THE UTAH STATE BOARD OF EDUCATION

Report to the Education Interim Committee

Utah's Early Intervention Reading Software Program Report

November 16, 2022

Melanie Durfee
Digital Teacher and Learning Specialist, USBE
Melanie Durfee@schools.utah.gov

Amber Wright
Software Initiatives Specialist, USBE
Amber.Wright@schools.utah.gov

Malia McIlvenna
Educational Specialist, USBE
Malia.McIlvenna@schools.utah.gov

David MacKay
Research Consultant, USBE
David.Mackey@schools.utah.gov

Stephanie Su
Research Consultant, USBE
Stephanie.Su@schools.utah.gov

STATUTORY REQUIREMENT

U.C.A. Section 53A-17a-167 requires the Utah State Board of Education (USBE) to submit a report on the final testing data regarding an interactive computer software program including student learning gains as a result of the interactive computer software program.

Utah's Early Intervention Reading Software Program Report

EXECUTIVE SUMMARY

The Early Intervention Software Program (EISP) was designed to increase the literacy skills of all students in K-3 through adaptive computer-based literacy software. The program provided Utah's Local Education Agencies (LEAs) with an option to select among four adaptive computer-based programs. State-wide program implementation provided the opportunity for large numbers of students to receive program benefits however, it was clear a notable portion of EISP students were unable to meet the minimum use recommendation as defined by the software vendors. It is therefore recommended that the state encourage consistency of use and continue to hold LEAs accountable for meeting vendors' recommendations to provide students the best opportunity to strengthen their literacy skills. The EISP was particularly impactful for kindergarteners. It is recommended that the state continue to explore the ways in which program participation can boost more advanced literacy skills for students.

RESEARCH REPORT

EVALUATION AND TRAINING INSTITUTE
www.eticonsulting.org

AN EVALUATION AND TRAINING INSTITUTE REPORT

Utah's Early Intervention Reading Software Program

2021-2022 Program Evaluation Findings

Submitted to the Utah State Board of Education
October 2022

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

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 Amber Wright (Software Initiatives 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, David Mackay, and Stephanie Su, research consultants, at USBE for all their efforts preparing and transferring the student data used for the analyses. Finally, the software vendor representatives played a key role in helping us understand their software programs, sharing their data, and working with us to prepare the data in a consistent and streamlined format. In particular, we give special thanks to Dezi Hendershot 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 software. 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 2021-2022 school year, EISP was implemented in 133 LEAs and to 155,222 students throughout the state of Utah. The proportion of students using the individual vendor's software reflected a similar pattern to previous years. Core5 was used by the most students (104,692), followed by Imagine Learning (35,640), i-Ready (9,383), and Waterford (5,507). 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). As demonstrated in the report, a sizable number of students were unable to meet the recommended minimum usage levels put forth by the software providers.

Program-Wide Impact on Acadience Achievement

We also studied the effectiveness of the EISP on end-of-year Acadience literacy achievement. 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) students who used the software for the recommended number of weeks and met the recommended average weekly minutes, (2) students who used the software for at least 80% of the minimum weeks and 80% of the average weekly minutes, and (3) those who used the software in any amount (Intent to Treat or "ITT"). Our impact analysis considered all three subgroups in order to capture a broad sample of program students. Lastly, we looked at program impacts across specific types of students including those classified as low-income, special education, or English Language Learners.

Literacy achievement was measured using the state provided Acadience Reading scores. We found for all students in kindergarten, first and third grade who met the recommended usage or 80% of the recommended usage, their predicted end-of-year Acadience scores were higher than their control counterparts. We found that treatment effects were largest for students who used the program as intended. 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 with a 0.27, and was the only grade to surpass the minimum threshold for a substantive effect.

Varied Program Usage and Literacy Outcomes

We also examined how different levels of program use (as defined above) influenced the way treatment students scored at the end of the year. Our findings indicate, across all grades, that students adhering closest to the vendors' recommendations for use, achieved higher predicted mean reading Acadience scores at the end-of-year.

EISP and Different Student Populations

We studied how the program may benefit students in specific demographic subgroups, such as English Language Learners, low-income, or special education designation status. Across kindergarten, first and third grade and for every subgroup, students in the EISP who met 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 and third grade students.

Recommendations

The current evaluation identified positive student literacy achievement outcomes, most notably for kinder students who met vendors' recommendations for weeks and average weekly minutes of 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.

- The EISP was particularly impactful for kindergarteners. We recommend that the state continue to explore the ways in which program participation can boost the more advanced literacy skills for students in the grades that follow.
- We also recommend that future evaluations continue to investigate the ways in which the EISP impacts students of all reading abilities so that the state can make informed decisions about the most optimal ways to support a population of students with diverse learning needs.

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, therefore, 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 were selected to provide software and training to schools through the EISP in 2021-2022. 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 2021-2022 academic year, the EISP's ninth 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.

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' Acadience scores across all vendors?
3. How did different program usage levels influence Acadience outcome scores?
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 including mean differences and effects sizes[1], 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, or special education. 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 2021-2022, four EISP software vendors were used in 133 LEAs, in 565 schools and by 155,222 students. As has been the case the last several years, Core5 was the most widespread program in the state relative to other EISP providers, reaching 61 LEAs, 358 schools, and 104,692 students (**Table 1**).

