RESEARCH REPORT

EVALUATION AND TRAINING INSTITUTE
www.eticonsulting.org

AN EVALUATION AND TRAINING INSTITUTE REPORT

Utah's Early Intervention Reading Software Program

2020-2021 Program Evaluation Findings

Submitted to the Utah State Board of Education
October 2021

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

ADA Compliant: 11/4/2021

ABOUT EVALUATION AND TRAINING INSTITUTE

Founded in 1974, the Evaluation & Training Institute (ETI) is a non-profit consulting firm, headquartered in Los Angeles, dedicated to working with schools, post-secondary institutions, public agencies, private foundations, community-based organizations and professional organizations. We specialize in third-party program evaluations covering many fields, including education, literacy, STEM, social services, health and prevention. Many of our evaluations have been instrumental in the development of public policy as well as state and federal legislation. Throughout, our focus is on helping clients improve their programs as well as maintain accountability to funders and oversight committees.

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

The Evaluation and Training Institute (ETI) thanks Melanie Durfee (Specialist, Digital Teaching and Learning) and Jimmy Hernandez 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 Alli Yeager from Imagine Learning, Haya Shamir from Waterford, Sarah Franzén and Kelley Pasatta 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 Intervention Software Program (EISP) was designed to increase the literacy skills of all students in K-3 through adaptive computer-based literacy programs. The program provided Utah's Local Education Agencies (LEAs) with an option to select among four adaptive computer-based 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) students' use of the program during the school year (program implementation); and 2) the effects of the program on increasing students' literacy achievement (program impacts). The current evaluation investigated the impact of the software programs across all four vendors (program-wide) and also the impact of each individual program (vendor-specific). This report captures all program-wide results. The vendor-specific findings can be found in separate, supplemental memos submitted along with this report.

Program Enrollment and Implementation

During the 2020-2021 school year, EISP was implemented in 142 LEAs and to 158,695 students throughout the state of Utah. The proportion of students using the individual vendor's software is a similar pattern to previous years. Core5 was used by the most students (97,566), followed by Imagine Learning (34,394), i-Ready (19,455), and Waterford (7,280). State-wide program implementation provided the opportunity for large numbers of students to receive program benefits, however, it was important for students to use the program for the intended amount of time (set by program vendors) in order to see the impact on students' literacy achievement.

Each year, program vendors provide LEAs with recommendations on weekly minutes, 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 met each vendors' minimum recommendations for use (evaluating both total weeks and weekly minutes). Using the most stringent definition of use, we looked at students who met or exceeded the week recommendation and who also met or exceeded the recommended number of minutes for the weeks they used the progam. Using this strict definition, we found very few students (regardless of vendor) who were able to engage at this level. This led us to expand the program use definitions for our impact analysis.

Program-wide Impact

After examining the implementation of the program, we studied the effectiveness of the EISP on literacy achievement. We took what we learned from the implementation study and compared several different groups of students. Most broadly, we examined the impact of the program on students who used the software vs. students who did not. The EISP students

were categorized into 3 subgroups (1) those who used the software in any amount (Intent to Treat or "ITT"), (2) students who used the software for at least 80% of the minimum weeks and 80% of the average weekly minutes, and (3) students who used the software for the recommended number of weeks and met the average weekly minutes. Our impact analysis considered all three subgroups in order to capture a more representative sample of program students and to provide a more realistic approach to how a larger population of students were actually able to engage with the program, given potential constraints on students' 'seat time'. Lastly, we looked at program impacts across specific types of students including those classified as low-income, special education, English Language Learners, or those attending a Title 1 school.

Literacy achievement was measured using the state provided Acadience Reading scores. We found statistically significant treatment effects for all grades (K-3) and across the three program usage levels. That is, students using the software program in any amount, scored higher at the end-of-year on measures of literacy compared to students not served by the program. Effect sizes (calculated using Hedges G) were used to describe the magnitude of the program impact and were interpreted as meaningful if they reached a minimum threshold of 0.26. Kindergarten had the highest effect size among all grades studied and was the only grade to surpass the 0.26 threshold. For kinder students using the software for the recommended weeks and average minutes or at least 80% of the recommendation, the program had a meaningful impact (effect sizes; 0.34 and 0.29, respectively).

Program Usage and Program Impacts

We also examined how treatment students compared to each other based on the 3 different levels of program use. Our findings indicate, across all grades, that students adhering closest to the vendors' recommendations for use (including average weekly minutes and total weeks), achieved higher mean reading composite scores at the end-of-year.

Student Characteristics and Program Impacts

Also of interest, was how the program may benefit students in specific demographic subgroups. We conducted a separate analysis of program impacts on students identified as English Language Learners, low-income, special education designation status, or those who attended a Title 1 school. Across all grades and for every subgroup, students in the EISP outperformed their non-program counterparts. The differential treatment effects were most pronounced in kindergarten, but still show positive impacts in end-of-year literacy scores for first, second and third grade students.

Discussion & Recommendations

First, we want to acknowledge the pandemic's lasting impact on the 2020-2021 school year. Delivery models for the EISP may have been altered to accommodate the unprecedented modifications made to instruction in some districts or schools across the state. We are aware that not all children had the same in-person or virtual classroom experience, making this program year unique from prior years.

Despite the limitations caused by Covid-19, we identified positive student literacy achievement outcomes, specifically for students who either met vendors' recommendations for weeks and average minutes or met at least 80% of the recommendations for use. Our findings underscore the importance of meeting minimum thresholds as well as the benefits of consistent program use from week-to-week.

A notable portion of EISP students were unable to meet the minimum use recommendation as defined by the software vendors. We therefore recommend that the state encourage consistency of use and continue to hold LEAs accountable for meeting vendors' recommendations so that students are provided the best opportunity to strengthen their literacy skills. We also recommend that future evaluations continue to explore the ways in which usage at different levels impacts literacy skill development and work to identify engagement patterns ideal for the skills acquired in each grade.

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 with students in kindergarten through the third grade. To participate in the Early Intervention Software Program (EISP), Local Education Agencies (LEAs) submit 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 2020-2021. 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 the results for both the combined impact of all the software programs used in Utah schools (program-wide) as well as the individual impact on literacy achievement for each of the software providers (vendor-specific). This report highlights the program-wide findings only. The vendor-specific results can be found in supplemental memos provided to USBE separate from this report.

The current evaluation includes findings from the 2020-2021 academic year, the EISP's eighth year of implementation. These findings are intended to help the USBE and Local Education Agencies (LEAs) understand how the program is working, to identify potential areas for program improvement, and to make evidence-based decisions about future iterations of the program.

In light of the Covid-19 pandemic, the 2020-2021 implementation and impact findings should be considered within the context of possible disruption to in-person instruction. We know that school districts were impacted differently and that not all students shared the same instructional delivery methods for the entire school year.

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

1. To what extent did students use the software program as intended?
2. How did the EISP impact students' literacy across all vendors?
3. How did different levels of program usage affect program impacts?
4. What impact did EISP have on specific student populations?

