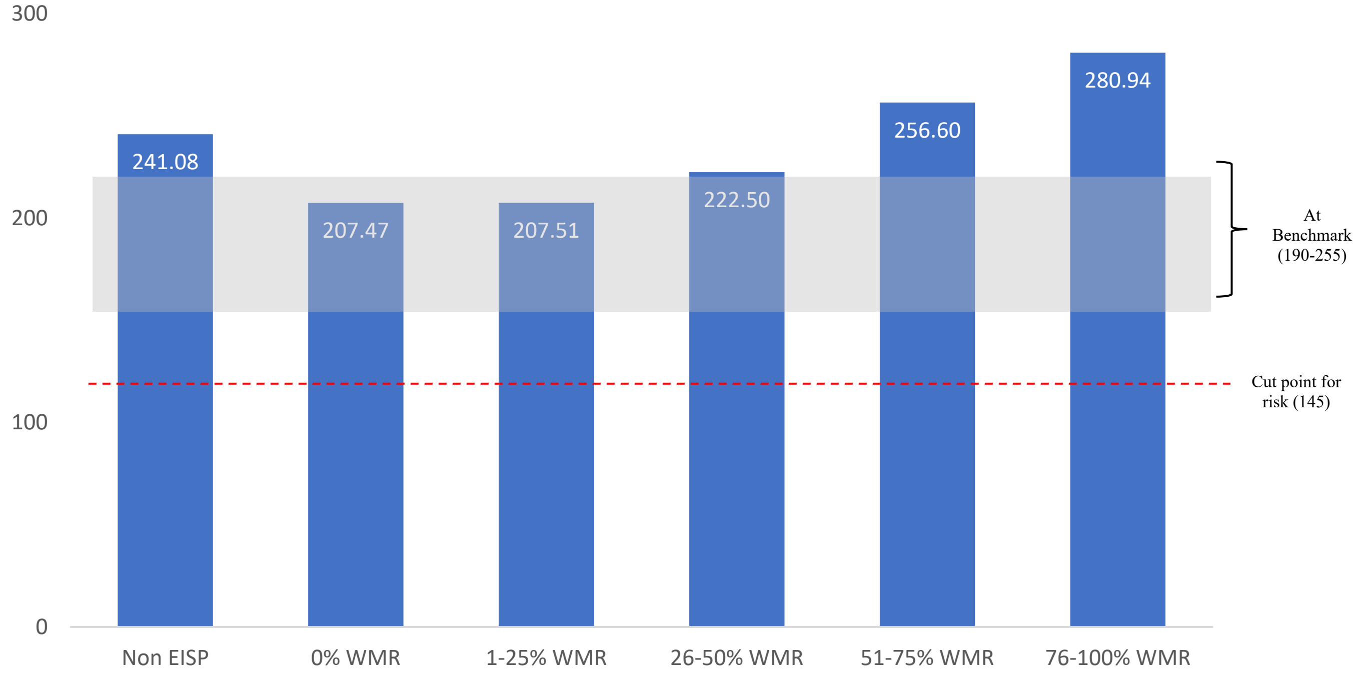
Figure 3. First Grade MOY Mean Scores



Note: First Grade sample size - Non- EISP n= 9,713; 0% WMR n= 2,419; 1-25% WMR n=9,416; 26-50% WMR n=10,506; 51-75% WMR n=11,607; 76-100% WMR n=2,439. Students scoring **At Benchmark** (130-176) or **Above Benchmark** goal (177 or greater) have the odds in their favor (approximately 80% to 90% overall) of achieving later important reading outcomes. Students scoring below the **cut point for risk** (100 and below) are unlikely (approximately 10%–20% overall) to achieve subsequent goals without receiving additional, targeted instructional support.

Figure 4 presents the average MOY mean scores for EISP second graders and non-EISP second graders. This graph shows that students in the two highest usage groups (51-75% and 76-100%) were the only two groups to have a MOY mean score that was "above benchmark". All usage groups, including the Non-EISP group, had a mean score that was in the "at benchmark" range.