Table 1. 2021-2022 Program Enrollment Overview

Program	LEAs	Schools	Students (K-3)
Core5	61	358	104,692
Imagine Learning	39	134	35,640
i-Ready	21	44	9,383
Waterford	12	29	5,507
Total	133	565	155,222

Data source: software vendor data, some LEAs and schools use more than one software vendor

Generally, student enrollment was similar across grades K-3 for three of the four vendors, with Waterford enrolling more students in earlier grades (**Table 2**).

[1] 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.

Table 2. 2021-2022 Program Enrollment by Grade

Program	Kinder	1st	2nd	3rd
Core5	24,312	26,868	27,074	26,438
Imagine Learning	8,496	9,272	9,312	8,560
i-Ready	1,607	2,452	2,749	2,575
Waterford	1,758	1,699	1,323	727
Total	36,173	40,291	40,458	38,300

Data source: software vendor data in K-3

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 2021-2022 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

* Core5 usage recommendations are automatically adjusted based on student need. Students working below grade level are assigned usage recommendations greater than those working 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.

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. 2021-2022 Program Use by Vendor and Grade

Program	Grade	N	Ave Weekly Min.	Ave Total Min.	Ave Wks of Use
Core5	K	24,312	48	1,232	24
	1	26,868	58	1,776	29
	2	27,074	54	1,655	29
	3	26,438	51	1,492	28
	Total	104,692	53	1,547	28
Imagine Learning	K	8,496	43	1,173	26
	1	9,272	52	1,603	30
	2	9,312	47	1,417	29
	3	8,560	45	1,245	26
	Total	35,640	47	1,547	28
i-Ready	K	1,607	33	753	21
	1	2,452	41	1,095	26
	2	2,749	40	1,123	26
	3	2,575	43	1,025	24
	Total	9,383	40	1,025	25
Waterford	K	1,758	50	1,404	26
	1	1,699	58	1,678	28
	2	1,323	47	1,187	23
	3	727	44	1,100	23
	Total	5,507	51	1,396	25

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

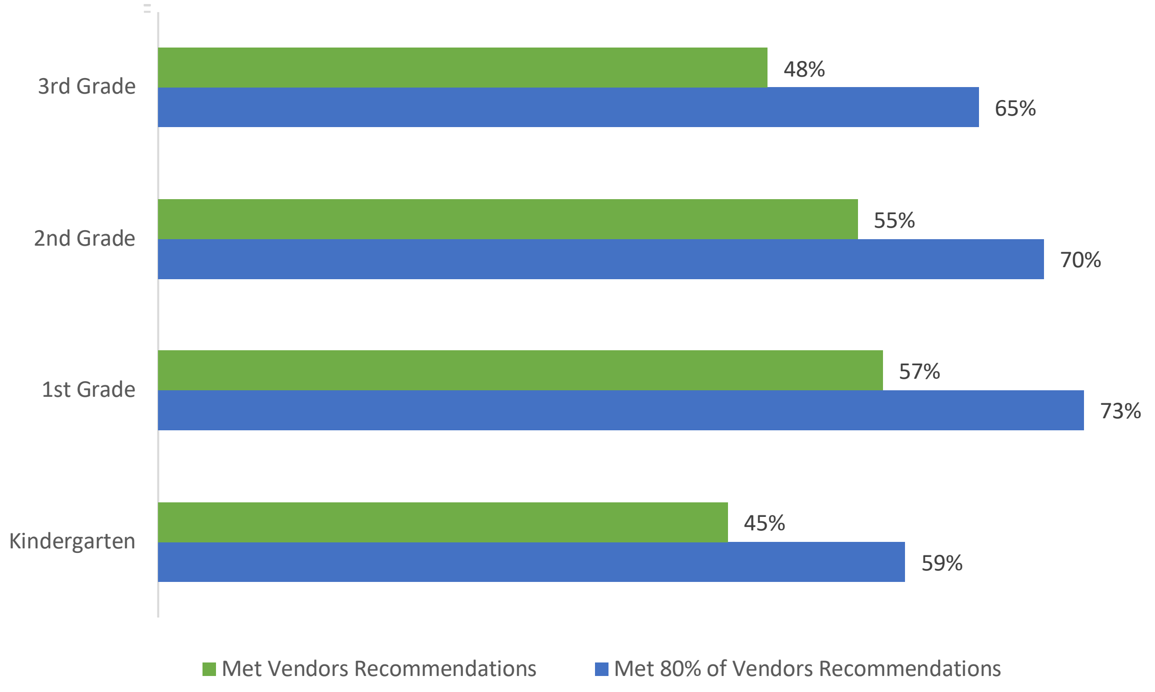
The data above represent the averages among all students who engaged with the EISP program (Intent to Treat) and should be viewed as descriptive in nature, not as a measure for meeting recommended program use.

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

Approximately 45% of kindergarteners and 48% of 3rd graders were able to adhere to the recommended weeks AND average weekly minutes, while just over half of 1st graders (57%) and 2nd graders (55%) met the vendor recommendations. (**Figure 1**; green bars).
This evaluation used two definitions of program use to capture students' EISP 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 participation varied among grades.

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 larger sample of students who engaged with the program. 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'

It warrants acknowledgement that just around half of the EISP student population achieved the levels of engagement put forth by the vendors. For the purposes of our impact evaluation, we analyzed both of the aforementioned groups of students.

Program Impacts on Acadience 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 students who used the program with students who did not. We have included a detailed methods section for technical reviewers in **Appendix A**.