The sections of this report include this year's program enrollment numbers across grade and vendor, program implementation findings including vendor recommendations and

participants' ability to meet them, the impact that the EISP had on literacy achievement, and the impact that different amounts of program use have on literacy outcomes. The report also shows the impact that the EISP has on specific populations of students including English Language Learners, those classified as low-income, special education, or Title 1 status. We summarize the key findings and study limitations in the final sections. A detailed summary of our research methods is included in **Appendix A**.

Program Enrollment

In 2020-2021, the four EISP software vendors were used in 142 LEAs, in 605 schools and by 158,695 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[1]. As outlined in **Table 1**, Core5 was the most widespread program in the state compared to the other EISP providers, reaching 56 LEAs, 335 schools, and 97,566 students.

Table 1. 2020-2021 Program Enrollment Overview

Program	LEAs	Schools	Students (K-3)
Core5	56	335	97,566
Imagine Learning	42	145	34,394
i-Ready	23	72	19,455
Waterford	21	53	7,280
Total	142	605	158,695

Data source: software vendor data.

Student participation by grade varied by program. Imagine Learning, Core5, and i-Ready had an even distribution of students across grades K-3, while Waterford was used more frequently in earlier grades (**Table 2**).

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

Table 2. 2020-2021 Program Enrollment by Grade

Program	Kinder	1st	2nd	3rd
Core5	22,169	25,848	25,359	24,190
Imagine Learning	8,213	9,253	9,086	7,842
i-Ready	4,433	4,840	5,107	5,075
Waterford	2,916	2,506	1,858	--
Total	37,731	42,447	41,410	37,107

Data source: software vendor data in K-3, Waterford 3rd grade had less than 5 students.

Program Implementation

Studying program implementation prior to measuring the program impact provided a better understanding of the way the program was ultimately used by students. Namely, students must use the program long enough to influence the outcomes under study. Critical to successful EISP implementation was the amount of time and how consistently a student used the program during the school year.

Each vendor provided recommendations for the amount of time that students should use the software program during the year, to have an impact on literacy achievement. As shown in **Table 3**, these recommendations differed by grade and by vendor.

Table 3. Vendor 2020-2021 Minimum Use Recommendations

Program	Kinder-garten	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	50 min/week	50 min/week	50 min/week	18 weeks
i-Ready	30 min/week	30 min/week	30 min/week	30 min/week	20-25 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.

Each software provider communicated both a range of minutes per week, and a minimum number of weeks for students to use the program. Across vendors, recommended weekly use ranged from 20 minutes to 80 minutes per week and total weeks ranged from 18 to 28 weeks.

There are various ways to measure how students used the program. **Table 4** presents a comprehensive summary of average usage for each vendor and grade. These numbers represent the overall average of all students in their respective grade, and include average weekly minutes of use, average total minutes of use, and average number of weeks of use through the end of the school year.

Table 4. Program Use by Vendor and Grade

Program	Grade	N	Ave Weekly Min.	Ave Total Min.	Ave Wks of Use
Core5	K	22,169	52	1,311	23
	1	25,848	61	1,785	28
	2	25,359	58	1,674	28
	3	24,190	52	1,415	26
	Total	97,566	56	1,557	26
Imagine Learning	K	8,213	48	1,288	25
	1	9,253	55	1,566	27
	2	9,086	52	1,470	27
	3	7,842	47	1,175	23
	Total	34,394	51	1,385	26
i-Ready	K	4,433	33	713	20
	1	4,840	43	1,191	26
	2	5,107	45	1,258	27
	3	5,075	45	1,142	25
	Total	19,455	42	1,087	25
Waterford	K	2,916	49	1,282	25
	1	2,506	58	1,675	27
	2	1,858	52	1,348	24
	3	-	-	-	-
	Total	7,280	53	1,434	25

Data source: K-3 vendor usage data after cleaning invalid SSIDs, duplicates and missing data

The data above represent the averages among all students who engaged with the EISP program (Intent to Treat) and should be viewed as descriptive in nature not as a measure for meeting recommended program use.

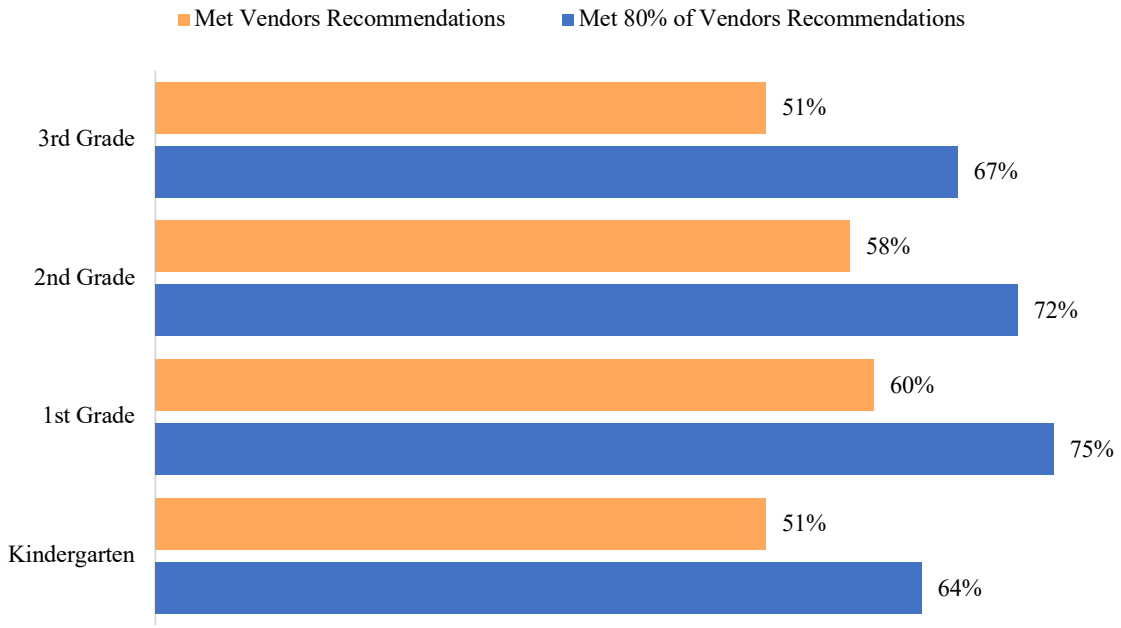
To what extent did students use the software program as intended?

We studied the extent to which students were able to satisfy both requirements of usage; meeting both the weekly minutes and also meeting the total weeks. Using the most stringent definition of use, we looked at students who met or exceeded the week recommendation and who also met or exceeded the recommended number of minutes for each week they used the program. Using this strict definition, we found very few students (regardless of vendor) who were able to engage at this level.