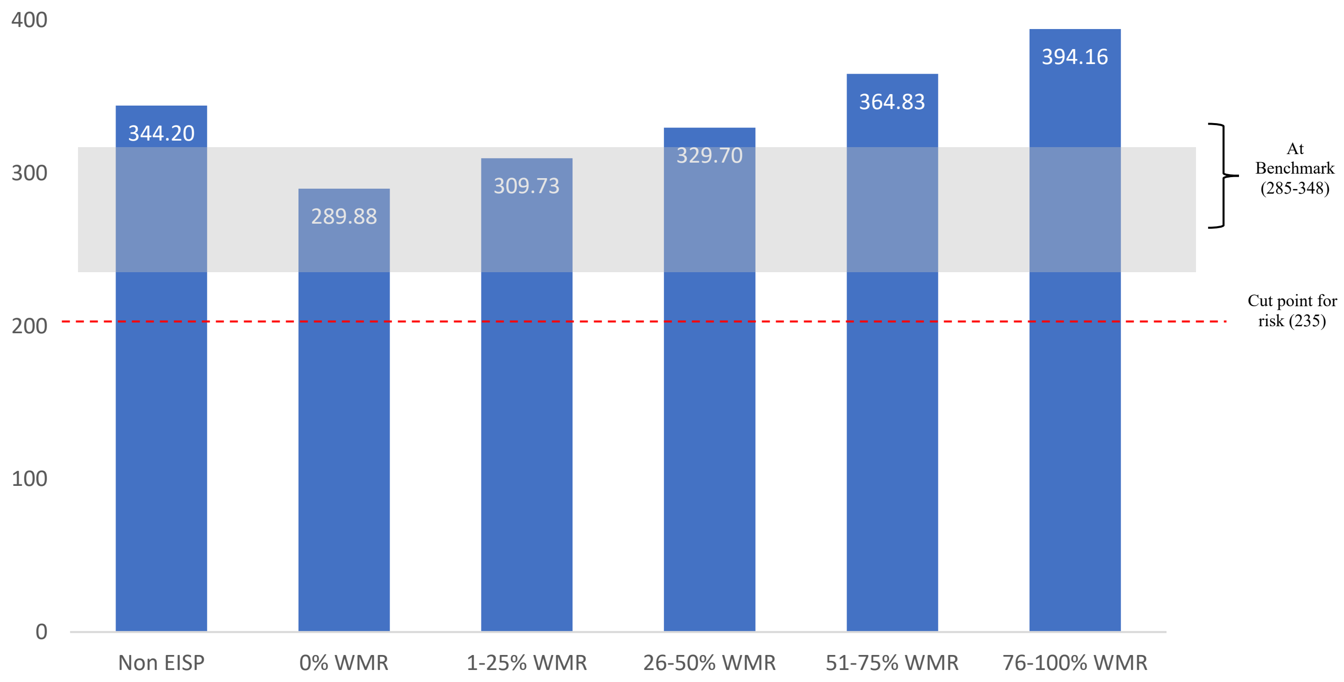
Figure 4. Second Grade MOY Mean Scores



Note: Second Grade sample size - Non- EISP n= 11,536; 0% WMR n= 2,177; 1-25% WMR n=8,208; 26-50% WMR n=9,060; 51-75% WMR n=11,272; 76-100% WMR n=2,962. Students scoring **At Benchmark** (190-255) or **Above Benchmark** goal (256 or greater) have the odds in their favor (approximately 80% to 90% overall) of achieving later important reading outcomes. Students scoring below the **cut point for risk** (145 and below) are unlikely (approximately 10%–20% overall) to achieve subsequent goals without receiving additional, targeted instructional support.

Figure 5 presents the middle-of-year mean scores for program and non-EISP students in the third grade. All dosage groups, including Non-EISP students, were “at benchmark” for their grade, which signifies a 70% to 80% likelihood of achieving later reading outcomes. Similar to trends observed in kindergarten and second grade, students in the highest two usage groups were the only groups with mean scores in the “above benchmark” range (349 or higher). These students are highly likely to achieve later reading outcomes.

