AMIRA

LEARNING

Utah's Early Interactive Reading Software Program Report

Amira Learning | Every Child Deserves the Chance to Become a Reader
5214F Diamond Heights Blvd, #3255, San Francisco, CA 94131
wwww.AmiraLearning.com | 866-883-7323 | info@amiralearning.com

THE UTAH STATE BOARD OF EDUCATION

Report to the Education Interim Committee

Utah's Early Interactive Reading Software Program Report

October 27, 2023

Amber Wright
Digital and Instructional Materials Specialist, USBE
Amber.Wright@schools.utah.gov

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

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

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

ADA Compliant: 10/30/2023

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 K3 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

2022-2023 Program Evaluation Findings

Submitted to the Utah State Board of Education
October 2023

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

ADA Compliant: 10/30/2023

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 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 Krislyn Mimnaugh and Rhiann Eddy from Amira, Ann Minckler and Erin Barnett from Imagine Learning, Haya Shamir, Steven Kiisel, and Erik Yoder from Waterford, Kelley Pasatta and Sarah Franzén from Lexia, Corey Fitzgerald, and Tracey Jackson 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 five adaptive computer-based programs: Amira Learning, Imagine Learning (Imagine Language & Literacy), Curriculum Associates (i-Ready), Lexia® (Core5), and Waterford. The Evaluation and Training Institute (ETI), the EISP program evaluator, studied two core aspects of the program: 1) students' use of the program during the school year (program enrollment and implementation); and 2) the effects of the program on increasing students' literacy achievement (program impacts). The current evaluation investigated program across all five vendors combined (program-wide) and 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 2022-2023 school year, five EISP software vendors were used in a total of 140 LEAs, in 692 schools and by 166,468 students throughout the state of Utah. Similar to previous years, Core5 was used by the most students (116,789), followed by Amira (24,127), Imagine Language & Literacy (17,042), i-Ready (7,802), and Waterford (708).

Our implementation study found a large number of students were unable to meet the recommended minimum usage levels put forth by the software providers. On one side, state-wide program implementation provides the opportunity for large numbers of students to receive program benefits, however, it is critical for students to use the program for the recommended amount of time in order to realize the benefit to literacy achievement.

Program-Wide Impact on Acadience Achievement Scores

Literacy achievement was measured using the state provided Acadience Reading scores. We found for all students in grades K-3 who met the recommended usage, their predicted end-of-year Acadience scores were higher than their control counterparts. We also found that treatment effects were largest for students who used the program as intended. And among all four grades, Kindergarten students were most significantly impacted by participation in EISP.

EISP and Different Student Populations. We additionally studied how the program benefitted students in specific demographic subgroups, such as English Language Learners, low-income, or special education designation status. We found that for every subgroup, students in the EISP who met the vendors' recommended use criteria, outperformed their non-program counterparts.

Recommendations

The current evaluation identified positive student literacy achievement outcomes, most notably for kinder students who met the vendors' usage recommendations. 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, specifically students who start the year reading below benchmark (high risk), 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. Six software vendors were selected to provide software and training to schools through the EISP in 2022-2023, however only five programs were used by LEAs. The five vendors used during the school year were (in alphabetical order): Amira Learning, Curriculum Associates ("i-Ready"), Imagine Language & Literacy, 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 2022-2023 academic year, the EISP's tenth 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**.

[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.

FINDINGS

Program Enrollment

In 2022-2023, five EISP software vendors were used in a total of 140 LEAs, in 692 schools and by 166,468 students. Core5 was the most widespread program in the state relative to other EISP providers, reaching 87 LEAs, 434 schools, and 116,789 students (**Table 1**).

Table 1. 2022-2023 Program Enrollment Overview

Program	LEAs	Schools	Students (K 3)
Amira	17	142	24,127
Core5	87	434	116,789
Imagine Language & Literacy	15	69	17,042
i-Ready	16	39	7,802
Waterford	5	8	708
Total	140	692	166,468

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

Program wide student enrollment was consistent across all grades, with a similar number of students enrolled in first, second and third grade. (**Table 2**).

Table 2. 2022-2023 Program Enrollment by Grade

Program	Kinder	1st	2nd	3rd
Amira	179	7,922	8,010	8,016
Core5	25,922	30,371	30,599	29,937
Imagine Language & Literacy	3,528	4,587	4,660	4,267
i-Ready	1,504	2,003	2,207	2,088
Waterford	323	287	67	31

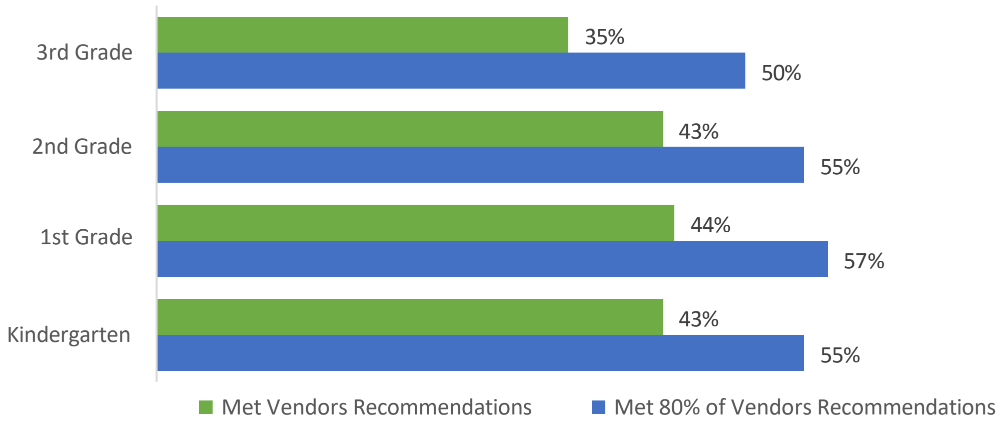
Program	Kinder	1st	2nd	3rd
Total	31,456	45,170	45,543	44,339
Data source: software vendor data in K-3				

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. In this section we answer the research question: *To what extent did students use the software program as intended?*

Just over 40% of kindergarteners, 1st graders and 2nd graders were able to adhere to the recommended weeks AND average weekly minutes, while roughly a third of 3rd graders (35%) met the vendor recommendations. (**Figure 1**; green bars).

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'

As depicted in **Figure 1**, 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 10-15% across all grades.

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 2022-2023 Minimum Use Recommendations

Program	Kindergarten	First Grade	Second Grade	Third Grade	Suggested Minimum Weeks
Amira	30 min/week	30 min/week	30 min/week	30 min/week	30 weeks

Program	Kindergarten	First Grade	Second Grade	Third Grade	Suggested Minimum Weeks
Core5	20 minutes to 60 min/week*	20 minutes to 60 min/week*	20 minutes to 60 min/week*	20 minutes to 60 min/week*	20 weeks
Imagine Language & Literacy	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 30 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. 2022-2023 Program Use by Vendor and Grade

Program	Grade	N	Ave Weekly Min.	Ave Total Min.	Ave Wks. of Use
Amira	K	179	16	150	8
	1	7,922	18	314	15
	2	8,010	20	354	16
	3	8,016	19	332	15
	Total	24,127	19	332	15
Core5	K	25,922	47	1,258	25

Program	Grade	N	Ave Weekly Min.	Ave Total Min.	Ave Wks. of Use
	1	30,371	56	1,707	29
	2	30,559	51	1,542	29
	3	29,937	49	1,415	27
	Total	116,789	51	1,489	28
Imagine Language & Literacy	K	3,528	38	850	20
	1	4,587	45	1,053	21
	2	4,660	41	955	20
	3	4,267	37	811	19
	Total	17,042	40	923	20
i-Ready	K	1,504	34	728	20
	1	2,003	38	926	23
	2	2,207	38	948	23
	3	2,088	43	901	21
	Total	7,802	38	887	22
Waterford	K	323	39	1,105	27
	1	287	49	1,415	27
	2	67	60	1,776	28
	3	31	67	1,836	26
	Total	708	46	1,325	27

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.

It warrants acknowledgement that just under 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 the "met recommendations" and "met 80% of recommendations" groups.