Methods Summary

In order to study EISP's impact on Acadience literacy test scores, we needed two samples of students, those who participated in the program (Treatment group) and those who were matched to the treatment students across characteristics that influence learning, such as socio-economic status, demographic information and beginning-of-year Acadience test scores, but who did not participate in the program (Control group). The students who made

[2] Please refer to the individual supplemental memos for vendor specific results.

up our treatment and control groups, within each grade K-3, were considered our analytic samples (i.e. the samples we used in the analysis).

Among the overall treatment sample, we created three subgroups of students to account for different levels of program usage. These subgroups were created to evaluate how different levels of use influenced the program's impact on literacy achievement. We considered three main factors in creating the subgroups for EISP students: (1) students who met the minimum weeks and average weekly use recommendations as defined by each vendor*,* (2) students who met at least 80% of the recommended weeks and average weekly minutes, and (3) the broadest use group, inclusive of those who used the program in any amount throughout the program year (Intent to Treat).

We then matched comparison (control) students who did not participate in the program to the three EISP usage groups using Coarsened Exact Matching (CEM). We used CEM to match students on grade (including full day vs half day for kinder), beginning-of-year achievement scores and benchmark levels, gender, race, English Language Learner (ELL) status, and poverty status. The baseline characteristics of the treatment and control samples can be found in **Appendix A and B**. The matched samples were statistically well-balanced as indicated by L1 coeficients. For more detail on our statistical matching process, please refer to **Appendix A**.

Statistical Modeling of Program Impacts on Acadience Test Scores. Ordinary least squares (OLS) regression models were computed for each analytic sample. The OLS models predicted the differences in treatment and control groups' end-of-year group mean scores, while controlling for students' beginning-of-year (BOY) reading scores and key demographics; gender, race, ELL status, SPED designation and poverty status. We examined treatment effects for each analytic sample based on their usage and grade.

Results

Key Takeaway. Substantive treatment effects were found in kindergarten among the students who met the vendors' usage requirements (Hedges'g = 0.27). Students in grades 1 and 3 achieved higher predicted literacy mean scores at the end-of-year compared to students not participating in the program, however, treatment effect sizes fell short of the relevant threshold (g = 0.26). The program additionally had a mixture of negative findings and statistically non-significant findings among other analytic samples.

Table 5 presents the treatment and control group mean scores and mean score differences across all three usage levels by grade. As shown, the highest predicted Acadience scores are among the EISP students who used the program as recommended by the software vendors. In all grades (with the exception of 2nd), students who participated in the program significantly exceeded their control group counterparts in predicted literacy outcome scores.

Table 5. Acadience Predicted EOY Mean Scores by Usage and Grade

Grade	Condition	Intent to Treat	Met 80% of Rec.	Met Rec.
		End-of-Year Predicted Mean Scores		
K	Treatment	145.24	154.06	157.94
	Control	140.90	145.32	147.45
(diff)		*4.34*	*8.74*	*10.49*
1	Treatment	NS	184.69	191.79
	Control		180.96	185.19
(diff)			*3.73*	*6.60*
2	Treatment	259.31	271.46	280.32
	Control	263.85	273.45	281.80
(diff)		*-4.54*	*-2.00*	*-1.48*
3	Treatment	NS	396.57	407.58
	Control		391.96	401.33
(diff)			*4.61*	*6.25*
Data source: Matched K-3 ITT, MRU80, MRU samples. All mean comparisons displayed between treatment and control were statistically significant at p≤ .05.				

Table 6 shows the most meaningful program impact was on kindergarten students who were able to meet the vendors recommendations for use (g = 0.27). 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. We defined this threshold as any effect size equal or greater than 0.26, which is the average effect size seen in similar intervention programs (Lipsey et. al, 2012)[3].

[3] 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.

Table 6. Effect Sizes by Grade and Usage Level

Grade	Condition	Intent to Treat	Met 80% of Rec.	Met Rec.
		Effect Sizes		
K	Treatment	0.110	0.224	**0.268**
	Control			
1	Treatment	NS	0.056	0.096
	Control			
2	Treatment	-0.079	-0.035	-0.026
	Control			
3	Treatment	NS	0.072	0.099
	Control			

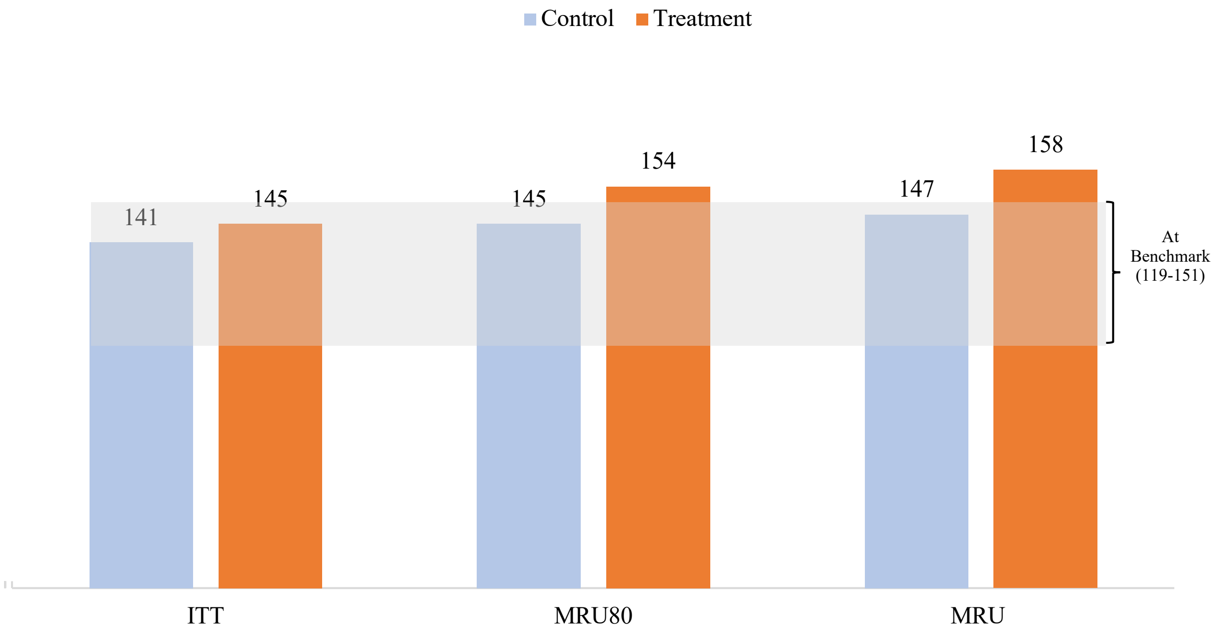
Data source: Matched K-3 ITT, MRU80, MRU samples. All effect sizes displayed were statistically significant at p≤ .05. Bold = Hedges' g exceeds the 0.26 threshold.