Figure 5. Third Grade MOY Mean Scores



Note: Third Grade sample size- Non- EISP n= 15,990; 0% WMR n= 2,419; 1-25% WMR n=8,664; 26-50% WMR n=8,334; 51-75% WMR n=8,690; 76-100% WMR n=1960. Students scoring **At Benchmark** (285-348) or **Above Benchmark** goal (349 or greater) have the odds in their favor (approximately 80% to 90% overall) of achieving later important reading outcomes. Students scoring below the **cut point for risk** (235 and below) are unlikely (approximately 10%–20% overall) to achieve subsequent goals without receiving additional, targeted instructional support.

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

Table 8 presents the mean score differences in Acadience Reading composite scores at middle-of-year for certain subgroups of students. Our findings indicate that students in certain groups did better if they use the program than if they did not. Among all the grades, differential treatment effects were the most pronounced in kindergarten, particularly for ELL students: program students scored 27 points higher than non-program ELL students. Second and third grade students also had mean score differences in the double digits, with scores ranging from 11-points (ELL third grade) to 15 points (SPED and low-income in first grade). In kindergarten and in third grade, program students in all four subgroups (low-income, special education (SPED), English Language Learners (ELL), and those who attended a Title 1 school) had higher mean scores than non-program students in the same subgroups. There were no statistically significant differences for English Language Learners in first grade and SPED students in second grade.

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

	Kindergarten	1st Grade	2nd Grade	3rd Grade
ELL	27 (1,334)	NS	8 (1,770)	11 (2,010)
Low-income	21 (6,002)	15 (6,458)	8 6,779	13 (6,932)
Special Education (SPED)	22 (N=1,638)	15 (1,683)	NS	15 (2,178)
Title I Schools	16 (5,872)	14 (5,804)	8 (5,791)	16 (6,071)

Note. NS (not significant) in a cell means the program did not have a significant effect.
Data source: Matched K-3 MRD sample.
All data points displayed in figure were statistically significant at p≤ .05.

Discussion, Limitations and Recommendations

There were two primary evaluation goals: 1) to study program implementation in relation to vendors' dosage recommendations; and, 2) to determine the impacts of the program on students' literacy achievement. We summarized the key findings for both goals in this section and provide recommendations to help improve the program within each section.

Implementation. The softened school year gave us an opportunity to focus on program implementation in a different way than in prior years. Instead of using average program use as a benchmark for successful implementation, we developed a measure that relied on the percentage of weeks a student met the vendor's minutes of use recommendation. Our adapted approach was designed to measure consistent program use. Based on this approach, we discovered fewer than half of the students met vendors' recommendation for more than 50 percent of the weeks they used the reading software. This trend was true across grade levels and indicates a need for further support to schools and students to increase their usage to a more consistent level.

Impacts. The need for more consistency is underscored by our findings, which show a clear link between more consistent program use and stronger program effects. The program had a positive impact on students who used it as intended for at least fifty-one percent of the weeks. Students in this dosage group had higher middle-of-year composite scores compared to a matched group of similar students who did not use the reading software across all grade levels. The link between program dosage and impacts was further supported through our descriptive analysis, that showed students' middle-of-year composite scores increased exponentially as the weeks they used the software program also increased.

The program was shown to have strong benefits in specific grades for certain subgroups of students: English Language Learners (ELL), special education, low-income, and Title-1 students. The treatment and control predicted mean score differences were all above 10 points in kindergarten, first and third grade. Kindergarten students in these groups benefited the most from program participation, with composite scores over 20 points higher than similar students from the following groups: SPED: 22 points; ELL: 27 points; low-income: 21 points.

Recommendations:

- We have shown the importance of consistent program use, and we recommend that the state continue to encourage schools to meet fidelity of program use through accountability measures such as the annual fidelity of use report.
- Vendors should provide frequent updates to schools to help them monitor their program use and communicate the importance of consistency in program use from week-to-week.
- The program is especially effective for children in certain groups, and the state should continue to offer it to students who are English Language Learners, special education, low-income, and in Title 1 schools.
- Future evaluations should continue to explore the relationship between program dosage and program outcomes.