With that, we created two additional definitions of use in order to more realistically capture students' program participation. Our goal was to align as closely as possible to the vendor's stated criteria for use. First, we calculated the percentage of students in each grade who met the total weeks as recommended by the vendor *AND* whose average weekly minutes (for those weeks) was at or above the recommended minimum. Throughout this report we refer to this group of students as "met vendors' recommendation." We found that approximately half of kindergarteners and 3rd graders and about 60% of 1st and 2nd graders were able to adhere to the recommended weeks AND average weekly minutes (**Figure 1**; orange bars).

Next, we calculated the percent of students who met at least 80% of the vendors' total week recommendation and met at least 80% of the average weekly minutes recommendation. We refer to this group of students as "met 80% of vendors recommendation." While this expanded the vendors' stated criteria for use, it increased the representativeness of the children we studied, and provided a more realistic approach to how a larger population of students were actually able to engage with the program, given potential constraints on students' 'seat time.' As illustrated in Figure 1 (blue bars), this adjustment increased the overall percentage of program students by nearly 15% across all grades.

Figure 1. Percentage of Students Meeting EISP Recommendations for Use



Note: Met Vendors Recommendations reflects 'Met minimum weeks and *average* weekly minutes'
Met 80% of Vendors Recommendations reflects 'Met 80% of weeks and 80% of *average* weekly minutes'

We included both of these use groups in our impact evaluation.

Program Impacts on Literacy Achievement

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[2]. We studied how the program impacted literacy achievement by comparing groups of students who used the program to groups of students who did not. We have included a detailed methods section for technical reviewers in **Appendix A**.

Program Impacts

To fully understand the extent to which the software program affected children's end-of-year literacy scores, we created control groups of students who did not participate in EISP and who also matched the program students on important factors such as beginning-of-year literacy scores and key demographics.

Based on what we learned from our implementation study, we created several analytic samples of program students (treatment) and non-program students (control)[3]. We

[2] For vendor specific findings, please refer to the supplemental memos with individual vendor results.
[3] Analytic samples refer to the those classified as our treatment and matched control students and whose data were used in our analysis.

considered three usage levels in creating the samples for program students and then matched each one to a control group on test scores and demographics. The three analytic samples were as follows, (1) program students who used the program in any amount throughout the program year (Intent to Treat, ITT) and their matched control counterparts, (2) program students who met at least 80% of the vendors' recommended weeks and average weekly minutes and their matched control counterparts, and (3) program students who met the total weeks and weekly average as defined by the vendor and their matched control counterparts. For more detail on our statistical matching process, please refer to **Appendix A**.

How did the EISP affect students' literacy across all vendors?

We conducted ordinary least squares (OLS) regression, to determine if those participating in the EISP had statistically higher literacy scores at the end of the year compared to those not in the program. We examined treatment effects for each analytic sample (based on their usage) and found that for all grades (K-3), the predicted literacy mean scores were higher for students participating in the EISP (regardless of usage level) than for those who did not. That is, students using the software program in any amount, scored higher at the end-of-year on measures of literacy compared to students not served by the program. **Table 5** presents the treatment and control group mean scores across all three usage levels by grade. As shown, the predicted mean scores for students using the software in any amount (ITT), were statistically higher than their control counterparts with differences ranging from 2 to 7 points.

Table 5. Acadience Predicted Means by Usage and Grade; Treatment and Control

Grade	Condition	Intent to Treat	Met 80% of Rec.	Met Rec.
		End-of-Year Predicted Mean Scores		
K	Treatment	140.11	148.79	150.38
	Control	132.62	136.58	136.95
(diff)		*7.49*	*12.21*	*13.43*
1	Treatment	173.75	183.19	187.81
	Control	171.27	176.65	172.83
(diff)		*2.48*	*6.54*	*14.98*
2	Treatment	256.02	267.77	269.90
	Control	250.77	259.86	262.90
(diff)		*5.25*	*7.91*	*7.00*

Grade	Condition	Intent to Treat	Met 80% of Rec.	Met Rec.
3	Treatment	377.67	392.76	398.35
	Control	375.14	385.37	387.65
(diff)		*2.53*	*7.39*	*10.7*

Data source: Matched K-3 ITT, MRU80, MRU samples. All mean comparisons displayed between treatment and control were statistically significant at p≤ .05.

Additionally, the differences between students who used the EISP at 80% of the vendors' recommendation or who met the recommendation, show even greater differences with predicted mean scores up to 15 points higher than their matched control students[4]. Generally, the largest differences between treatment and control students are seen in the younger grades and among the highest use groups.

In addition to mean scores, we looked at the effect sizes (ES) between treatment and control students. Effect sizes 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). Additional information on effect sizes can be found in **Appendix D**.

Table 6. Effect Sizes by Grade and Usage Level

Grade	Condition	Intent to Treat	Met 80% of Rec.	Met Rec.
			Effect Sizes	
K	Treatment	0.183	**0.289**	**0.341**
	Control			
1	Treatment	0.036	0.092	0.226
	Control			
2	Treatment	0.085	0.123	0.125
	Control			

[4] See Appendix C for complete table of analytic statistics

Grade	Condition	Intent to Treat	Met 80% of Rec.	Met Rec.
3	Treatment	0.037	0.106	0.167
	Control			

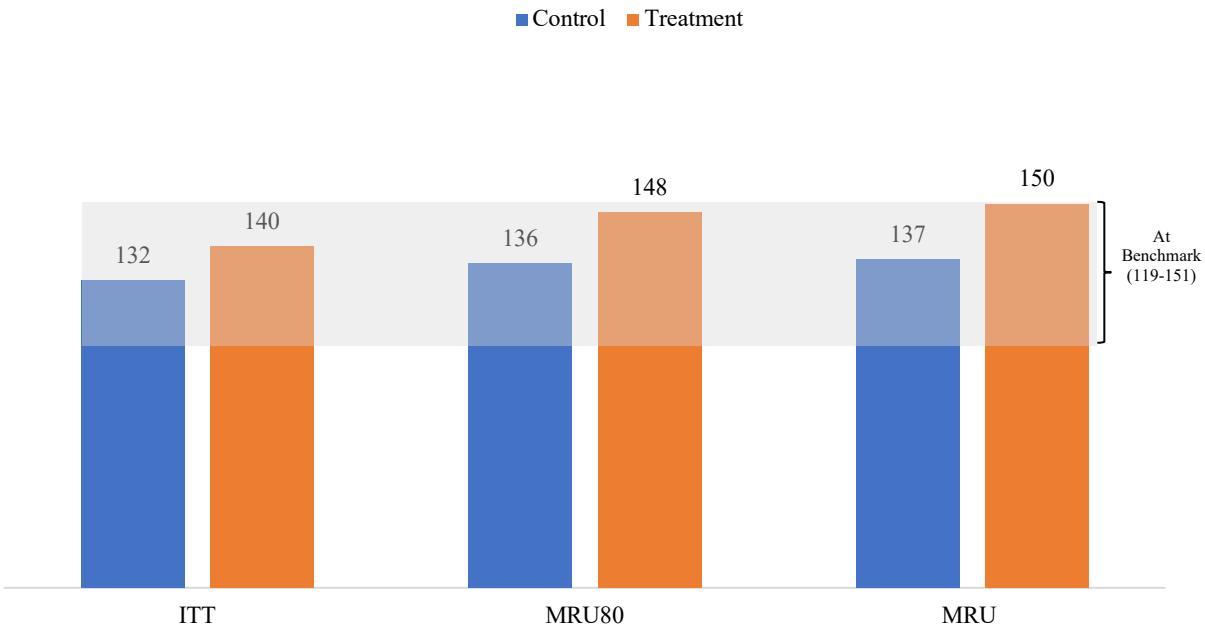
Data source: Matched K-3 ITT, MRU80, MRU samples. All effect sizes displayed were statistically significant at p≤ .05. Bold = ES above the 0.26 threshold.