Kindergarteners who met 80% of vendor recommendations approached the meaningful threshold for effect size but fell just short (g = 0.22). Despite significant predicted mean differences, all other grades and usage levels had effect sizes below the 0.26 threshold. Additional information on effect sizes can be found in **Appendix F**.

Program Impacts on Acadience Literacy Scores in Context

It is also important to understand how the EISP impacted students' progress relative to grade level expectations. The following graphs depict not only the elevated 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 Mean Scores by Usage Level and Matched Sample

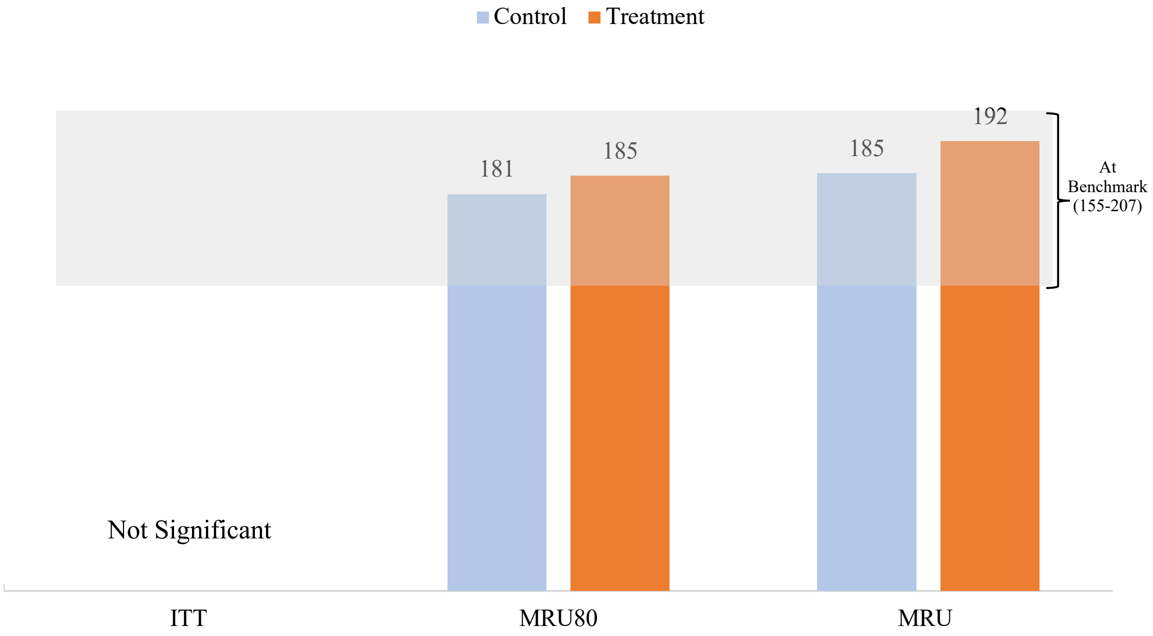


Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendation Kindergarten sample size –ITT n=37,354 (ctrl= 8,619, tr= 28,735); MRU80 n=25,694 (ctrl= 7,636, tr= 18,058); MRU n=21,499 (ctrl= 7,495, tr= 14,004); Students scoring ***At Benchmark*** *(119-151) or* ***Above Benchmark*** *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 figure 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 two highest usage subgroups (those that met vendors recommendations and those that met 80% of the recommendations) had the highest end-of-year mean score (158 and 154, respectively), putting them in the "above 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.