Limitations. To understand the effect of the program on literacy achievement we compare program students to a group of similar non-program students. In recent years, we have learned that LEAs have increased their use of digital technology intervention programs in the state. Therefore, it is possible that some of our control students used similar intervention programs, which may underestimate the strength of the program impacts. It is also possible that some LEAs used the same reading interventions with students using a non-EISP funding source. Future evaluations would benefit from the USBE and vendors tracking and sharing this information.

This year we used middle-of-year composite scores to measure program impacts and it is possible that students' impacts were underestimated given the shorter program duration. We recommend that future evaluation continue to use end-of-year scores in order to maximize the amount of time students use the software and its potential impacts on students.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix A. Evaluation Methods

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

Our evaluation methods were adapted from previous years to account for the disruption to 2019-2020 school year due to the coronavirus pandemic. We used middle of the year (MOY) Acadience test scores instead of end of year scores as our program outcomes, and we adjusted program entrance and exit dates based on the lack of continuous in-person attendance during the academic year (more details below). In the remainder of this section 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 B-C** provide additional details on our methods, data processing procedures and samples.

Which program participants were included in our study?

Implementation Study Evaluation Participant Samples

The goal of the implementation study was to examine the extent to which students used the software as intended by each program vendor. We included as many students who used the programs as possible to provide the most accurate depiction of students' program use, and the samples used for the implementation analyses were the most inclusive of all the samples. For K-3 students, we used the vendor data, and did not remove students with inaccurate SSIDs, students who used multiple software providers, or students with incomplete Acadience data.

Impact Study Evaluation Participant Samples

In a normal program year, the impact analyses relied on students using the program for the duration of the school year. Due to the lack of a normal school year and our use of MOY test scores to determine the program's impact, in conjunction with the USBE we elected to use a sample of student program participants *who met a minimum threshold of program use* from the larger pool of total program students to create an "met recommended dosage analytic sample," (MRD*; see **Appendix B** for descriptive statistics of the students included MRD).* We created MRD samples based on the specific combination of vendor and grade of students being analyzed. To be included in our analytic samples, students needed to have accurate state student SSIDs (unique identification numbers used by the state to track students in K-12) and complete Acadience test score data (outcome data). Further, we excluded students who may have used multiple software programs during the year to reduce "treatment cross-program contamination" effects, and we removed potential non-program control students who had been exposed to the

EISP program during the 2018-2019 school year to reduce "control program exposure contamination."

Control Student Matching Process. Our impact study compared Acadience literacy test scores between EISP program students (the treatment group) to a group of non-program students (the control group). Since we were not able to randomly assign students to treatment or control groups, we matched preexisting program 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).

We employed a CEM approach designed to retain as many treatment cases as possible. There were fewer control students than treatment students, which resulted in slight pretest imbalances between our matched treatment and control groups (these imbalances were statistically corrected by using weighting to balance the differences in mean values of the covariates between groups; see the below description about linear regression models). Despite these slight differences, our approach led to a well-balanced analytic sample, as indicated by an L1 score[5] of 0.0000000000000022960. Lower values indicate less imbalance, and the closer to zero the better the two samples were balanced across covariates.

We explored how program dosage impacted students' literacy skill development. Because vendor recommendations for dosage were not adjusted for the soft school closure in March, we created an adjusted usage definition for our program wide analyses. First, we determined if children met the weekly minute recommendation for each week that they used the program. We were then able to calculate what percentage of weeks children met the target number of minutes, which varied by vendor and grade. We defined our MRD usage threshold as all students who met 51% or more of the weekly recommended minute target. For example, if a child was in the program for twenty weeks, they were considered having optimal usage if they met the minute target for at least 51% of the weeks.

We created two matched treatment and control samples based on two dosage thresholds. The first matched sample was comprised of students who used the software based on our optimal usage definition. The recommended weekly minutes dosage was based on vendors recommendations for how much time students should use the program before benefits are observed, and we wanted to determine how literacy outcomes were affected for students who met these recommendations.

[5] The L1 statistic is a comprehensive measure of global imbalance (Iacus, King and Porro, 2008). It is based on the L1 difference between the multidimensional histogram of all pretreatment covariates in the treated group and that in the control group.