Table 6 shows the most meaningful program impact was on kindergarten students who were able to meet the vendors recommendations for use (ES 0.34) followed by those who met at least 80% of the recommendation (ES 0.29). First graders in the top usage group approached the meaningful threshold for effect size but fell just short (ES 0.23). Despite significant predicted mean differences, all other grades and usage levels had effect sizes below the .26 threshold.

Program Impacts in Context

Given the uniqueness of the past academic year and the disruptions that have been caused due to the global pandemic, it is equally important to understand not just how the EISP impacted students' progress this year, but also how the entire student population performed relative to grade level expectations. The following graphs depict not only the increased performance of the EISP students, but also provide evidence that all students generally performed as expected for grade level regardless of program participation.

Figure 2. Kindergarten Predicted Means Scores by Usage Level and Matched Sample

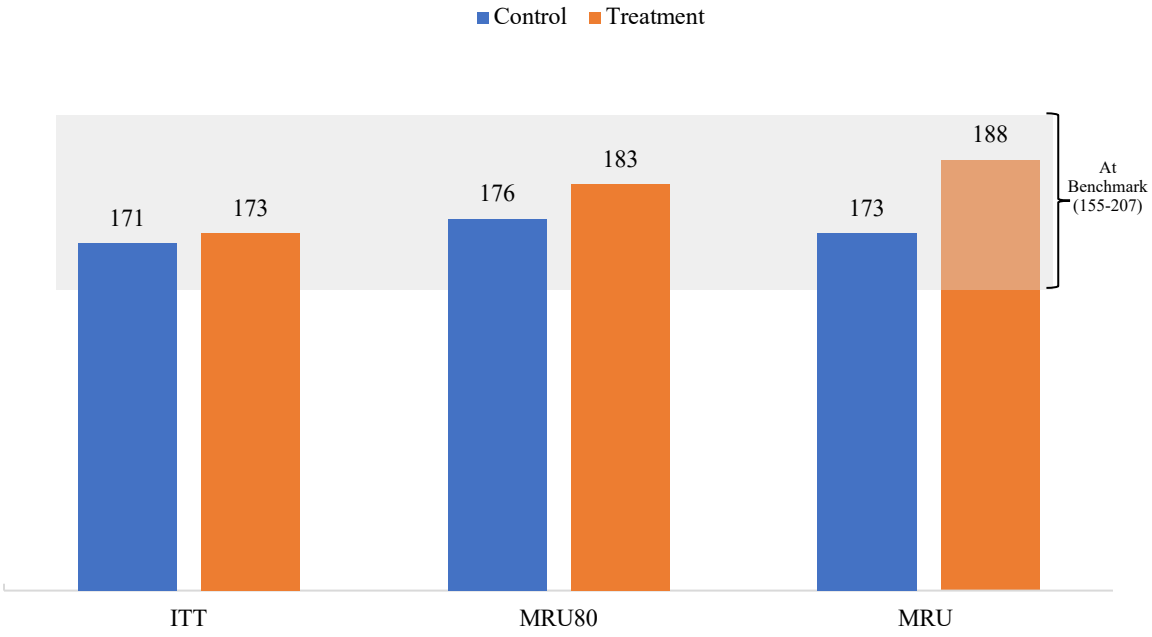


Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendation
*Kindergarten sample size –ITT n=37,972 (ctrl= 6,779, tr= 31,193); MRU80 n=26,763 (ctrl= 6,722, tr= 20,041); MRU n=31,496 (ctrl= 15,424, tr= 16,072; Students scoring **At Benchmark** (119-151) or **Above Benchmark** goal (152 or greater) have the odds in their favor (approximately 80% to 90% overall)of achieving later important reading outcomes. Data source: Matched K-3 ITT, MRU80 and MRU samples. All mean comparisons displayed in the table were statistically significant at p≤ .05.*

Figure 2 presents the predicted end-of-year mean scores for kindergarten students who used the EISP at different levels, side by side with their matched control counterparts. Students in the highest usage group (those that met the weeks and average weekly minutes) had the highest end-of-year mean score (150), putting them in the "at benchmark" score range. Further supporting that when the program is used consistently, students receive the highest program benefits.

That said, the end-of-year mean scores for all kindergarten students depicted here (both treatment and control) show literacy performance within expected levels for their grade.