Program Impacts on Acadience Literacy Test Scores

We analyzed the program's impact on Acadience test scores 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**[2]. In this section we answer the research questions: *How did the EISP impact students' Acadience scores? And how did different program usage levels influence Acadience outcome scores?*

Key Takeaway. **EISP Students in grades K-3 achieved higher predicted literacy mean scores at the end-of-year compared to students not participating in the program, however, large substantive treatment effect sizes were only found in kindergarten. Additionally, Acadience scores were highest among those using the program as recommended.**

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, 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	153.38	164.63	168.81
	Control	147.07	151.93	154.14
(diff)		*6.31*	*12.7*	*14.67*
1	Treatment	82.35	88.18	91.38
	Control	79.68	84.07	86.47
(diff)		*2.67*	*4.12*	*4.92*

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

Grade	Condition	Intent to Treat	Met 80% of Rec.	Met Rec.
2	Treatment	NS	NS	293.35
	Control			291.74
(diff)				*1.6*
3	Treatment	387.44	411.69	425.09
	Control	384.66	405.09	416.83
(diff)		*2.78*	*6.61*	*8.27*

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

Effect sizes describe the magnitude of the difference between two groups on an outcome measure. We adapted a set of effect size benchmarks based on categories from Kraft (2020) that were adjusted for early literacy outcome measures: less than 0.10 is *small*, 0.10 to less than .30 is *medium* and .30 or greater is *large* (M. Kraft, personal communication, October 13, 2023). **Table 6** shows the effect sizes where the most meaningful program impact was on kindergarten students in the highest two usage groups, those who were able to meet the vendors recommendations for use (g = 0.37) and for those who met 80% of the vendors recommendations (g= 0.32). All other grades and usage levels had effect sizes that reflected medium or small treatment effects.

Table 6. Effect Sizes by Grade and Usage Level

Grade	Intent to Treat	Met 80% of Rec.	Met Rec.
K	*0.16*	**0.32**	**0.37**
1	*0.10*	*0.15*	*0.18*
2	NS	NS	0.03
3	0.04	*0.10*	*0.13*

Data source: Matched K-3 ITT, MRU80, MRU samples[3]. All effect sizes displayed represent statistically significant mean differences at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater.

[3] Kindergarten sample size –ITT ctrl=6,117.252 , tr= 26,604; MRU80 ctrl= 6082.588, tr= 15,678; MRU ctrl= 6,287.8985, tr= 12,639; First Grade- ITT - ctrl= 8356.334, tr= 35,725; MRU80 - ctrl= 8824.249, tr=22,318;

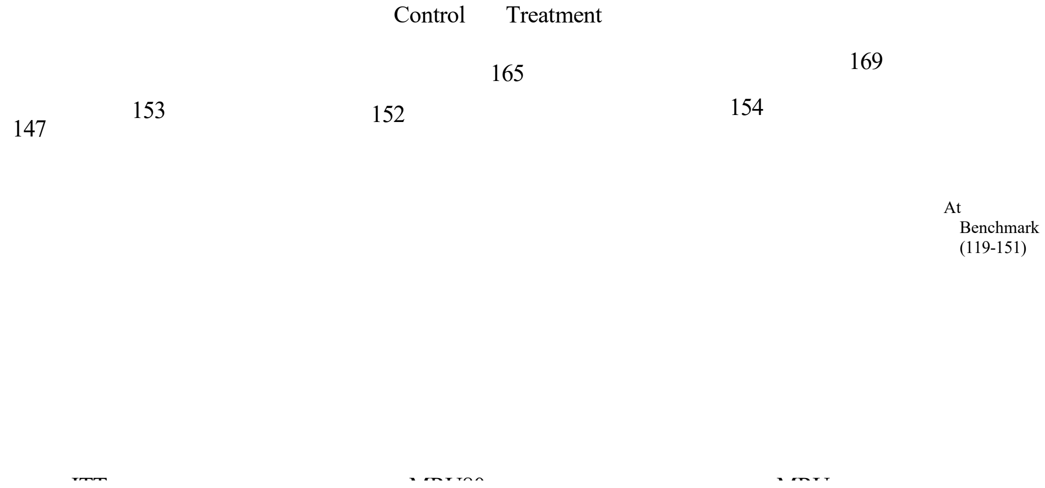
There are multiple ways to interpret effect sizes, including the use of categories such as small, medium, or large (e.g., Cohen, 1988; Kraft, 2020), or using a minimum threshold (Hill 2008). Variations of both approaches are widely used and accepted, yet both require careful considerations of the research design and key study components (such as sample, measures, etc.) Our effect size interpretation approach uses a categorical range based on effect sizes for similar types of research, studying similar interventions (early literacy programs) and with similar populations (elementary students). Specifically, the range used in the current study represents the benchmarks for early literacy found in a summary of meta-analyses of relevant and similar educational studies, as well as the direct recommendation from the author (Kraft, 2020; M. Kraft, personal communication, October 13, 2023).

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.

MRU- ctrl= 8,932.693, tr= 17,600; Second Grade sample size - ITT ctrl= 8,786.8247, tr= 38,214; MRU80 ctrl= 8823.594, tr=22,743; MRU ctrl= 8,897.284, tr= 17,884; Third Grade sample size – ITT ctrl= 8,233.366, tr=35,807; MRU80 ctrl=7611.577, tr=19,619; MRU ctrl= 7,138.125, tr=14,348.

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

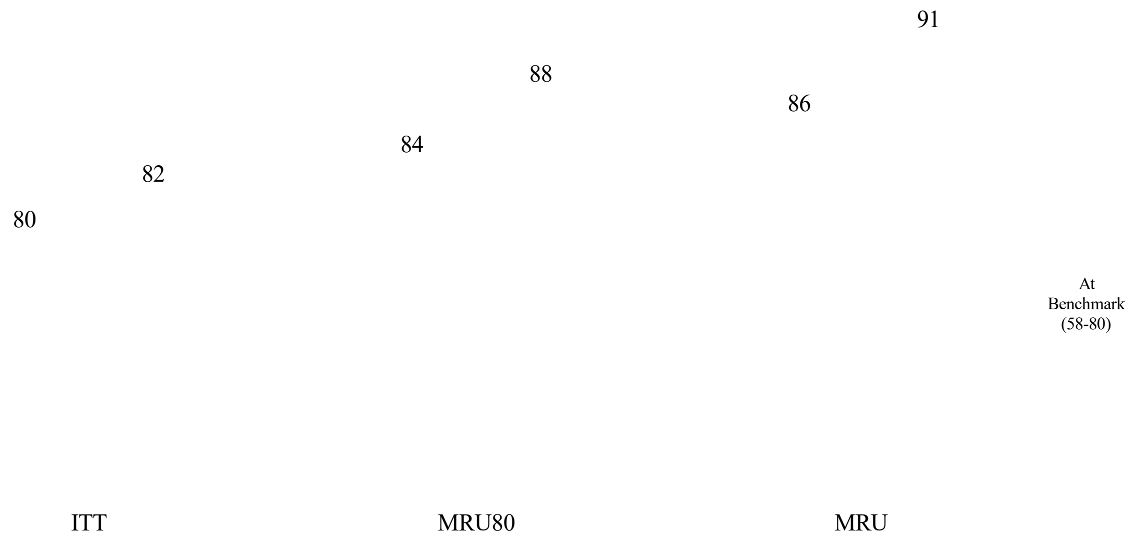


*Note: Students scoring **At Benchmark** (119-151), or **Above Benchmark** goal (152 or greater) have the odds in their favor (approximately 80% to 90% overall) of achieving later important reading outcomes. Data source: Matched K-3 ITT, MRU80 and MRU samples. All mean comparisons displayed in the figure 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, along 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 (169 and 165, respectively), putting them in the "above benchmark" score range. Though the matched control students for the higher usage groups had predicted mean scores in the "above benchmark" range, treatment students had end-of-year mean scores 13-15 points higher than the control students. These findings further support that when the program is used consistently in kindergarten, 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.