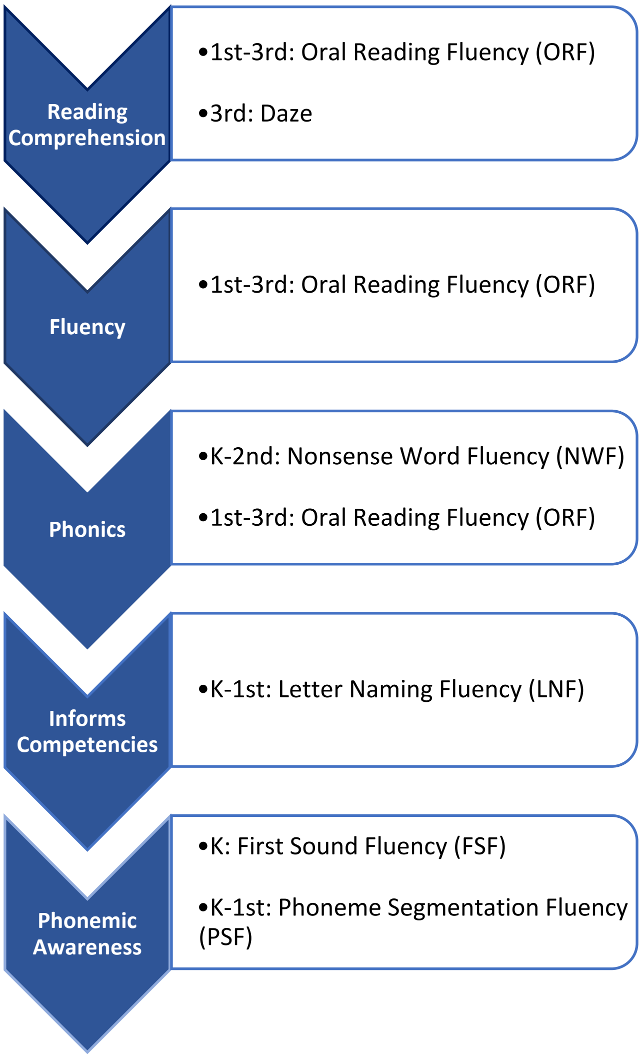
The second matched sample, the Intent to Treat (ITT) sample, included all students who used the program for any amount of time and showed how effective the program was for students, irrespective of use.

Finally, we used the MRD variable to create program dosage quintiles, which we used to study the effects of increased program use on students' test scores across vendors.

What sources of data were used in our analyses?

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

Figure A1: Acadience Indicator & Literacy Skill Measures



Reading Comprehension
- 1st-3rd: Oral Reading Fluency (ORF)
- 3rd: Daze

Fluency
- 1st-3rd: Oral Reading Fluency (ORF)

Phonics
- K-2nd: Nonsense Word Fluency (NWF)
- 1st-3rd: Oral Reading Fluency (ORF)

Informs Competencies
- K-1st: Letter Naming Fluency (LNF)

Phonemic Awareness
- K: First Sound Fluency (FSF)
- K-1st: Phoneme Segmentation Fluency (PSF)

Which instruments did we use to measure literacy achievement?

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

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, through the end of in person instruction (March 13th) due to COVID-19.

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

Our study relied on statistical analyses to measure program impacts, which included linear regression modeling (OLS), and descriptive analyses of trends related to levels of program use and Acadience benchmark category outcomes.

Linear regression models. We studied the program impacts on students' Acadience test scores by comparing a sample of treatment group students drawn from all vendors to a matched sample of control students. We determined that using an ordinary least squares (OLS) regression model allowed us to study the differences in treatment and control group test scores, while controlling for other important predictors of reading achievement. We used OLS 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, school district, economic disadvantaged status, and ethnicity 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 applied the use of weights to our regression analysis to balance he differences in mean values of the covariates between treatment and control groups. The control observations were given weights such that the joint distribution of the multidimensional analytic sample achieved balance. Sometimes this meant the controls were given more weight and sometimes it means they were given less weight.