Figure 3. First Grade Predicted Means Scores by Usage Level and Matched Sample

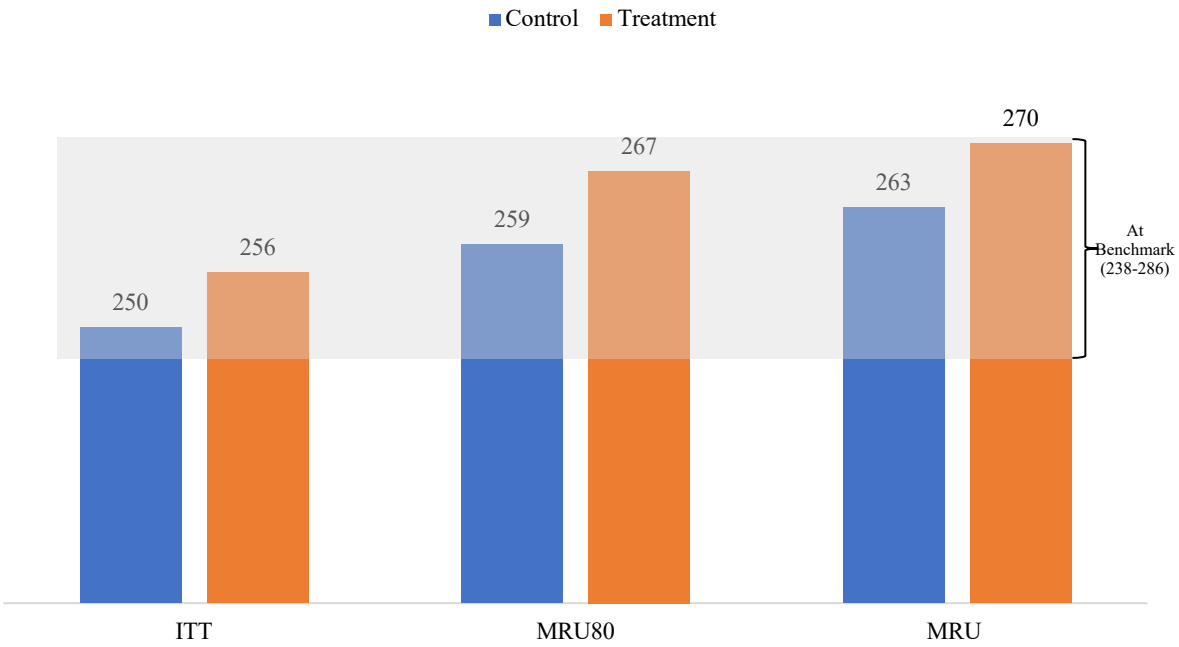


Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendiation
*First Grade sample size- ITT n= 40,457 (ctrl= 5,475, tr= 34,982); MRU80 n= 31,554 (ctrl= 5,460, tr=26,094); MRU n=35,574 (ctrl= 14,159, tr= 21,415); Students scoring **At Benchmark** (155-207) or **Above Benchmark** goal (208 or greater) have the odds in their favor (approximately 80% to 90% overall)of achieving later important reading outcomes. Data source: Matched K-3 ITT, MRU80 and MRU samples. All mean comparisons displayed in the table were statistically significant at p≤ .05.*

Figure 3 presents the predicted end-of-year mean scores for first grade students who used the EISP at different levels, along with their matched control counterparts. Similar to kindergarten, students who used the program closest to the vendors' intention, had the

highest end-of-year mean score (188). While all treatment groups outperformed their control counterparts, all first graders averaged literacy levels within the expected range.

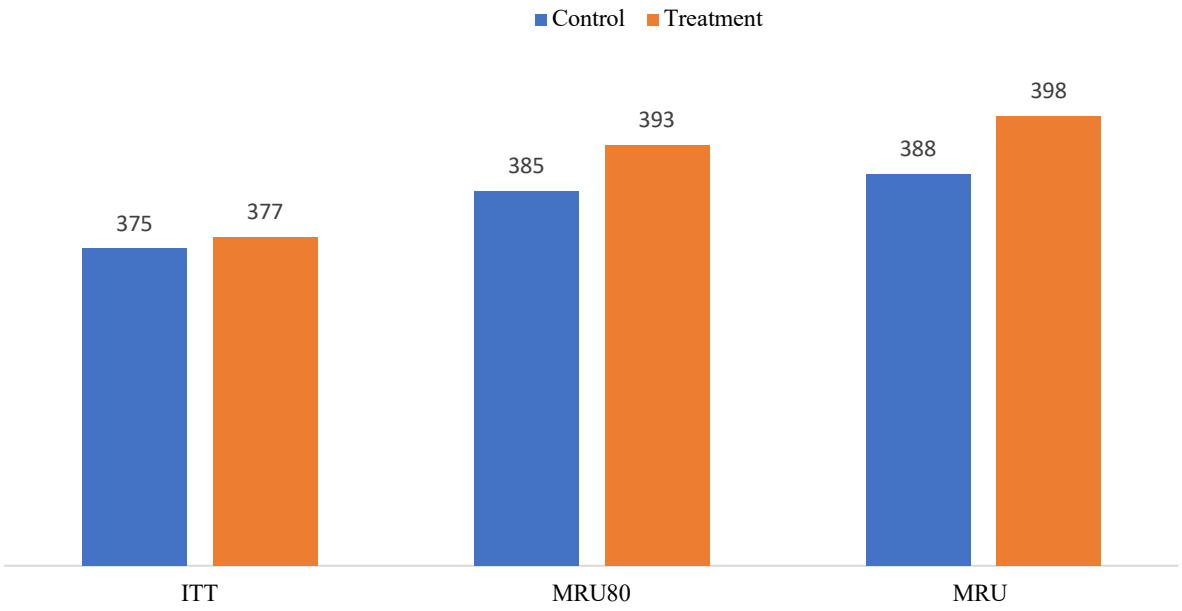
Figure 4. Second Grade Predicted Means Scores by Usage Level and Matched Sample



Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendiation
Second Grade sample size - ITT n= 41,761 (ctrl= 7,366, tr= 34,395); MRU80 n=32,189 (ctrl= 7,346, tr=24,843);MRU n=34,791 (ctrl=14,540, tr= 20,251); Students scoring At Benchmark (238-286) or Above Benchmark goal (287 or greater) have the odds in their favor (approximately 80% to 90% overall)of achieving later important reading outcomes. Data source: Matched K-3 ITT, MRU80 and MRU samples. All mean comparisons displayed in the table were statistically significant at p≤ .05.

Figure 4 presents the predicted end-of-year mean scores for second grade students who used the EISP at different levels. As with the younger grades, students who used the program with more consistency, had the higher end-of-year mean scores (267 and 270, respectively). Again, while all treatment groups outperformed their control counterparts, all second graders averaged literacy levels within the expected range.

Figure 5. Third Grade Predicted Means Scores by Usage Level and Matched Sample



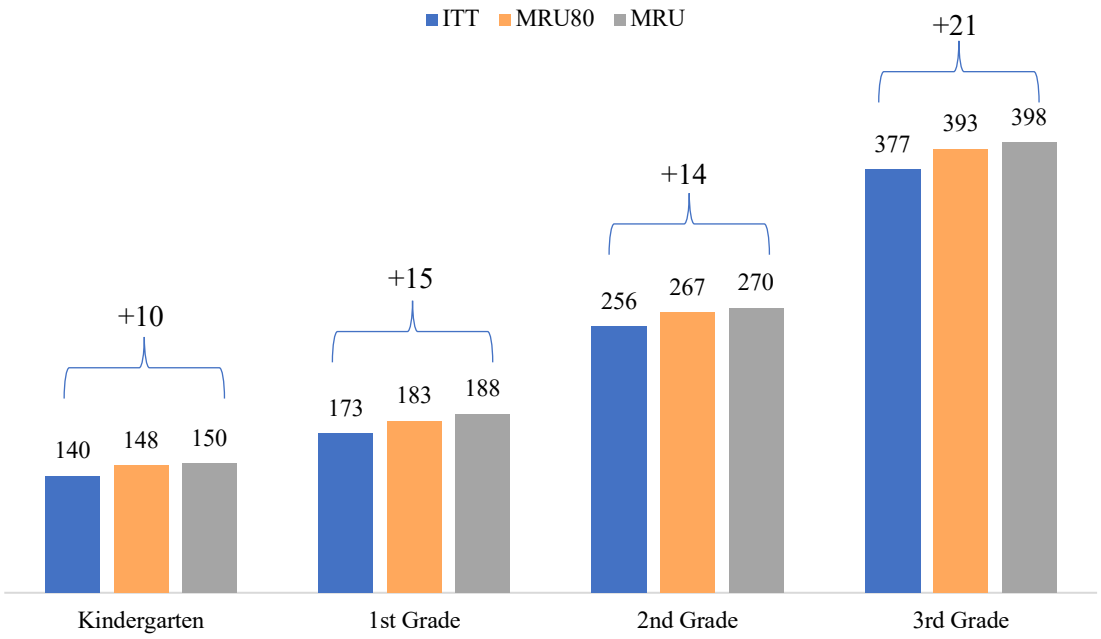
Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendiation
*Third Grade sample size – ITT n= 40,364 (ctrl=10,348, tr=30,016); MRU80 n=30,349 (ctrl=10,314, tr=20,035; MRU n=30,229 (ctrl= 14,851, tr=15,378); Students scoring **At Benchmark** (330-404) or **Above Benchmark** goal (405 or greater) have the odds in their favor (approximately 80% to 90% overall)of achieving later important reading outcomes. Data source: Matched K-3 ITT, MRU80 and MRU samples. All mean comparisons displayed in the table were statistically significant at p≤.05.*