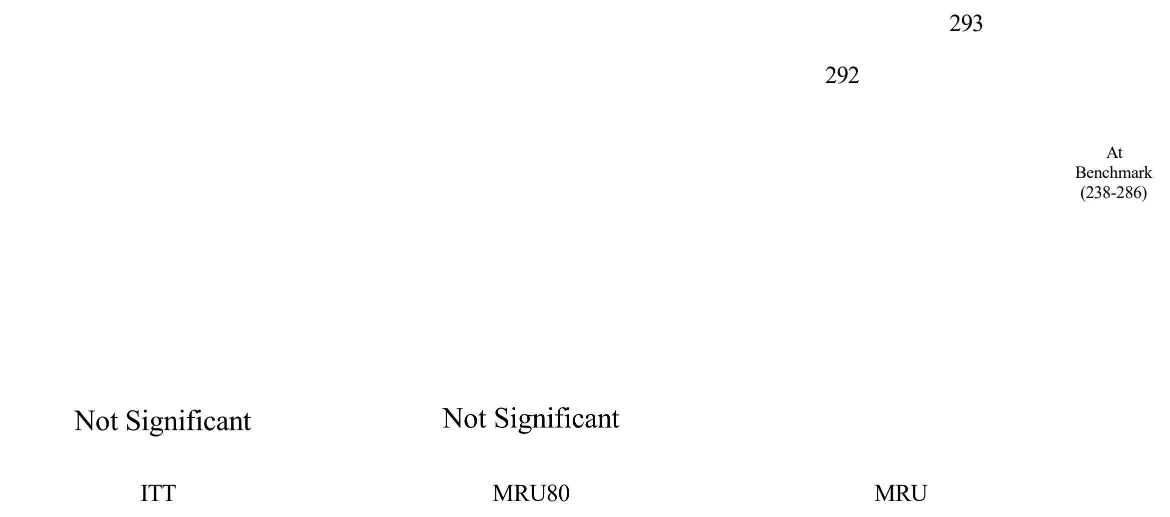
Figure 3. First Grade Predicted Mean Scores by Usage Level and Matched Sample



*Note: First grade end-of-year predicted outcomes were measured with the Nonsense Word Fluency- Correct Letter Sounds scale and has a different range than the reading composite scale. Students scoring **At Benchmark** (58-80), or **Above Benchmark** goal (81 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 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 (91). First grade students using the software in any amount had predicted end-of-year mean scores higher than the comparison students. All first graders averaged literacy levels at or above the benchmark.

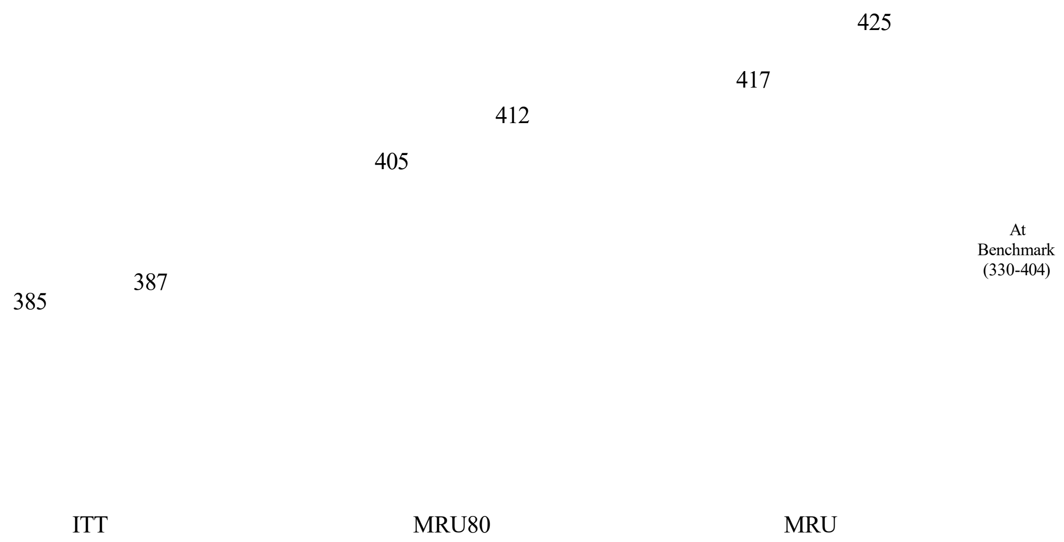
Figure 4. Second Grade Predicted Mean Scores by Usage Level and Matched Sample



Note: 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. The program did not have statistically significant results for those who used the program in the lower two usage groups- intent to treat and met 80% of recommended use. In the highest usage group, treatment students had predicted mean scores of 293, which was only one point higher than comparison students. Both treatment and control students had predicted end of year scores that were above benchmark. 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: 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 highest achievement scores were aligned to the students who used the program as the vendor intended (425). All 3rd graders averaged literacy levels within the expected range or above benchmark.

EISP and Different Demographic Groups

We were also interested in studying how the program may benefit students in specific demographic subgroups. In this next section we answer the research question: *What impact did EISP have on specific student populations?* We conducted a separate analysis of program impacts on students identified as English Language Learners, low-income, and special education designation status. **Table 8** 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	155.06	83.64	278.19	401.39
	Control	140.38	78.72	276.59	393.12
ELL	Treatment	167.79	84.74	276.61	414.72
	Control	153.12	79.82	275.00	406.45
Low-Income	Treatment	168.07	88.78	289.51	420.06
	Control	153.40	83.87	287.91	411.79
	Data source: Matched K-3 MRU sample. All data points displayed in figure were statistically significant at p≤ .05.				

Across all grades and demographic subgroups, students in the EISP who were able to meet 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.

A Within Treatment Comparison

As shown in the analysis sections above, our evaluation sought to show differences *between* treatment and control students, but equally important was understanding how different levels of program participation specifically among EISP students impacted literacy outcomes. **Table 9** 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. This finding is especially pronounced in 2nd and 3rd grade.

Table 8. EISP Students' Predicted Mean Scores by Grade and Usage Level

Grade	Intent to Treat	Met 80% of Rec.	Met Rec.	Diff ITT to MRU
K	153	165	169	**+16**
1	82	88	91	**+9**
2	264	284	293	**+29**
3	387	412	425	**+38**

Note: ITT: Intent to Treat; MRU80: Met 80% of recommendation; MRU: Met recommendation. Kindergarten, second and third grade students predicted means were measured with the reading composite scale and first grade end-of-year predicted outcomes were measured with the Nonsense Word Fluency- Correct Letter Sounds scale, which has a different range than the reading composite scale.

Like in previous school years, the greatest benefits of consistent program use are seen among the 2nd and 3rd grade students. As seen in Table 8, the point difference in 2nd and 3rd literacy outcomes was 29 and 38, 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.

DISCUSSION, LIMITATIONS, AND RECOMMENDATIONS

There were two primary goals for the 2022-2023 EISP evaluation: (1) to study program implementation, and (2) to determine the program's impact on Acadience literacy scores. In this section, we summarize those findings, and present the known limitations, as well as our recommendations for program improvement.

Implementation

An average of 41% of all EISP students (across grades K-3), were able to meet the recommended minimum usage levels put forth by program vendors, which is a slight drop compared to previous years of the program. 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 less than half of the students are unable to adequately use the program. This drop in usage follows a pattern we have observed since the 2020-2021 school year, where we postulated that the challenges stemmed from the COVID-19 pandemic and disruptions to in-person learning. We can no longer attribute declines in usage to any major disruptions to the school year. The responsibility of successful implementation and adhering to the recommended usage set by the program vendors falls to the LEAs. 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 positively impact year-end literacy scores.

Impacts

Large and substantive treatment effects were found in kindergarten among the students who met the vendors' usage requirements ($g = 0.37$) and for students who met 80% of the recommendations ($g=0.32$). In the highest usage group, students in grades 1-3 achieved higher predicted literacy mean scores at the end-of-year compared to students not participating in the program, however, the treatment effect sizes were not as strong (medium or small), compared to those in kindergarten.

We included several different usage definitions in our impact analysis to help stakeholders understand the effect that varying usage levels 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 pattern in all grades K-3. In 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. This relationship was less clear in second grade (as has been the case in previous years), however evaluating the impacts of a program used at levels *below* the recommendation to impact literacy, inherently creates results that are difficult to interpret.

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

We do our best to control for all possible influences on student reading outcomes 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. 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 prevalence 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.