Benchmark Category Outcome Descriptive Analyses. To present our findings in an intuitive and applicable context, we measured the change in students' reading proficiency benchmark levels based on different categories of program exposure, or "dosage." Dosage categories ranged from zero (i.e. no program exposure/control students) to the highest category of meeting minimum program requirements for 75%-100% of weeks the program was used. As a complement to our OLS regression (causal) analysis, we used the benchmark descriptive analysis to show the association between levels of program use for all students in the program. Students in the benchmark descriptive analysis were not matched and the results support an inference of association not causation.

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[6], or "d") to help readers understand the magnitude of treatment effects. Presenting effect sizes enabled us to provide a standardized scale to compare results based on different samples and measure the relative strengths of program impacts. Descriptive statistics, such as percentages, were presented to describe students' program use and change in reading proficiency benchmark status.

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

For the purposes of this study, we have chosen to contextualize our findings using the average of all three methods as our benchmark. The mean effect size for similar instructional programs is .26, and we consider this the standard by which to compare our results. Effect sizes larger than this are stronger than average, which we note in our results.[7] More information on how we selected our ES benchmark is provided in **Appendix E**.

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

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

Appendix B: Analyses Samples

Tables B1 – B3 present the characteristics of the population sample, and treatment and control group for each matched sample used in our analyses.

Table B1. Study Population by Grade

Grade	N	Female	Caucasian	SPED	Low-income	ELL	BOY Comp
K	43,666	48%	75%	9%	31%	8%	36.30
1	46,203	49%	75%	11%	34%	9%	125.60
2	45,315	49%	75%	12%	34%	10%	188.09
3	46,156	49%	74%	14%	34%	11%	272.32

Table B2. MRD Sample by Grade[8]

	Grade	N	Female	Caucasian	SPED	Low-income	ELL	BOY Comp
Control	K	12,136	49%	78%	9%	29%	7%	36.92
	1	6,646	49%	76%	10%	33%	9%	125.58
	2	6,609	49%	76%	12%	34%	9%	194.06
	3	9,955	48%	78%	12%	33%	9%	280.36
Treatment	K	7,876	48%	77%	7%	32%	7%	42.17
	1	13,585	48%	79%	8%	31%	7%	137.62
	2	13,825	49%	77%	9%	33%	8%	210.98
	3	10,430	50%	75%	10%	35%	11%	300.39

[8] The matched sample had an L1 score of 0.0000000000000022960. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates.

Table B3. ITT Sample by Grade [9]

	Grade	N	Female	Caucasian	SPED	Low-income	ELL	BOY Comp
Control	K	12,559	48%	76%	10%	29%	7%	36.47
	1	6,697	49%	75%	10%	33%	9%	124.76
	2	6,708	49%	75%	12%	34%	9%	193.50
	3	10,037	48%	78%	12%	33%	9%	279.23
Treatment	K	30,747	48%	75%	9%	32%	8%	36.05
	1	34,289	48%	79%	9%	33%	8%	125.55
	2	32,105	49%	77%	11%	33%	9%	188.80
	3	28,344	49%	77%	13%	34%	10%	279.23

[9] The matched sample had an L1 score of 0.00000000000002323.

Appendix C. Data Processing & Merge Summary

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

Software Program Data

Each software program provider provided student level data with the time students spent in the software for each week of school, through the week of March 9th, 2020. To help vendors provide quality data and ensure consistency across software program providers, vendors received an example data file, a description of the correct format for each variable, and a checklist to conduct a final review of their data. Our cleaning process for the program vendor data files included making sure all program schools that received licenses were included in the data, identifying and processing duplicate IDs within vendors' data, and formatting variables as needed, among other steps. We reviewed existing variables and created additional variables to use in our analyses, such as total weeks of use, average minutes of use, and other program fidelity measures.

When cleaning duplicate IDs within each vendors' data, we deleted cases that were the same student with different usage reported and kept any unique cases after removing exact replicas. We did not count weeks, or include minutes, when there were fewer than five minutes recorded in a given week. After removing these instances, we updated the usage variables, such as total minutes, to reflect the change in use, and then removed students who had fewer than five minutes of total use from the data. After we cleaned and processed the vendors data, the total count of students went from 158,991 to 150,169 students. We used this data to study program implementation.