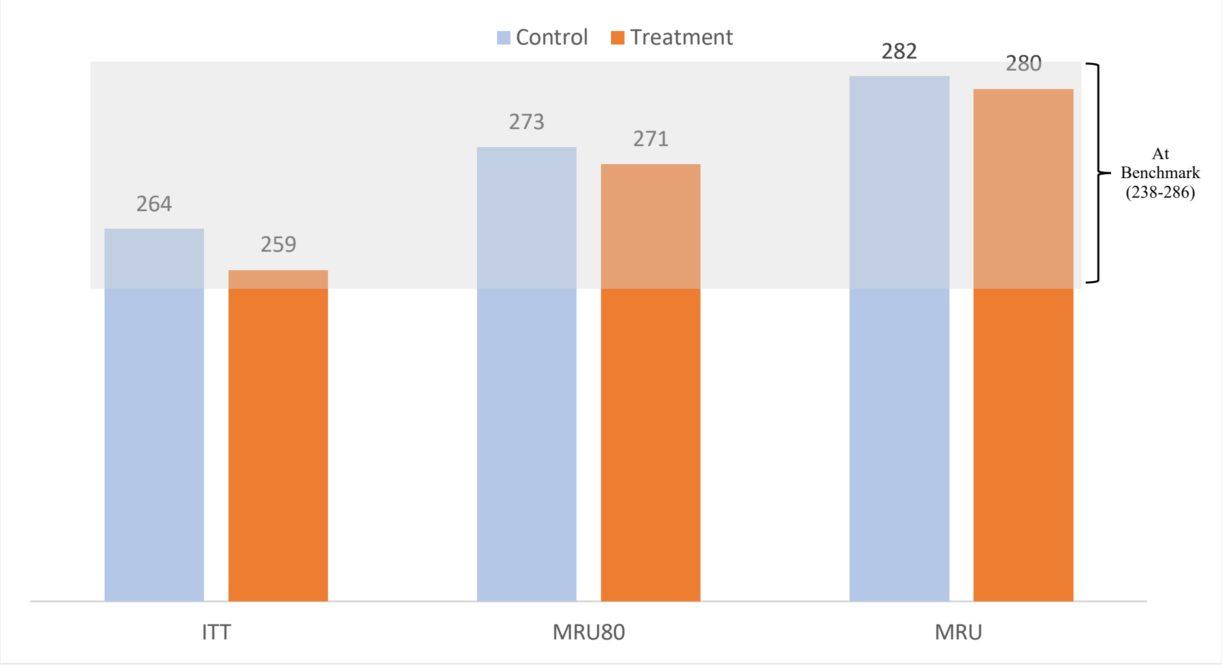
Figure 3. First Grade Predicted Mean Scores by Usage Level and Matched Sample



Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendiation
First Grade sample size- ITT n= 43,165 (ctrl= 9,960, tr= 33,205); MRU80 n= 34,087 (ctrl= 10,725, tr=23,362); MRU n=31,084 (ctrl= 10,836, tr=20,248); 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. The MRU80 and MRU mean comparisons displayed in the figure were statistically significant at p≤ .05.*

Figure 3 shows 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 (192). First grade students who used the program in any amount (ITT), did not differ from their matched comparison group. Again, all first graders averaged literacy levels within the expected range.

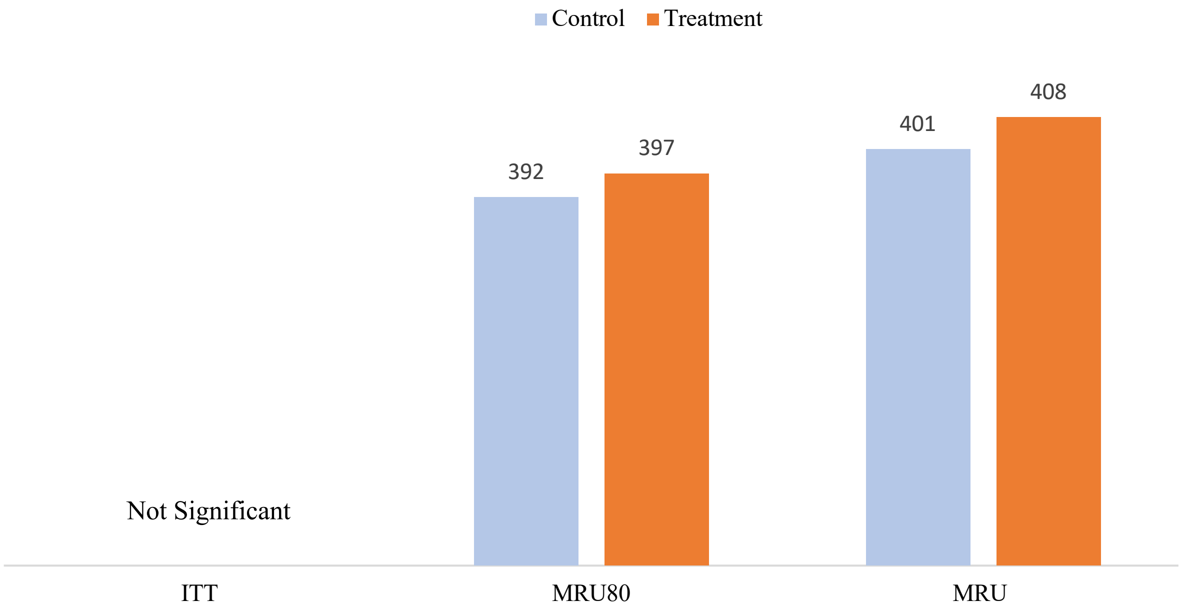
Figure 4. Second Grade Predicted Mean Scores by Usage Level and Matched Sample



Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendiation
Second Grade sample size - ITT n= 43,852 (ctrl= 10,118, tr= 33,734); MRU80 n=35,223 (ctrl= 10,468, tr=24,755);MRU n=30,478 (ctrl=10,625, tr= 19,853); 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 illustrates the predicted end-of-year mean scores for second grade students who used the EISP at different levels. While increased use of the program was reflected in greater end-of-year scores, the treatment students did not statistically outperform their matched control counterpart for this grade. This finding is addressed further in the discussion section of the report.