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.

Recommendations

The results of the evaluation underscore the importance of supporting students' literacy development and creating opportunities for our youngest learners. Generally 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:

- To boost the number of students adhering to the minimum recommended usage levels, we encourage all EISP software vendors (new and veteran) to clearly define usage recommendations for LEAs at the beginning of the school year. A new approach to communicating these requirements may be needed.
- 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.
- We also recommend that future evaluations continue to investigate the ways in which the EISP impacts students of all reading abilities, specifically students who start the year reading below benchmark (high risk), 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.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189x20912798

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, 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; 0.00000000000005116; MRU80; 0.000000000000003423 and MRU;

[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.

0.00000000000002693. 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?

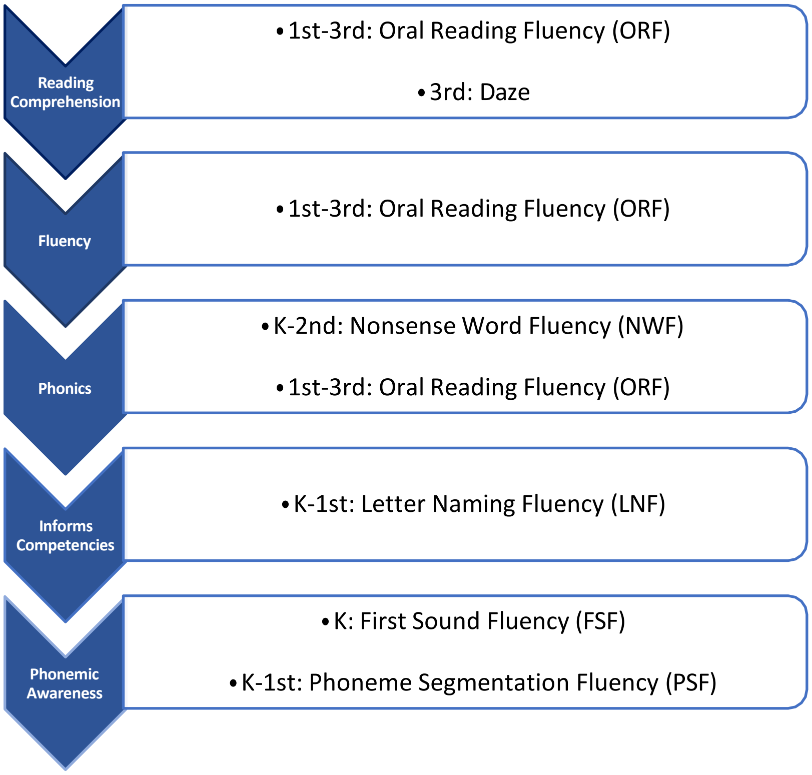
We collected data from ten different sources to create our master dataset for the EISP analyses. The data sources included: five program vendors, who provided us with usage information for each student who used their programs; state Acadience Learning (Acadience Reading) testing data; and student information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

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, 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 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.

There are multiple ways to interpret effect sizes, including the use of categories such as small, medium, or large (e.g., Cohen, 1988; Kraft, 2020), or using a minimum threshold (Hill 2008). Variations of both approaches are widely used and accepted, yet both require careful considerations of the research design and key study components (such as sample, measures, etc.) Our effect size interpretation approach uses a categorical range based on effect sizes for similar types of research, studying similar interventions (early literacy programs) and with similar populations (elementary students). We adapted a set of effect size benchmarks based on categories from Kraft (2020) that were adjusted for early literacy outcome measures: less than 0.10 is *small*, 0.10 to less than .30 is *medium* and .30 or greater is *large* (M. Kraft, personal communication, October 13, 2023).Specifically, the range used in the current study represents the benchmarks for early literacy found in a summary of meta-analyses of relevant and similar educational studies, as well as the direct recommendation from the author (Kraft, 2020; M. Kraft, personal communication, October 13, 2023).

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 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 (MRU), (2) students who met at least 80% of the recommended weeks and average weekly minutes (MRU80), 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, beginning-of-year achievement scores and benchmark levels[5], 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 coefficients. 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. For kindergarten, 2nd and 3rd grade end of year group mean scores, we used the reading composite score to measure student outcomes

[5] Students in kindergarten, 2nd and 3rd grade were matched on reading composite scores (BOY Comp) and students in 1st grade were matched on nonsense word fluency, correct letter sounds (NWF-CLS) scores.

and for 1st grade students, we used the nonsense word fluency, correct letter sounds as our outcome variable.

APPENDIX B. ANALYTIC SAMPLES

Tables B1 – B3 present the characteristics for the population sample, as well as the matched sample used in our analyses. We also present the L1 statistic for each covariate in the matches ample. Lower values indicate less imbalance, and the closer to zero the better the two samples were balanced across covariates.

Table B1. Matched Treatment ITT Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	26,895	49%	73%	9%	30%	6%	36.04
	1	37,814	49%	73%	12%	31%	73%	37.98
	2	39,098	49%	73%	13%	31%	8%	178.39
	3	36,515	49%	73%	15%	31%	9%	261.53
Matched ITT Treatment Sample	K	26,604	49%	73%	9%	30%	6%	36.12
	1	35,725	49%	77%	11%	30%	6%	36.99
	2	38,214	49%	75%	13%	31%	7%	179.37
	3	35,807	49%	74%	14%	30%	8%	262.82

Note: The matched sample had a multivariate L1 score of 0.00000000000005116. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.000000000000079), White (L1= 0.000000000000014), SPED (L1 = 0.0000000000000077), Low-Income (L1= 0.000000000000079), and ELL (L1= 0.0000000000000045).

Table B2. Matched Treatment MRU80 Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	15,856	49%	74%	8%	28%	6%	39.87
	1	23,624	49%	76%	11%	28%	6%	41.77
	2	23,196	49%	75%	11%	29%	7%	195.42
	3	19,973	49%	74%	13%	29%	8%	282.06
Matched MRU 80 Treatment Sample	K	15,678	49%	75%	8%	28%	5%	39.98
	1	23,188	49%	79%	10%	28%	5%	40.58
	2	22,743	49%	76%	11%	29%	6%	196.40
	3	19,619	49%	75%	12%	29%	8%	283.40

Note: The matched sample had a multivariate L1 score of 0.000000000000003423. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.00000000000002), White

(L1= 0.0000000000000065), SPED (L1 = 0.0000000000000063), Low-Income (L1= 0.000000000000011), and ELL (L1= 0.00000000000000019).

Table B3. Matched Treatment MRU Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	12,784	48%	75%	8%	27%	5%	41.72
	1	18,650	49%	77%	10%	28%	5%	43.98
	2	18,224	50%	76%	11%	28%	6%	204.50
	3	14,574	49%	75%	12%	28%	8%	294.71
Matched MRU Treatment Sample	K	12,639	48%	76%	7%	26%	5%	41.85
	1	17,600	49%	81%	9%	26%	4%	42.71
	2	17,884	50%	77%	10%	27%	5%	205.42
	3	14,348	49%	76%	11%	28%	7%	295.97

Note: The matched sample had a multivariate L1 score of 0.00000000000002693. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.000000000000029), White (L1= 0.000000000000028), SPED (L1 = 0.0000000000000031), Low-Income (L1= 0.000000000000026), and ELL (L1= 0.0000000000000072).

APPENDIX C. REGRESSION STATISTICS AND EFFECT SIZES BY SAMPLE

Table C1. ITT Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Intent to Treat	K	Treatment	0.000	153.38	0.25	6.31	*0.16*
		Control		147.07	0.52		
	1	Treatment	0.000	82.35	0.14	2.67	*0.10*
		Control		79.68	0.30		
	2	Treatment	NS				
		Control					
	3	Treatment		387.44	0.34		0.04
		Control		384.66	0.71		

Note. Hedges' g effect size (ES) benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater. Data source: Matched K-3 ITT sample. Kindergarten sample size –ctrl=6,117.252, tr= 26,604; First Grade sample size- ctrl= 8356.334, tr= 35,725; Second Grade sample size - ctrl= 8,786.8247, tr= 38,214; Third Grade sample size – ctrl= 8,233.366, tr=35,807.