To create the vendor data used in our outcome analyses, we identified and removed duplicate IDs across vendors[10] (approximately 2,467 cases) and any IDs that did not comply with the state student ID (SSID) format (5,023 cases). The duplicate IDs across vendors indicated students used more than one software program, either because they moved to a different district, or because the LEA administered multiple programs to the same students. In either case, we did not include these students in order to report the individual impacts for each software provider. This left us with a file of 142,679 cases.

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

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 2019-2020 participants were included in the data. The SIS data file consisted of 211, 563 cases, of which approximately 1 percent were duplicate records. After cleaning the data of duplicates, our SIS data consisted of 211, 539 records.

Acadience Reading Data

In 2019-2020, the USBE prepared and transferred an Acadience Reading data file (n=188,074). After cleaning the IDs (e.g. deleting missing IDs and IDs that were not in a valid format) and removing duplicates, we were left with a master Acadience file containing 182,649 cases. This master file contained outcome data for our pool of treatment and control cases.

Master Merged Data File

We merged the SIS data from the USBE into our master Acadience Reading file and were left with 182,615 cases. Next, we merged our master vendor data into the Acadience and SIS data and removed duplicate cases between vendors and missing data (e.g. beginning and middle-of-year composite scores). This left us with 131,384 complete treatment cases and 51,233 control cases.

Lastly, we identified (where possible) schools or students with program exposure, either using one of the four program vendors through non-EISP funding or using the program in the 2018-2019 school year. We removed these cases from our pool of potential controls[11]. This included excluding students who used Imagine Learning through a separate state-wide grant[12] prior to reporting the program impacts for similar reasons. After processing the data, our final, pre-matched dataset consisted of 167,929 cases, of which, 131,384 were treatment and 36,545 were potential controls.

Matched Data Files

Before we could run our analyses, the final step was to create our matched control groups. Control students were drawn from a group of children who were not exposed to an early intervention software program (EISP) in 2019-2020. We needed to create a comparison group that matched the students in our treatment sample. We drew controls from a pool of non-program participants in the state of Utah, and in general, lost very few cases when creating our matched samples for individual vendors and the program-wide analyses which consisted of fewer students (e.g. the Met Recommended Dosage samples). However, for our largest sample of program students, the Intent to Treat (ITT) program-wide sample, there were more program students than control students. This automatically reduced the size of this particular sample.

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

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

Appendix D: Acadience Reading Measures

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

Table D1. Acadience Reading Scales

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

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

Appendix E: Determining Effect Size Benchmark

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

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

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

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

Appendix F. Program Use by Vendor and Grade

Table F1 presents a comprehensive summary of usage for each vendor and grade. The table includes usage frequencies, such as average weekly minutes of use, average total minutes of use, and average number of weeks of use through the week of March 9th, 2020.

Table F1. Program Use by Vendor and Grade

	Grade	N	Ave Weekly Min.	Ave Total Min.	Ave Wks. of Use
Core5	K	21916	51	1,047	19
	1	25270	62	1,449	23
	2	24736	56	1,272	22
	3	23776	51	1,107	21
	Total	**95698**	**55**	**1,226**	**21**
Imagine Learning	K	9,925	40	825	20
	1	10,995	48	1,062	21
	2	9953	48	1,062	21
	3	8093	43	882	19
	Total	38966	44	964	20
i-Ready	K	1,694	35	589	16
	1	2,514	39	795	20
	2	2,672	43	930	21
	3	2,530	43	821	19
	Total	**9,410**	**40**	**803**	**19**
Waterford	K	2,878	54	1,229	22
	1	2,503	59	1,402	23
	2	766	51	1,050	17
	3	5	24	153	5
	Total	**6,152**	**56**	**1,277**	**22**

Note. K-3 Data source: vendor usage data before cleaning invalid SSIDs, duplicates, missing data, contamination with other programs, etc.

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

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

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

ADA Compliant: 1/7/2021