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



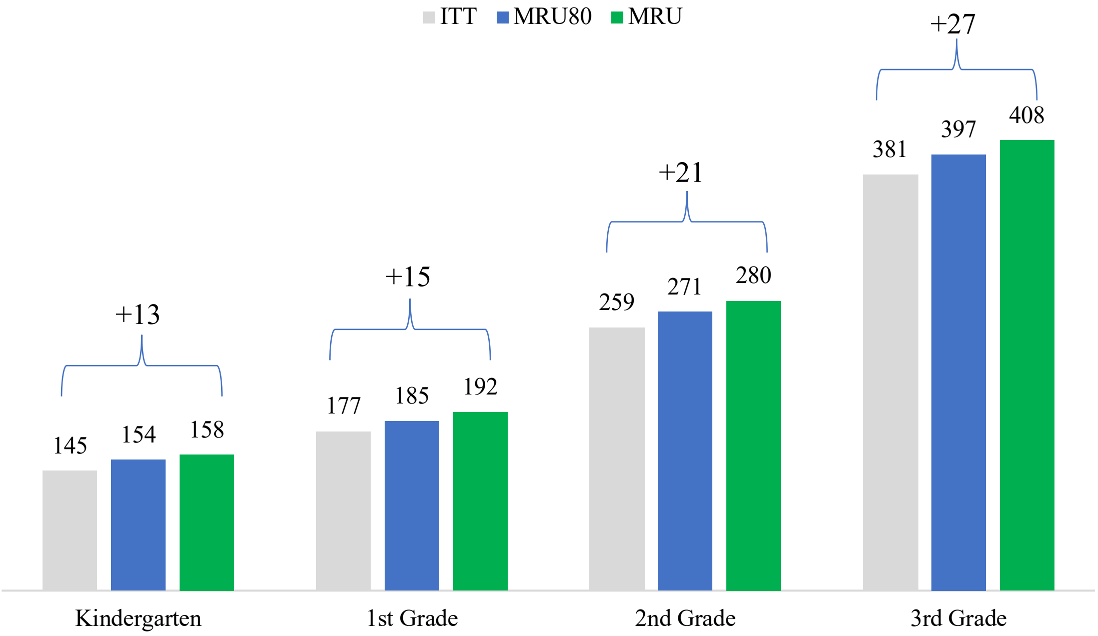
Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendiation
Third Grade sample size – ITT n= 42,178 (ctrl= 9,732, tr=32,446); MRU80 n=31,516 (ctrl=9,366, tr=22,150); MRU n=25,743 (ctrl= 8,974, tr=16,769); Students scoring ***At Benchmark*** *(330-404) or* ***Above Benchmark*** *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. The MRU and MRU80 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 highest achievement scores were aligned to the students who used the program with more consistency (397 and 408, respectively). Students who were able to fully meet the vendors' use requirements performed above benchmark. Similar to first grade, the ITT group did not differ statistically from their matched peers, yet all 3rd graders averaged literacy levels within the expected range.

How did different program usage levels influence Acadience outcome scores?

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 outcomes. **Figure 6** shows a side-by-side view of each grade and the three defined usage levels among treatment students who (1) met the recommendation for weeks and average minutes, (2) met 80% of the recommendation, and (3) who had any use, ITT. 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. EISP Students' Predicted Mean Scores by Grade and Usage Level



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

Similar to the prior school year, the greatest benefits of consistent program use are seen among the older grades (2nd and 3rd). Namely, the difference in literacy outcomes was highest in 2nd-3rd grade (+21 and +27 points, respectively) when comparing students engaged in casual program use to those engaged in vendor-recommended use. Results also suggest that as more advanced reading skills are practiced and acquired, adequate use of supplemental literacy interventions provide beneficial support within the classroom.

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, and special education designation status. **Table 7** presents the predicted mean scores for the Acadience Reading composite.

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

		Kindergarten	First Grade	Second Grade	Third Grade
Special Education	Treatment	143.28	178.84	262.47	411.10
	Control	132.80	172.24	263.47	404.85
ELL	Treatment	152.57	187.84	273.51	400.84
	Control	142.08	181.24	274.99	394.59
Low-Income	Treatment	155.13	185.40	274.32	404.37
	Control	144.65	178.79	275.80	398.12

Data source: Matched K-3 MRU sample. All data points displayed in figure were statistically significant at p≤ .05.

Across kindergarten, first and third grade and for every demographic subgroup, students in the EISP who were able to met 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 and third grade students.

DISCUSSION, LIMITATIONS, AND RECOMMENDATIONS

There were two primary goals for the 2021-2022 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' Acadience literacy achievement. We summarize here those findings, and present the known limitations, as well as our recommendations for improvement.

Implementation

An average of 51% of all EISP students (across grades K-3), were able to meet the recommended minimum usage levels put forth by program vendors. These use thresholds are shared with LEAs each year as guideposts to help facilitate the needed levels of engagement to effectively impact literacy achievement outcomes. Expectations for literacy gains should be tempered, if nearly half of the students are unable to adequately use the program. We noted a similar pattern during the 2020-2021 school year, where we postulated that the challenges stemmed from the COVID-19 pandemic and disruptions to in-person learning. It is less clear this school year, what barriers existed for achieving the recommended program usage (where the vast majority of Utah schools were open for the full academic year). 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 in order to positively impact year-end literacy scores.