Table C2. MRU 80 Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Met 80% of Recommended Use	K	Treatment	0.000	164.63	0.32	12.70	**0.32**
		Control		151.93	0.51		
	1	Treatment	0.000	88.18	0.18	4.12	*0.15*
		Control		84.08	0.29		
	2	Treatment	NS				
		Control					
	3	Treatment	0.000	411.69	0.45	6.61	*0.10*
		Control		405.09	0.73		

Note. Hedges' g effect size (ES) benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater. Data source: Matched K-3 MRU80 sample. Kindergarten sample size ctrl= 6082.588, tr= 15,678, First Grade sample size- ctrl= 8824.249, tr=22,318; Second Grade sample size ctrl= 8823.594, tr=22,743; Third Grade sample size ctrl=7611.577, tr=19,619.

Table C3. MRU Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Met Recommended Use	K	Treatment		168.81	0.35	14.67	<u>**0.37**</u>
		Control		154.14	0.50		
	1	Treatment		91.38	0.21	4.92	*0.18*
		Control		86.47	0.29		
	2	Treatment	0.025	293.35	0.41	1.60	0.03
		Control		291.74	0.58		
	3	Treatment	0.000	425.09	0.52	8.27	*0.13*
		Control		416.83	0.74		

Note. Hedges' g effect size (ES) benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, <u>**Large**</u>: bold and underlined text: .30 or greater. Data source: Matched K-3 MRU sample. Kindergarten sample size ctrl= 6,287.8985, tr= 12,639, First Grade sample size ctrl= 8,932.693, tr= 17,600, Second Grade sample size ctrl= 8,897.284, tr= 17,884; Third Grade sample size ctrl= 7,138.125, tr=14,348.

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 **172, 944** to **166,468** 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[6] (approximately **5,957** cases) and any IDs that did not comply with the state student ID (SSID) format (**1,437** 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 **159,074** cases.

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 2022-2023 participants were included in the data. The SIS data file consisted of **206,578** cases, of which approximately three percent were duplicate records. After cleaning the data of duplicates, our SIS data consisted of **200,762** records.

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

Acadience Reading Data

In 2022-2023, the USBE prepared and transferred an Acadience Reading data file (n= **186,810**). 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 **176, 357** 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 **176, 255** cases. Next, we merged our master vendor data into the Acadience and SIS data and removed duplicate cases between vendors. This left us with **144,781** complete treatment cases and **31,829 control** cases.

Lastly, we identified (where possible) schools or students with program exposure, using one of the five program vendors through non-EISP funding. We removed these cases from our pool of potential controls[7]. This included excluding students who used Imagine Learning through a separate state-wide grant[8] prior to reporting the program impacts for similar reasons. After processing the data, our final, pre-matched dataset consisted of **172,151** cases, of which, **140,322** were treatment and **31,829** were potential controls.

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

[8] 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.

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 2022-2023. 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 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 E1** provides a summary of the Acadience subscales used in our analyses.

Table E1. 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

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

RESEARCH REPORT

EVALUATION AND TRAINING INSTITUTE
www.eticonsulting.org

AN EVALUATION AND TRAINING INSTITUTE REPORT

Utah's Early Intervention Reading Software Program

2022-2023 Impact on At-Risk Students

Submitted to the Utah State Board of Education
October 2023

For more information contact:
Jon Hobbs, Ph.D.
jhobbs@eticonsulting.org

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

INTRODUCTION OF AT-RISK ANALYSIS

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 USBE was specifically interested in understanding how the program was impacting students who were at-risk of falling behind in reading, defined as reading below grade level at the start of the school year. The Evaluation and Training Institute (ETI), the EISP independent evaluator, was contracted to conduct this additional analysis with the EISP data from the 2022-2023 program year. The current analysis investigated the impact of the two software programs specifically contracted to serve at-risk students, Lexia, and Amira.

The following report includes the 2022-2023 EISP enrollment numbers of at-risk students across grade and vendor, the methods used, and the impact of the program on test scores. We additionally looked at literacy growth rates across the school year and compared program students to their non-program counterparts for each of the two software providers independently. The at-risk analysis, including all results shown in this report, included only students reading below/well below grade level at the beginning of the school year and were engaged with the software for at least the minimum recommended use.

METHOD

Sampling. During the 2022-2023 school year, *Lexia* and *Amira* served a combined total of 140,956 students in the EISP program, representing 85% of the entire EISP population of students. Among the total number of participating students served by either *Lexia* or *Amira*, 44,134 (31%) were considered 'at-risk' readers at the beginning of the school year as defined by their below grade level performance on the Acadience literacy assessment. **Table 1** shows the total number of EISP students and the number of at-risk EISP students served by each vendor by grade.

Table 1. 2022-23 Total Number of Lexia and Amira students & At-Risk by Vendor and Grade

Program	Grade	Total EISP Participants*	At Risk EISP Participants**
Lexia	K	25,922	8,972
	1	30,371	10,635
	2	30,599	8,177
	3	29,937	8,314
	Total	116,829	36,098
Amira	K	179	--
	1	7,922	3,144
	2	8,010	2,541
	3	8,016	2,351
	Total	24,127	8,036

*Data source: K-3 vendor usage data after cleaning duplicates and missing data
**At-Risk numbers are based on Acadience beginning of year literacy level (below or well below grade level)

In order to study EISP's impact on at-risk students' Acadience literacy test scores, we needed two broad samples, at-risk students who participated in the program (Treatment group) and at-risk students who did not participate in the program (Control group). Students in the control group were matched to the treatment students across characteristics that influenced learning. We matched our treatment to control students using socio-economic status, demographic information and beginning-of-year Acadience test scores. Our analytic samples (i.e., the samples we used in the analysis) were composed of students in our treatment and control groups within grades K-3. ETI created several sets of matched samples of at-risk students by grade for both *Lexia* and *Amira*.

Not every EISP student engaged with the program as it was recommended by the software providers. The findings included here, therefore, are focused on the at-risk students who used the vendors' software as it was intended[1]. **Table 2** illustrates the total number of EISP students by grade and vendor who were at-risk and met the vendors' recommended usage.

Table 2. Number of Lexia and Amira Students At-Risk and Met Vendor Recommendations

Program	Grade	Students At Risk & Met Rec. Usage
Lexia	K	3,558
	1	5,329
	2	3,426
	3	3,019
	Total	**15,332**
Amira	K	--
	1	101
	2	75
	3	73
	Total	**249**

We've provided a high-level summary of all outcome data in **Appendix A** for reviewers who would like to see the results among at-risk students using the program in lesser amounts.

Matching. We matched control students who did not participate in the program to students who did using Coarsened Exact Matching (CEM). We used CEM to match students on grade, beginning-of-year achievement scores and benchmark levels[2], gender, race, English Language Learner (ELL) status, and poverty status. All matched samples were statistically well-balanced as indicated by L1 coefficients. The baseline characteristics of the treatment and control samples can be found in **Appendix B and C**.

[1] Usage recommendations vary by grade and vendor. Each software provider recommended both a range of minutes per week, and a minimum number of weeks for students to use the program.

[2] Students in kindergarten, 2nd and 3rd grade were matched on reading composite scores (BOY Comp) and students in 1st grade were matched on nonsense word fluency, correct letter sounds (NWF-CLS) scores.

Research Questions. The following research questions were used to guide the at-risk analysis:

1. How did the individual software vendors impact at-risk students' Acadience scores at the end of the year compared to students not participating in the vendor's program?
2. What impact did each software vendor have on students' literacy growth over the course of the school year, compared to students not participating in the program?

Statistical Modeling of Program Impacts on Acadience Test Scores. Ordinary least squares (OLS) regression models were computed for each analytic sample of at-risk students. 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 (effect sizes) for both vendors for each grade (K-3) served by the program. Effect sizes describe the magnitude of the difference between two groups on an outcome measure. We adapted a set of effect size benchmarks based on categories from Kraft (2020) that were adjusted for early literacy outcome measures: less than 0.10 is *small*, 0.10 to less than .30 is *medium* and .30 or greater is *large* (M. Kraft, personal communication, October 13, 2023).