Figure 5 presents the predicted end-of-year mean scores for third grade students. The pattern held true, students who used the program with more consistency, had the higher end-of-year mean scores (393 and 398, respectively). Again, while all treatment groups outperformed their control counterparts, all third graders averaged literacy levels within the expected range.

How did different usage levels effect program impacts?

Our evaluation sought to show differences between treatment and control students, but equally important was understanding how different levels of program participation within the treatment group, impacted literacy. **Figure 6** shows a side-by-side view of each grade and the three defined usage levels among treatment students (1) any use, ITT, (2) 80% of the recommendation, and (3) recommendation met for weeks and average minutes. The data suggest that as usage of the program increased within each grade (i.e. more adherence to the way program use was intended), predicted end-of-year mean scores also increased.

Figure 6. Predicted Means Scores by Grade and Usage Level



Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendiation

Interestingly, while using the program as recommended appeared to be the most beneficial usage level regardless of grade, treatment students in third grade benefited the most by meeting both average weekly minutes and total weeks (mean = 398), compared to third graders who engaged with the program more casually (mean = 377). That is, for treatment students in third grade, engaging with the software more closely to how it was intended showed the greatest increase to literacy outcomes compared to less consistent use.

What impact did EISP have on specific student populations?

We were also interested in studying how the program may benefit students in specific demographic subgroups. We conducted a separate analysis of program impacts on students identified as English Language Learners, low-income, special education designation status, or those who attended a Title 1 school. **Table 7** presents the Acadience Reading composite mean scores.

Table 7. Subgroup Analysis of Predicted End-of-Year Mean Scores

		Kindergarten	First Grade	Second Grade	Third Grade
Special Education	Treatment	133.70	170.95	247.90	365.40
	Control	121.49	164.41	239.99	358.01
ELL	Treatment	145.04	179.96	250.72	376.26
	Control	132.83	173.42	242.81	368.87
Low-Income	Treatment	146.14	173.74	261.88	386.17
	Control	133.93	167.20	253.96	378.77
Title 1	Treatment	152.84	180.08	265.46	392.17
	Control	140.63	173.54	257.55	384.78
	Data source: Matched K-3 MRU80 sample. All data points displayed in figure were statistically significant at $p \leq .05$.				

Across all grades and for every subgroup, students in the EISP who were able to meet at least 80% of the vendors' recommended use criteria, outperformed their non-program counterparts. The differential treatment effects were most pronounced in kindergarten, but still show positive impacts in end-of-year literacy scores for first, second and third grade students.

DISCUSSION, LIMITATIONS, AND RECOMMENDATIONS

There were two primary goals for the 2020-2021 EISP evaluation, 1) to study program implementation as defined by vendors' software use recommendations and, 2) to determine the impacts of the program on students' literacy achievement. We summarize here those findings and present the known limitations as well as our recommendations for improvement.

Implementation

Each year, program vendors provide LEAs with recommendations on the amount of time the program should be used by students. These usage recommendations varied across grade and software vendor. To guage successful implementation from a program-wide perspective, the implementation study was designed to determine the extent to which students met each vendors' minimum recommendations for use. We found that approximately half of kindergarteners and 3rd graders and about 60% of 1st and 2nd graders were able to meet both total weeks and average weekly minutes as vendors' recommended, regardless of software program.

A notable number of students, therefore, missed the threshold as suggested by program vendors. In the context of the evaluation during the ongoing challenges faced from the COVID-19 pandemic, it is difficult to measure whether the usage challenges were due to disruptions to in-person learning or possible changes in delivery method for the program. Not all districts and schools were able to offer in-class instruction for the full school year, and as a result of the pandemic many were forced to use virtual teaching. This may have disrupted their ability to track and ensure that students used the software programs as planned. That said, regardless of why minimum use requirements could not be met by all students, the data suggest the importance of helping students use the program consistently to benefit their year-end literacy scores.

We included several different use groups in our impact analysis to help stakeholders understand the effect that program use had on student outcomes. We studied all students who used the program in any amount, those meeting at least 80% of the vendors' recommended weeks and average minutes, and those who met the recommended weeks and average minutes.

Impacts

We identified positive student literacy achievement outcomes across all three usage groups as compared to matched groups of similar students who did not use the program. The difference in outcome scores were most pronounced for students who either met vendors' recommended weeks and average minutes or met at least 80% of the recommendations for use. We further explored this pattern within the treatment students only and found a link between more consistent program use and stronger program effects. That is, we saw an

increase in literacy scores across grades K-3 as program use more closely adhered to the recommendations.

Additionally, the EISP was shown to have strong benefits for students classified as English Language Learners (ELL), special education, low-income, and Title-1 as compared to matched counterparts not served by the program.

Limitations

Pandemic Year Influences. Delivery models for the EISP may have been altered to accommodate the unprecedented modifications made to instruction in some districts or schools across the state. We are aware that not all children had the same in-person or virtual classroom experience, making this program year unique from prior years. This year districts and schools were differentially impacted by closures or shifts in instructional methods due to spikes in Covid-19 cases. It is possible that certain vendors served districts or populations of students that were more negatively impacted by the pandemic. It is also possible that not all students in our evaluation were administered the state-wide Acadience end-of-year assessment in the same format. While we are aware that these events and circumstances can impact the engagement and outcomes with the EISP across the school year, we acknowledge that we were unable to control for all possible scenarios in our analysis.

Individual Teacher Influences. As a classroom based intervention, the variability in the way teachers implement the program plays a role in our ability to determine and understand program-wide impacts. With more than a hundred thousand students participating across thousands of classrooms, we are unable to control for the extent to which different teachers actively support students' use of the software. This is particularly true for a year where some may have experienced unexpected disruptions. More detailed information about the way in which teachers are implementing the intervention could shed light on the usage data that we analyze and the impacts that we measure.

Comparison Students. Lastly, we know that the use of digital technology in educational interventions is on the rise in the state of Utah. Therefore, the number of students exposed to and leveraging these software programs increases every year. Our control students are made up of children not participating in the EISP, however, with the growing prevelance of educational technology, it is possible that some of the control students may have been exposed to different non-EISP reading interventions. Future evaluations would benefit from the USBE and vendors tracking and sharing this information.