Impacts

Substantive treatment effects were found in kindergarten among the students who met the vendors' usage requirements (g = 0.27). Students in grades 1 and 3 achieved higher predicted literacy mean scores at the end-of-year compared to students not participating in the program, however, treatment effect sizes fell short of the relevant threshold (g = 0.26). The EISP program additionally had a mixture of negative findings and statistically non-significant findings among other analytic samples.

We included several different usage subgroups in our impact analysis to help stakeholders understand the effect that program use had on student outcomes. Generally, EISP students who used the program as it was intended outperformed their control counterparts on predicted end-of-year Acadience outcomes. We observed this across kindergarten, first and third grade. EISP students also outperformed their fellow treatment peers who used the program less consistently. That is, we found a link between more consistent program use and stronger program effects.

Second grade was the exception, where results showed control students outperforming treatment students on predicted end-of-year Acadience scores. The driving factor is not fully clear. One possible explanation is the alignment of the skills taught by the program at second grade and what the EOY Acadience composite measured. It is important to note that the literacy assessment used in the current evaluation is a state chosen method and may not sufficiently align to the EISP vendors' intended outcome goal for each grade and each skill domain. It is possible that the skills students acquire by using a specific vendor's curriculum (in a given grade) may not be captured by the Acadience composite measure, and subsequently produce nonsignificant, difficult to interpret results. This may have been the case for second grade this program year.

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

Limitations

Additional Literacy Programs. New literacy programs and interventions do not always occur one at a time or in isolation, particularly when a state-wide educational priority is boosting literacy skills among students in K-3. We know that there are different types of programs simultaneously implemented across the state and across school districts. We do our best to control for these factors in our sampling approaches and statistical techniques, however, research conducted in live educational environments is inevitably susceptible to influences outside of the specific program under study.

Individual Teacher Influences. The variability in teachers' implementation of 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. 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 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 program 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. Genereally speaking, 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.
- Schools and LEAs may consider ways in which teachers can make the necessary modifications in their classrooms in order to provide the time and space for students to use the program at the recommended levels.
- Beyond achieving usage thresholds, understanding how the successful teachers are using the software in their classrooms could benefit all program participants. Perhaps a shared best-practice platform by grade could provide a helpful resource for teachers to access.
- 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.
- We also recommend that future evaluations continue to investigate the ways in which the EISP impacts students of all reading abilities so that the state can make informed decisions about the most optimal ways to support a population of students with diverse learning needs.

With intentional effort behind accountability, improving consistency of use, and the ability to marry multiple formats of literacy-focused programs, 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. To do so, 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 (and full vs half day for kinder), 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,[4] ITT; 0.000000000000003869; MRU80; 0.00000000000014 and MRU; 0.0000000000001137. 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?

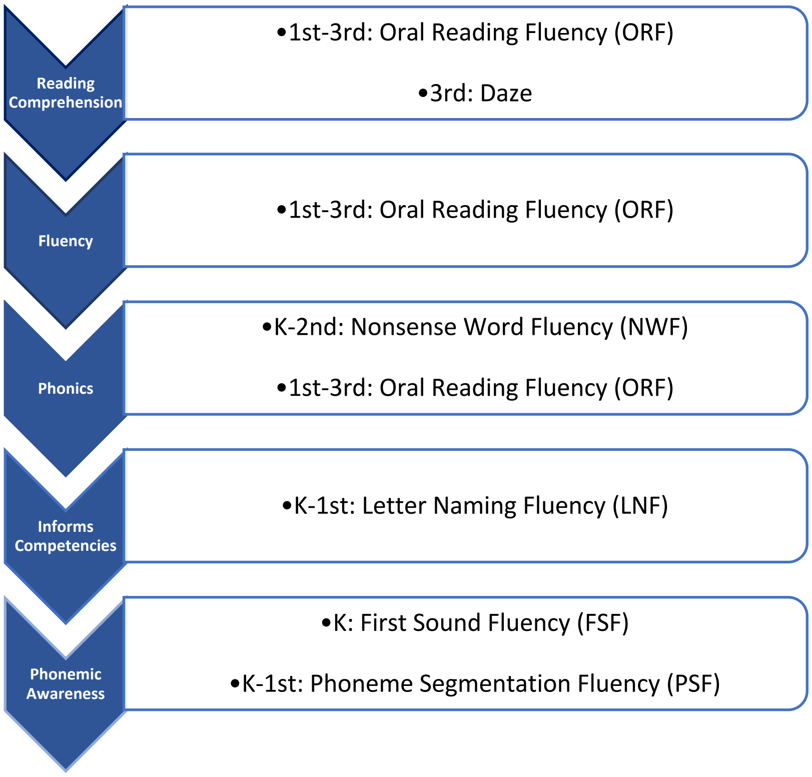
We collected data from nine different sources to create our master dataset for the EISP analyses. The data sources included: four program vendors, who provided us with usage information for each student who used their programs; state Acadience Learning (Acadience Reading) testing data; and student information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

[4] 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


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)

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, if a student was enrolled in a LETRS district, economic disadvantaged status, and ethnicity to adjust for their influence on end-of-year reading scores. In our kindergarten regression model, we also included the type of kindergarten (full or half day) students were enrolled in. 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 the 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 (0.2), medium (0.5) and large (0.8). 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. g= 0.25), intervention type (Avg. g= 0.13), intervention target (Avg. g= 0.40), or averaging all three methods (g= 0.26) (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 0.26, 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.[5] More information on how we selected our ES benchmark is provided in **Appendix F**.