RESULTS

In the following section we report the impact on Acadience literacy achievement, providing a separate look for how *Lexia* and *Amira* performed with at-risk students. All analyses compared at-risk students who used the vendor's software with matched at-risk students who did not. The vendor-specific analysis was designed to help USBE understand the effectiveness of the individual providers' ability to serve at-risk students.

Lexia

At-risk students in all grades K-3 were positively impacted by participation with *Lexia*, compared to the at-risk students not using the program, as evidenced by the higher predicted mean scores among the Treatment students compared to the Control students (**Table 3**). Treatment effects were most pronounced for at-risk kindergartners who met the vendor recommendations (Hedges'g= 0.51), where the effect size exceeded the threshold for large treatment effects. Treatment effects, although to a lesser degree, fell in the medium effect size category at first grade (Hedges' g=0.23) and third grade (g= 0.13), and in the small range for second grade (Hedges' g=0.09). **Table 3** presents the predicted means, mean score differences and effect sizes of *Lexia's* at-risk students who met the recommendations for usage and their matched comparison peers.

Table 3. Lexia Acadience Predicted EOY Mean Scores by Grade for at-risk students

Grade	Tr		Ctrl		Dif.	ES
	Mean	SE	Mean	SE		
Kindergarten	139.16	0.70	118.05	0.76	21.10	**<u>0.51</u>**
First Grade	64.99	0.38	58.86	0.41	6.13	*0.23*
Second Grade	155.83	1.09	149.98	1.19	5.86	0.09
Third Grade	262.53	1.32	253.14	1.44	9.39	*0.13*

Note. Model covariates were gender, White, special education, low-income, ELL, and BOY reading score. All data points displayed were statistically significant at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **<u>Large</u>**: bold and underlined text: .30 or greater.

Figures 1-4 show the learning growth over time among *Lexia* and comparison students for each grade. For kindergarten, even though the matched students started in a similar place, *Lexia* students who used the software as intended show a greater growth rate, about 21 points higher, in literacy achievement by the end of the year compared to their matched non-program counterparts. Additionally, *Lexia* kindergartens scored within the ‘at benchmark’ range (139)[3], compared to control students who averaged literacy levels below the expected range (118). First grade students also benefited from the program, ending the year just about 6 points higher in their predicted literacy Acadience scores compared to the matched peers. Both treatment and control first graders had end of year scores within the ‘at benchmark’ range (58-80)[4], with control students just making the cut (at 59).

Figure 1-4. At-Risk Students K-3 Acadience Scores Over Time Based on Lexia Participation

Figure 1. Lexia Kindergarten Growth

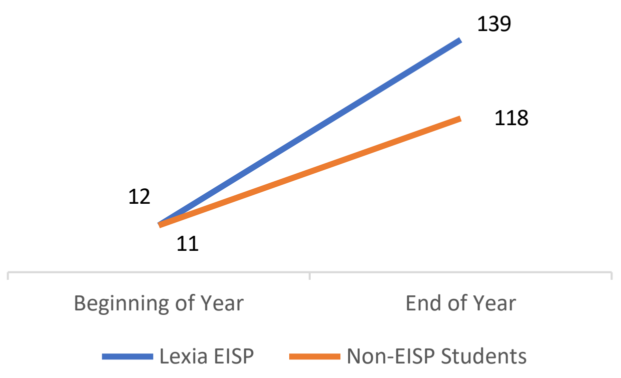
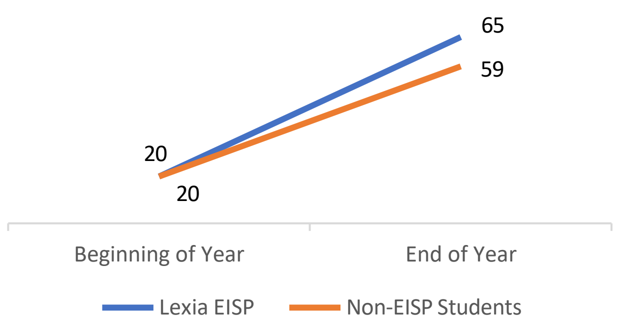


Figure 2. Lexia 1st Grade Growth



Second and third grade at-risk students also show growth over time with *Lexia*. Second grade students ended the year 6 points higher than the control students, and third grade *Lexia* students outperformed their control peers by 10 points. All at-risk second and third grade students (treatment and control), however, fell short of the ‘at benchmark’ range at the end of the year.[5]

[3] (Grade K) **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. (Dynamic Measurement Group, Inc. 2016).

[4] First grade **At Benchmark** (58-80) (for NWF-CLS).

[5] 2nd grade **At Benchmark** range of 238-286 and 3rd grade **At Benchmark** range of 330-404.

Figure 3. Lexia 2nd Grade Growth

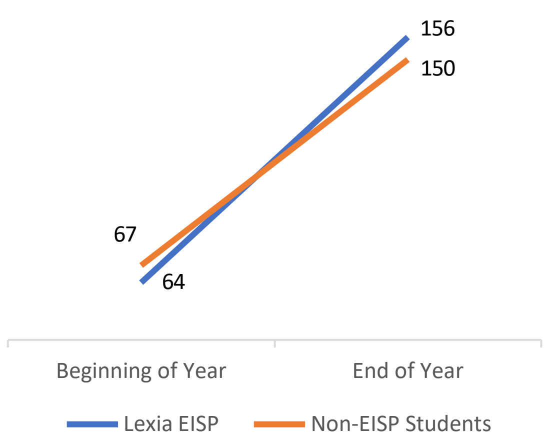
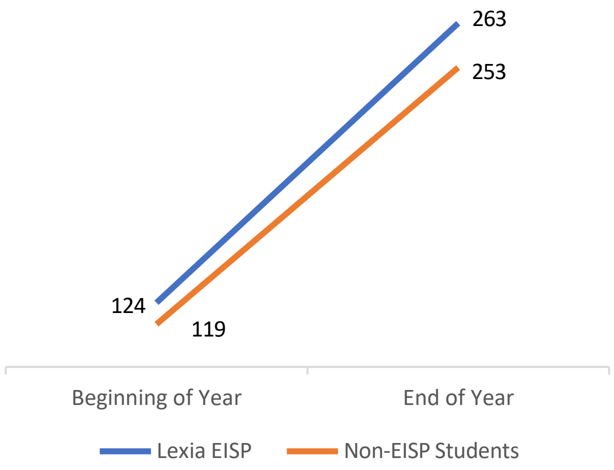


Figure 4. Lexia 3rd Grade Growth



Amira

We found that the at-risk students who used *Amira*, had higher predicted mean scores compared to at-risk students who did not use the reading software. Large treatment effects in first through third grade were demonstrated among students who met the usage requirements (Hedges' g = 0.47, 0.55, and 0.34) all exceeding the effect size threshold for large treatment effects (Hedges'g = 0.30 or greater). It should be noted that *Amira* had very small samples of at-risk students in each grade who were able to use the program as recommended, therefore these results should be interpreted with sample size in mind.

Table 4 presents the predicted means, mean score differences and effect sizes of *Amira's* at-risk students who met the recommendations for usage and their matched counterparts. As shown, at-risk students in grades 1-3 were positively impacted by their participation in *Amira*, compared to the at-risk students not using the program. *Amira* did not serve at-risk Kindergarten students this program year.

Table 4. Amira Acadience Predicted EOY Mean Scores by Grade for at-risk students

Grade	Tr		Ctrl		Dif.	ES
	Mean	SE	Mean	SE		
Kindergarten	--	--	--	--	--	--
First Grade	67.68	2.60	55.50	0.66	12.19	**0.47**
Second Grade	189.97	7.32	156.06	1.65	33.91	**0.55**
Third Grade	263.40	8.65	239.24	1.96	24.16	**0.34**

Note. Model covariates were gender, White, special education, low-income, ELL, and BOY score. All data points displayed were statistically significant at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater.

Figures 5-7 show the learning growth over time among *Amira* students and their comparison peers. All three grades of at-risk students were significantly impacted, outperforming their control peers in each grade. First grade *Amira* students who were at-risk achieved 'at benchmark' levels at the end of the year (68) while the control students remained below (56). Second and third graders dramatically outperformed their control peers over time, however, regardless of condition (treatment and control) both fell short of scoring within the 'at benchmark' range for their respective grade.