Recommendations

The results of the evalution underscore the importance of supporting students' literacy development and creating opportunities for our youngest learners. Students served by the

EISP outperformed the students who were not. Further, the students who were able to engage with the software as it was intended by the vendors also showed greater end-of-year literacy scores relative to those participating more casually in the program. These benefits were seen across grades K-3.

Several recommendations surfaced from our findings:

- With evidence supporting consistency of use, we suggest that vendors identify and meet with LEAs who have usage below the recommended levels in order to cultivate ways to improve student engagement with the software.
- It is recommended that vendors emphasize the importance of meeting both aspects of the intended use recommendations: weekly minutes and total weeks. Results suggest that both components are critical to maximizing literacy outcomes.
- Beyond time spent with the software, it is recommended that the vendors consider alternatives to time (minutes/weeks) as a measure of program implementation. Not all 'seat time' is equivalent across students.
- As we navigate uncertainty with the possibility of future disruption to in-person instruction, vendors may consider providing stakeholders with alternative solutions if delivery methods need to pivot during the school year.
- Regular communication with schools may help keep all stakeholders apprised of instructional modifications happening at the school or district level that require vendor-related adjustments in real time.
- We recommend that future evaluations continue to explore the ways in which usage at different levels impacts literacy skill development and work to identify engagement patterns ideal for the skills acquired in each grade.

With intentional effort behind accountability, improving consistency of use, and the ability to remain agile, more and more students will benefit from the *Early Intervention Software Program*.

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. (2014-2020, October). *Early Intervention Software Program Evaluation: Results.* Culver City, CA

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.

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.

APPENDIX A. EVALUATION METHODS

The following is 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.

Program Participants

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. All students captured in the vendors' usage data were included in our implementation study. Our goal was to provide the most accurate depiction of students' program use, regardless of how much students engaged with the program. For K-3 students, we used the vendor data, and did not remove students with incomplete Acadience data.

Impact Study Evaluation Participant Samples

To study program impact, we created three different groups of treatment students based on their level of program usage, (1) those who used the software in any amount (Intent to Treat or "ITT"), (2) students who used the software for at least 80% of the minimum recommended amount, and (3) students who used the software as intended by the vendors including weekly minutes and total weeks. 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.

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; Iacus 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 samples, as indicated by the following three L1 scores,[5] ITT; 0.00000000000005417; MRU80; 0.00000000000008100 and MRU; 0.00000000000004161. Lower values indicate less imbalance, and the closer to zero the better the two samples were balanced across covariates.

To summarize, we created and matched three treatment and control samples based on three different levels of usage. The EISP students were categorized into 3 subgroups (1) those who used the software in any amount (Intent to Treat or "ITT"), (2) students who used the software for at least 80% of the minimum recommended amount, and (3) students who used the software as intended by the vendors including weekly minutes and total weeks. Each of these groups had matched control counterparts.

What sources of data were used in our analyses?

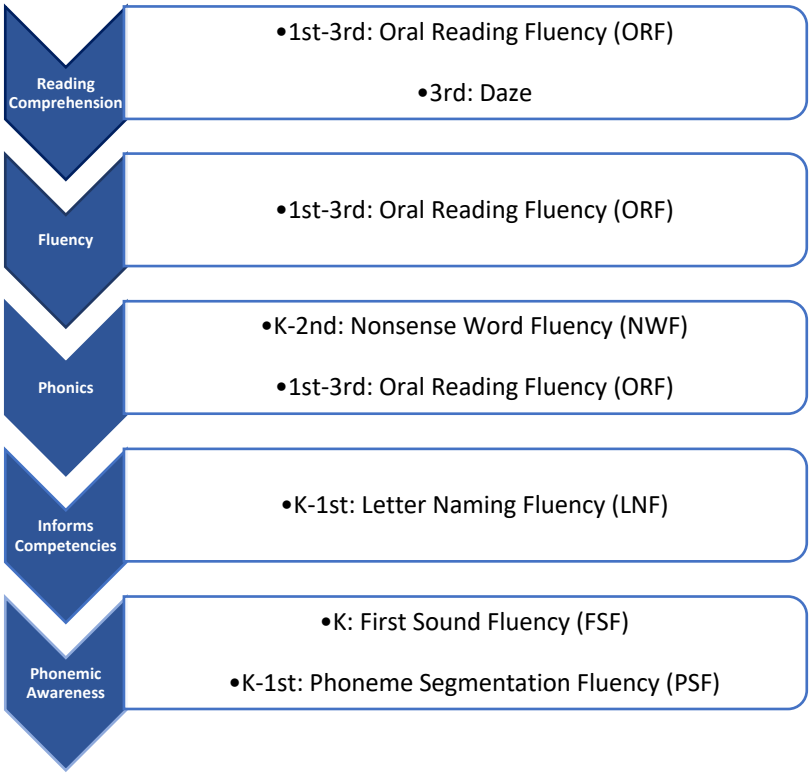
We collected data from nine different sources to create our master dataset for the EISP analyses. The data sources included: four program vendors, who provided us with usage information for each student who used their programs; state Acadience Learning (Acadience Reading) testing data; and student information 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.

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.

[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.

Figure A1: Acadience Indicator & Literacy Skill Measures



How did we study program implementation?

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

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

Our study relied on 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.

Treatment Outcome Descriptive Analyses

To present our findings in an intuitive and applicable context, we measured the differences in students' reading scores at the end-of-year based on different categories of program exposure, or use. Use categories ranged from any use (i.e. Intent to Treat) to the highest category of meeting vendors' minimum recommended use requirement. As a complement to our OLS regression (causal) analysis, we used the descriptive analysis to show the association between levels of program use and outcomes for all students in the program.

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 Hedges G) to help readers understand the magnitude of treatment effects. Presenting effect sizes enabled us to provide a standardized scale to compare results based on different samples and measure the relative strengths of program impacts.

When interpreting our findings, it is important to note that effect sizes can be used to measure the strength of program impacts in multiple ways. A commonly used method is Cohen's (1988) characterization of effect sizes as small (.2), medium (.5) and large (.8). However, recent studies have suggested using a more targeted approach for determining the magnitude of the program impacts. For example, Lipsey et. al (2012) suggested effect size comparisons should be based on "*comparable outcome measures from comparable interventions targeted on comparable samples*", and notes that effect sizes in educational program research are rarely above .3, and that an effect size of .25 may be considered large (pg. 4). 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.[6] More information on how we selected our ES benchmark is provided in **Appendix F**.