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

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. MRU80 Sample by Grade[6]

	Grade	N	Female	Caucasion	SPED	Low-Income	ELL	BOY Comp
Control	K	7,636	49%	77%	6%	23%	5%	38.45
	1	10,725	49%	76%	9%	28%	6%	123.42
	2	10,468	48%	77%	10%	27%	7%	187.72
	3	9,366	49%	75%	12%	28%	9%	274.69
Treatment	K	18,058	49%	77%	6%	23%	5%	39.01
	1	23,362	53%	83%	10%	30%	6%	124.67
	2	24,755	48%	77%	10%	27%	7%	187.00
	3	22,150	49%	75%	12%	28%	9%	272.55

[6] The matched sample had an L1 score of 0.000000000000014000.Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates.

Table B2. ITT Sample by Grade [7]

	Grade	N	Female	Caucasion	SPED	Low-Income	ELL	BOY Comp
Control	K	8,619	49%	76%	7%	24%	6%	36.09
	1	9,960	49%	75%	10%	28%	7%	120.18
	2	10,118	49%	76%	11%	28%	8%	179.06
	3	9,732	49%	74%	14%	28%	9%	263.22
Treatment	K	28,735	49%	76%	7%	24%	6%	36.04
	1	33,205	49%	75%	10%	28%	7%	120.73
	2	33,734	49%	76%	11%	28%	8%	177.59
	3	32,446	49%	74%	14%	28%	9%	259.94

Table B3. MRU Sample by Grade [8]

	Grade	N	Female	Caucasion	SPED	Low-Income	ELL	BOY Comp
Control	K	7,495	49%	78%	6%	22%	5%	40.09
	1	10,836	49%	77%	9%	26%	5%	127.12
	2	10,625	48%	77%	9%	26%	7%	195.52
	3	8,974	49%	76%	11%	27%	8%	283.73
Treatment	K	14,004	49%	78%	6%	22%	5%	40.93
	1	20,248	49%	77%	9%	26%	5%	129.02
	2	19,853	48%	77%	9%	26%	7%	195.31
	3	16,769	49%	76%	11%	27%	8%	282.70

[7] The matched sample had an L1 score of 0.000000000000003869.

[8] The matched sample had an L1 score of 0.00000000000001137.

APPENDIX C. REGRESSION STATISTICS AND EFFECT SIZES BY SAMPLE

Table C1. ITT Regression Summary, by grade

	Grade	Condition	N	P-value	Marginal Mean	St. Error	Diff.	ES
Intent to Treat	K	Treatment	28,735	0.000	145.24	0.23	4.33	0.110
		Control	8,619		140.90	0.42		
	1	Treatment	33,205	0.634	176.58	0.36	0.40	0.006
		Control	9,960		176.18	0.74		
	2	Treatment	33,734	0.000	259.31	0.31	-4.54	-0.079
		Control	10,118		263.85	0.57		
	3	Treatment	32,446	0.150	380.53	0.36	1.07	0.017
		Control	9,732		379.46	0.65		

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 Summary, by grade

	Grade	Condition	N	P-value	Marginal Mean	St. Error	Diff.	ES
Met 80% of Recommended Use	K	Treatment	18,058	0.000	154.06	0.29	8.74	0.224
		Control	7,636		145.32	0.45		
	1	Treatment	23,362	0.000	184.69	0.41	3.73	0.056
		Control	10,725		180.96	0.73		
	2	Treatment	24,755	0.002	271.46	0.36	-2.00	-0.035
		Control	10,468		273.45	0.55		
	3	Treatment	22,150		396.57	0.43		0.072
		Control	9,366		391.96	0.66		

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 Summary, by grade

	Grade	Condition	N	P-value	Marginal Mean	St. Error	Diff.	ES
Met Recommended Use	K	Treatment	14,004	0.000	157.94	0.33	10.49	0.332
		Control	7,495		147.45	0.45		
	1	Treatment	20,248	0.000	191.79	0.46	6.60	0.126
		Control	10,836		185.19	0.73		
	2	Treatment	19,853	0.028	280.32	0.40	-1.48	-0.033
		Control	10,625		281.80	0.54		
	3	Treatment	16,769	0.000	407.58	0.49	6.25	0.123
		Control	8,974		401.33	0.67		

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 **156,682** to **155,222** 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[9] (approximately **1,767** cases) and any IDs that did not comply with the state student ID (SSID) format (**3,147** 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 **150,308** cases.

[9] 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 2021-2022 participants were included in the data. The SIS data file consisted of **208,378** cases, of which approximately three percent were duplicate records. After cleaning the data of duplicates, our SIS data consisted of **202,382** records.

Acadience Reading Data

In 2021-2022, the USBE prepared and transferred an Acadience Reading data file (n=**188,714**). 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 **177,688** 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 **177,622** cases. Next, we merged our master vendor data into the Acadience and SIS data and removed duplicate cases between vendors. This left us with **136,616** complete treatment cases and **40,988** 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[10]. This included excluding students who used Imagine Learning through a separate state-wide grant[11] prior to reporting the program impacts for similar reasons. After processing the data, our final, pre-matched dataset consisted of **173,057** cases, of which, **133,895** were treatment and **39,162** 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 2021-2022. 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

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

[11] 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

	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

APPENDIX F. DETERMINING EFFECT 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 0.25.
- *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 0.13.
- *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 0.40. Interventions that targeted individual students had the highest observed effect sizes, on average.

For the current research, 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 0.26.

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