Figure 5-7. At-Risk Students 1st-3rd Acadience Scores Over Time Based on Amira Participation

Figure 5. Amira 1st Grade Growth

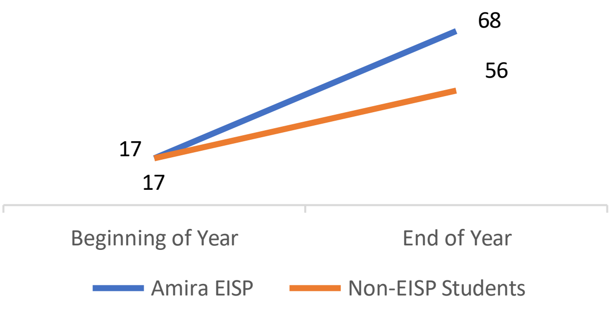


Figure 6. Amira 2nd Grade Growth

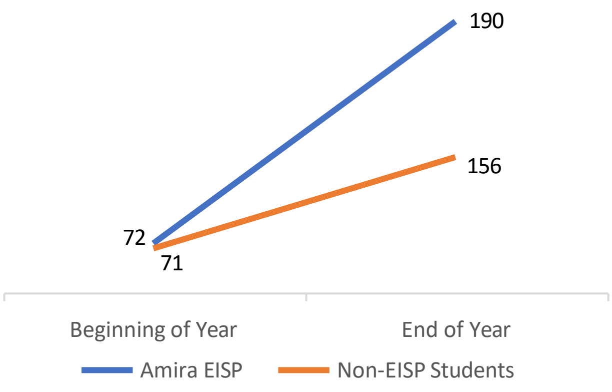
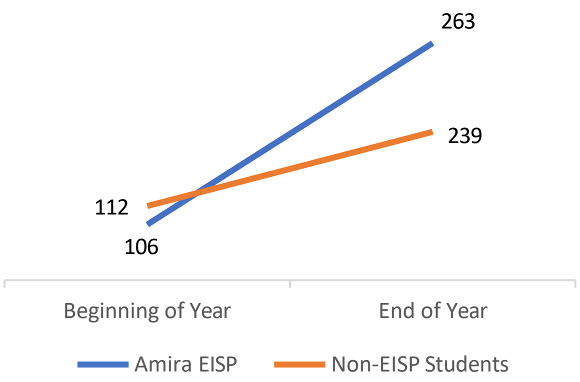


Figure 7. Amira 3rd Grade Growth



SUMMARY AND CONCLUSION

Often interventions are developed with at-risk students in mind and then quickly expand into populations of students less in need of additional resource. The current analysis reminds us that serving those with the greatest need can have positive and impactful outcomes.

There were two primary goals for the at-risk analysis from the 2022-2023 EISP program year: (1) to determine the impacts of the individual software providers on at-risk students' Acadience literacy achievement, and (2) examine the literacy growth rates over the course of the year.

Impacts on At-Risk Students

By Vendor

Results varied for each of the individual software providers. *Lexia* served at risk students in all grades and demonstrated positive trends in predicted mean scores across all 4 grades when comparing treatment students to control students. *Lexia's* treatment effects were considered to be large in kindergarten and medium in first and third grade. *Amira* did not serve kindergarten risk students during this program year, but had large impacts across first, second and third grade. The number of students served by *Amira* in general was much lower than Lexia, but among those who used the program as it was intended, the software had a significant effect on end of year literacy achievement for at-risk students.

Growth Rates

All students (treatment and control) in our matched samples started the school year with scores that were ‘below’ or ‘well below’ benchmark. In all cases where our model was statistically significant, we found that at-risk students who used the software program as intended, had a greater growth rate in literacy achievement by the end of the year compared to their matched at-risk peers. In kindergarten and first grade, both *Lexia* and *Amira* program students achieved scores within the ‘at benchmark’ range by the end of the school year. At-risk second and third grade treatment students also outperformed their control counterparts, however, fell short of the ‘at benchmark’ Acadience threshold at the end of the year.

Limitations

We do our best to control for all possible influences on student reading outcomes 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.

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.

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

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189x20912798

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.

APPENDIX A. Effect Sizes by Usage Group

Effect sizes describe the magnitude of the difference between two groups on an outcome measure. We adapted a set of effect size benchmarks based on categories from Kraft (2020) that were adjusted for early literacy outcome measures: less than 0.10 is *small*, 0.10 to less than .30 is *medium* and .30 or greater is *large* (M. Kraft, personal communication, October 13, 2023). There are multiple ways to interpret effect sizes, including the use of categories such as small, medium, or large (e.g., Cohen, 1988; Kraft, 2020), or using a minimum threshold (Hill 2008). Variations of both approaches are widely used and accepted, yet both require careful considerations of the research design and key study components (such as sample, measures, etc.). Our effect size interpretation approach uses a categorical range based on effect sizes for similar types of research, studying similar interventions (early literacy programs) and with similar populations (elementary students). Specifically, the range used in the current study represents the benchmarks for early literacy found in a summary of meta-analyses of relevant and similar educational studies, as well as the direct recommendation from the author (Kraft, 2020; M. Kraft, personal communication, October 13, 2023).

Table A1. MRU Effect Sizes

	Kindergarten	1st Grade	2nd Grade	3rd Grade
Lexia	**0.51**	*0.23*	0.09	*0.13*
Amira	--	**0.47**	**0.55**	**0.34**

Data source: Matched K-3 ITT, MRU80, MRU samples. All effect sizes displayed were statistically significant at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater. Kindergarten sample size –Lexia MRU (ctrl= 2,944.66, tr= 3,501); First Grade sample size- Lexia MRU (ctrl= 4,322.96, tr=5,039), Amira MRU (ctrl= 1,525.58, tr=98); Second Grade sample size Lexia MRU (ctrl= 2,780.64, tr=3,306), Amira MRU (ctrl= 1,368.29, tr=70); Third Grade sample size Lexia MRU (ctrl= 2,458.51, tr=2,923), Amira MRU (ctrl= 1,290.10, tr=66).

APPENDIX B. Amira Match Summary and Regression Results by Usage Group

Table B1. Amira Matched Treatment MRU Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	ISS						
	1	101	46%	70%	25%	43%	6%	17.75
	2	75	59%	61%	8%	47%	19%	71.49
	3	73	51%	44%	15%	71%	28%	109.96
Matched MRU Treatment Sample	K	ISS						
	1	98	47%	71%	24%	41%	5%	17.46
	2	70	59%	66%	9%	46%	17%	72.29
	3	66	53%	48%	14%	70%	26%	106.94

Note: Kindergarten had an insufficient sample size. The matched sample had a multivariate L1 score of 0.000000000000006945. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.0000000000000022), White (L1= 0.0000000000000044), SPED (L1 = 0.0000000000000022), Low-Income (L1= 0.0000000000000012), and ELL (L1= 0.0000000000000046).

Table B2. Amira MRU Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Met Recommended Use	K	Treatment	ISS				
		Control					
	1	Treatment	0.00	67.68	2.60	12.19	**<u>0.47</u>**
		Control		55.49	0.66		
	2	Treatment		189.97	7.32		
		Control		156.06	1.65		
	3	Treatment	0.006	263.40	8.65	24.16	**<u>0.34</u>**
		Control		239.24	1.96		

Note. Kindergarten had an insufficient sample size. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **<u>Large</u>**: bold and underlined text: .30 or greater. Data source: Matched K-3 MRU sample. First Grade sample size- ctrl= 1525.576, tr= 98; Second Grade sample size - ctrl= 1368.2906, tr= 70; Third Grade sample size – ctrl- 1290.103, tr=66.

Table B3. Amira Matched Treatment MRU80 Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	ISS						
	1	385	48%	69%	19%	50%	9%	17.68
	2	341	57%	60%	13%	49%	16%	66.67
	3	308	51%	50%	19%	56%	27%	116.13
Matched MRU 80 Treatment Sample	K	ISS						
	1	375	49%	70%	19%	49%	9%	17.55
	2	325	58%	62%	13%	49%	15%	66.38
	3	289	52%	53%	18%	56%	26%	114.95

Note: Kindergarten had an insufficient sample size. The matched sample had a multivariate L1 score of 0.00000000000000184. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.00000000000000014), White (L1= 0.000000000000006), SPED (L1 = 0.0000000000000016), Low-Income (L1= 0.0000000000000043), and ELL (L1= 0.0000000000000035).