[6] 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: ANALYTIC 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	Caucasion	SPED	Low-Income	ELL	BOY Comp
K	31,823	49%	75%	9%	28%	6%	34.52
1	35,841	48%	75%	10%	30%	8%	111.26
2	35,108	48%	75%	12%	30%	9%	171.79
3	30,577	49%	75%	14%	30%	10%	257.77

Table B2. MRU80 Sample by Grade[7]

	Grade	N	Female	Caucasion	SPED	Low-Income	ELL	BOY Comp
Control	K	6,722	48%	74%	10%	29%	5%	33.69
	1	5,460	50%	72%	9%	30%	9%	113.08
	2	7,346	49%	76%	11%	30%	9%	179.64
	3	10,314	49%	78%	12%	29%	8%	269.10
Treatment	K	20,041	49%	78%	7%	27%	5%	37.37
	1	26,094	48%	78%	9%	28%	6%	114.96
	2	24,843	48%	77%	10%	28%	7%	180.76
	3	20,035	49%	77%	12%	29%	9%	269.71

[7] The matched sample had an L1 score of 0.000000000000008100. 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 [8]

	Grade	N	Female	Caucasion	SPED	Low-Income	ELL	BOY Comp
Control	K	6,779	47%	73%	10%	29%	6%	33.53
	1	5,475	50%	72%	9%	30%	9%	113.09
	2	7,366	49%	75%	11%	31%	9%	179.44
	3	10,348	49%	77%	12%	29%	8%	268.91
Treatment	K	31,193	49%	77%	8%	28%	5%	34.63
	1	34,982	48%	77%	9%	30%	7%	111.47
	2	34,395	49%	76%	11%	30%	8%	172.55
	3	30,016	49%	76%	13%	30%	9%	258.66

Table B4. MRU Sample by Grade [9]

	Grade	N	Female	Caucasion	SPED	Low-Income	ELL	BOY Comp
Control	K	15,424	49%	75%	10%	30%	7%	30.41
	1	14,159	49%	73%	12%	34%	9%	101.48
	2	14,540	49%	73%	14%	33%	10%	151.91
	3	14,851	49%	74%	16%	32%	11%	238.45
Treatment	K	16,072	49%	77%	7%	26%	5%	38.65
	1	21,415	48%	78%	9%	28%	7%	117.69
	2	20,251	48%	77%	10%	28%	7%	186.41
	3	15,378	49%	77%	12%	28%	9%	277.58

[8] The matched sample had an L1 score of 0.00000000000005417.
[9] The matched sample had an L1 score of 0.00000000000004161.

APPENDIX C. REGRESSION STATISTICS AND EFFECT SIZES BY SAMPLE

Table C1. ITT Regression Model, by grade

	Grade	Condition	N	P-value	Marginal Mean	St. Error	Diff.	ES
Intent to Treat	K	Treatment	31,193	0.00	140.11	0.23	7.49	0.183
		Control	6,779		132.62	0.53		
	1	Treatment	34,982	0.011	173.75	0.36	2.48	0.036
		Control	5,475		171.27	0.85		
	2	Treatment	34,395	0.00	256.02	0.32	5.25	0.085
		Control	7,366		250.77	0.80		
	3	Treatment	30,016	0.015	377.67	0.38	2.53	0.037
		Control	10,348		375.14	0.91		

Note. ES: Effect Size (based on Hedges G). ES's greater than .26, the average for similar intervention programs Data source: Matched K-3 ITT sample.

Table C2. MRU 80 Regression Model, by grade

	Grade	Condition	N	P-value	Marginal Mean	St. Error	Diff.	ES
Met 80% of Recommended Use	K	Treatment	20,041	0.00	148.79	0.29	12.21	0.289
		Control	6,722		136.58	0.57		
	1	Treatment	26,094	0.00	183.19	0.43	6.54	0.092
		Control	5,460		176.65	0.85		
	2	Treatment	24,843	0.00	267.77	0.39	7.91	0.123
		Control	7,346		259.86	0.80		
	3	Treatment	20,035	0.00	392.76	0.48	7.39	0.106
		Control	10,314		385.37	0.94		

Note. ES: Effect Size (based on Hedges G). ES's greater than .26, the average for similar intervention programs Data source: Matched K-3 MRU80 sample.

Table C3. MRU Regression Model, by grade

	Grade	Condition	N	P-value	Marginal Mean	St. Error	Diff.	ES
Met Recommended Use	K	Treatment	16,072	0.00	150.38	0.31	13.43	0.341
		Control	15,424		136.95	0.35		
	1	Treatment	21,415	0.00	187.81	0.45	14.98	0.226
		Control	14,159		172.83	0.51		
	2	Treatment	20,251	0.00	269.90	0.39	7.00	0.125
		Control	14,540		262.90	0.44		
	3	Treatment	15,378	0.00	398.35	0.51	10.70	0.167
		Control	14,851		387.65	0.58		

Note. ES: Effect Size (based on Hedges G). ES's greater than .26, the average for similar intervention programs Data source: Matched K-3 MRU sample.

APPENDIX D. DATA PROCESSING & MERGE SUMMARY

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

Software Program Data

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

When cleaning duplicate IDs within each vendors' data, we deleted cases that were the same student with different usage reported and kept any unique cases after removing exact replicas. We did not count weeks, or include minutes, when there were fewer than five minutes recorded in a given week. After removing these instances, we updated the usage variables, such as total minutes, to reflect the change in use, and then removed students who had fewer than five minutes of total use from the data. After we cleaned and processed the vendors data, the total count of students went from 165,361 to 158,695 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 4,889 cases) and any IDs that did not comply with the state student ID (SSID) format (901 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 153,806 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 2020-2021 participants were included in the data. The SIS data file consisted of 208,476 cases, of which approximately four percent were duplicate records. After cleaning the data of duplicates, our SIS data consisted of 199,426 records.

Acadience Reading Data

In 2020-2021, the USBE prepared and transferred an Acadience Reading data file (n=183,787). After cleaning the IDs (e.g. deleting missing IDs and IDs that were not in a valid format), removing duplicates and removing cases with missing outcome data, we were left with a master Acadience file containing 168,625 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 168,549 cases. Next, we merged our master vendor data into the Acadience and SIS data and removed duplicate cases between vendors. This left us with 135,306 complete treatment cases and 33,243 control cases.

Lastly, we identified (where possible) schools or students with program exposure, using one of the four program vendors through non-EISP funding. 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 163,479 cases, of which, 133,349 were treatment and 30,130 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 2020-2021. 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. However, for our largest sample of program students, the Intent

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.

to Treat (ITT) program-wide sample, there were more program students than control students. This automatically reduced the size of this particular sample.

APPENDIX E: 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
Composite Score	Acadience Composite Score is a combination of multiple Acadience scores	Overall estimate of reading proficiency	K-6

APPENDIX F: DETERMINING EFECT SIZE BENCHMARK

A commonly used metric for identifying the strength of treatment effects is Cohen's (1998) Z 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.

ETI calculated effect sizes using the standardized mean difference calculation known as "Hedges' g" based on What Works Clearinghouse recommendations (WWC, 2020). For group design studies, this effect size is defined as the difference between the mean outcome for the intervention group and the mean outcome for the comparison group. Our interpretation of effect sizes and student impacts is focused solely on the intervention's impacts on student achievement.

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.

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