Table B4. Amira MRU80 Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Met 80% of Recom mended Use	K	Treatment					
		Control	ISS				
	1	Treatment	0.002	60.57	1.35	4.65	*0.18*
		Control		55.91	0.57		
	2	Treatment		172.32	3.45		
		Control		148.26	1.41		
	3	Treatment	0.00	261.77	4.13	20.56	*0.29*
		Control		241.17	1.69		

Note. Kindergarten had an insufficient sample size. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater. Data source: Matched K-3 MRU80 sample. First Grade sample size- ctrl= 2099.85, tr= 375; Second Grade sample size - ctrl= 1947.704, tr= 325; Third Grade sample size – ctrl= 1731.959, tr=289.

Table B5. Amira Matched Treatment ITT Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	ISS						
	1	3,144	50%	63%	17%	48%	13%	17.37
	2	2,541	52%	60%	22%	52%	16%	56.13
	3	2,351	51%	58%	26%	50%	20%	109.19
Matched ITT Treatment Sample	K	ISS						
	1	2,997	51%	66%	16%	47%	12%	17.20
	2	2,444	53%	62%	21%	51%	15%	55.41
	3	2,264	51%	60%	25%	50%	19%	108.62

Note: The matched sample had a multivariate L1 score of 0.000000000000006291. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.0000000000000062), White (L1= 0.0000000000000056), SPED (L1 = 0.0000000000000015), Low-Income (L1= 0.0000000000000038), and ELL (L1= 0.000000000000018).

Table B6. Amira ITT Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Intent to Treat	K	Treatment	ISS				
		Control					
	1	Treatment	0.002	56.48	0.47	2.02	0.08
		Control		54.45	0.46		
	2	Treatment	0.000	145.62	1.26	9.86	*0.16*
		Control		135.76	1.28		
	3	Treatment	0.002	239.77	1.49	6.72	0.09
		Control		233.04	1.52		

Note. ES: Effect Size (based on Hedges G). Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater. Data source: Matched K-3 ITT sample. First Grade sample size - ctrl= 3,069.95, tr= 2,997; Second Grade sample size - ctrl= 2,352.21, tr= 2,444; Third Grade sample size – ctrl= 2,178.97, tr=2,264.

APPENDIX C. Lexia Match Summary and Regression Results by Usage Group

Table C1. Lexia Matched Treatment MRU Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	3,558	47%	64%	13%	41%	9%	11.80
	1	5,329	50%	70%	17%	39%	9%	20.10
	2	3,426	51%	68%	22%	42%	12%	67.01
	3	3,019	48%	66%	25%	43%	14%	124.03
Matched MRU Treatment Sample	K	3,501	47%	65%	13%	40%	9%	11.84
	1	5,039	50%	73%	16%	38%	9%	19.82
	2	3,306	51%	70%	21%	42%	11%	66.81
	3	2,923	47%	68%	24%	43%	14%	123.79

Note: The matched sample had a multivariate L1 score of 0.000000000000013. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.00000000000001), White (L1= 0.0000000000000045), SPED (L1 = 0.0000000000000037), Low-Income (L1= 0.0000000000000098), and ELL (L1= 0.0000000000000059).

Table C2. Lexia MRU Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Met Recom mended Use	K	Treatment	0.000	139.16	0.70	21.10	<u>**0.51**</u>
		Control		118.05	0.76		
	1	Treatment	0.000	64.99	0.38	6.13	*0.23*
		Control		58.86	0.41		
	2	Treatment	0.000	155.83	1.09	5.86	0.09
		Control		149.98	1.19		
	3	Treatment	0.000	262.53	1.32	9.39	*0.13*
		Control		253.14	1.44		

Note. ES: Effect Size (based on Hedges G). Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, <u>**Large**</u>: bold and underlined text: .30 or greater. Data source: Matched K-3 MRU sample. Kindergarten sample size – ctrl=2,944.66, tr= 3,051; First Grade sample size - ctrl= 4,322.96, tr= 5,039; Second Grade sample size - ctrl= 2,780.64, tr= 3,306; Third Grade sample size – ctrl= 2,458.51, tr=2,923.

Table C3. Lexia Matched Treatment MRU80 Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	4,730	48%	63%	13%	42%	11%	11.29
	1	7,183	50%	68%	18%	40%	10%	19.43
	2	4,896	51%	67%	22%	43%	13%	63.84
	3	4,590	48%	65%	26%	44%	14%	120.91
Matched MRU 80 Treatment Sample	K	4,666	48%	63%	13%	41%	10%	11.31
	1	6,814	50%	71%	18%	39%	9%	19.25
	2	4,728	51%	69%	22%	43%	12%	63.70
	3	4,428	48%	67%	25%	44%	14%	120.40

Note: The matched sample had a multivariate L1 score of 0.00000000000001052. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.000000000000025), White (L1= 0.0000000000000022), SPED (L1 = 0.00000000000001), Low-Income (L1= 0.000000000000014), and ELL (L1= 0.0000000000000025).

Table C4. Lexia MRU80 Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Met 80% of Recom mended Use	K	Treatment	0.000	134.10	0.81	17.06	**<u>0.34</u>**
		Control		117.04	0.78		
	1	Treatment	0.000	63.06	0.33	4.95	*0.18*
		Control		58.11	0.42		
	2	Treatment	0.032	149.44	0.91	3.17	0.05
		Control		146.27	1.16		
	3	Treatment	0.003	254.73	1.07	5.24	0.07
		Control		249.49	1.37		

Note. ES: Effect Size (based on Hedges G). Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **<u>Large</u>**: bold and underlined text: .30 or greater. Data source: Matched K-3 MRU80 sample. Kindergarten sample size – ctrl=2,827.20, tr=4,666; First Grade sample size - ctrl= 4,218.67, tr= 6,814; Second Grade sample size - ctrl= 2,864.77, tr= 4,728; Third Grade sample size – ctrl= 2,682.99, tr=4,428.

Table C5. Lexia Matched Treatment ITT Sample Demographics

	Grade	N	Female	Caucasian	SPED	Low Income	ELL	BOY Score
Total Treatment Sample	K	8,972	47%	62%	15%	43%	11%	10.63
	1	10,635	50%	66%	20%	41%	12%	18.47
	2	8,177	51%	66%	25%	43%	14%	60.33
	3	8,314	49%	66%	29%	43%	14%	116.21
Matched ITT Treatment Sample	K	8,872	47%	62%	15%	43%	11%	10.64
	1	10,031	50%	70%	19%	40%	11%	18.31
	2	7,872	51%	68%	24%	43%	13%	60.09
	3	8,035	48%	68%	29%	43%	13%	115.76

Note: The matched sample had a multivariate L1 score of 0.0000000000000258. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.000000000000045), White (L1= 0.000000000000031), SPED (L1 = 0.0000000000000076), Low-Income (L1= 0.000000000000045), and ELL (L1= 0.0000000000000069).

Table C6. Lexia ITT Regression Summary, by grade

	Grade	Condition	P value	Marginal Mean	St. Error	Diff.	ES
Intent to Treat	K	Treatment	0.000	122.55	0.45	7.43	*0.18*
		Control		115.12	0.75		
	1	Treatment	0.000	59.62	0.26	3.20	*0.12*
		Control		56.42	0.44		
	2	Treatment	0.835	140.96	0.70	0.28	0.01
		Control		140.68	1.16		
	3	Treatment	0.949	243.12	0.79	-0.10	-0.001
		Control		243.22	1.32		

Note. ES: Effect Size (based on Hedges G). Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater. Data source: Matched K-3 ITT sample. Kindergarten sample size – ctrl=3,212.15, tr= 8,872; First Grade sample size - ctrl= 3,716.95, tr= 10,031; Second Grade sample size - ctrl= 2,850.10, tr= 7,872; Third Grade sample size – ctrl= 2,909.11, tr=8,035.

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