2025

Implementation and Outcome Evaluation of the *Early Interactive Reading Software Program (EISP)*

Prepared for
the Utah State Board of Education

THE UNIVERSITY OF UTAH
UTAH EDUCATION POLICY CENTER

Bridging Research, Policy, & Practice

The Utah Education Policy Center (UEPC) is an independent, non-partisan, not-for-profit research-based center at the University of Utah founded in the Department of Educational Leadership and Policy in 1990 and administered through the College of Education since 2007. The UEPC's mission is to bridge research, policy, and practice by conducting rigorous and comprehensive research and evaluations, and providing expert and research-informed technical assistance and professional learning. We empower educators, policymakers, and leaders to make research actionable and impactful to transform education across early childhood education, K-12 schools, and higher education.

We are committed to supporting the understanding of whether educationalpolicies, programs, and practices are being implemented as intended, whether they are effective and impactful, and how they may be improved and scaled-up and become sustainable.

Please visit our website for more information about the UEPC:
http://uepc.utah.edu

Andrea K. Rorrer, Ph.D., Director
andrea.rorrer@utah.edu

Cori Groth, Ph.D., Associate Director
cori.groth@utah.edu

Ellen Altermatt, Ph.D., Assistant Director for Research and Evaluation
ellen.altermatt@utah.edu

T. W. Altermatt, Ph.D., Assistant Director and Lead Data Scientist
bill.altermatt@utah.edu

Citation: Altermatt, T. W., Gallyer, J., Altermatt, E. R., Yildiz, M., & Rorrer, A. K. (2025). *Implementation and Outcome Evaluation of the Early Interactive Reading Software Program (EISP).* Salt Lake City, UT: Utah Education Policy Center.

Acknowledgements

The Utah Education Policy Center (UEPC) is grateful to Amber Wright, Digital and Instructional Materials Specialist at the USBE, for her insights, responsiveness, and collaboration throughout this this project. We also acknowledge the participating Early Intervention Software Program vendors for their cooperation in providing relevant data needed to complete this evaluation.

Table of Contents

List of Figures

List of Tables

1 | Executive Summary

1.1 Overview

This report presents findings from the Utah Education Policy Center's (UEPC's) 2024-2025 evaluation of the Early Interactive Software Program (EISP), a state-supported initiative to strengthen literacy skills among students in grades K through 3. The evaluation focused on both implementation (number of students, amount of usage) and impact (change in reading proficiency as a result of using early literacy software).

1.2 Methods

Using data sharing agreements between vendors and the Utah State Board of Education (USBE) and a Master Data Sharing Agreement between the UEPC and USBE, the UEPC connected data provided by vendors on weekly student use of early literacy software with data provided by the USBE on student demographics, school information, and scores on the Acadience Reading standardized assessment. Implementation was evaluated by tabulating the number of students, schools, and LEAs served by each vendor as well as the mean level of student engagement with the software per week and the mean number of weeks per year that the software was used. These were compared to vendor-supplied cutoffs for the minimum recommended use, and reported as the percentage of students who met 80%, 100%, 200%, and 300% of vendor recommendations. In addition to examining base rates of implementation, we also tested whether implementation varied significantly across student and school-level characteristics. Impact was evaluated using a method designed to reduce the correlation between student characteristics and early literacy software use: covariate balancing propensity score weighting. Like matching and random assignment, weighting increases confidence in cause-and-effect conclusions between early literacy software use and learning gains by controlling for other variables that might systematically co-vary with reading software use. The relationship between "dose" (level of software use) and "response" (learning gains) was modeled using statistical regression tailored to the weighting process. Finally, we examined how the dose-response relationship varied across student- and school-level characteristics.

1.3 Key Findings

1. **EISP shows strong evidence of effectiveness at improving early literacy**. Compared to non-users, students who used early literacy software at 100% of vendor recommendations showed increases in reading that were "large" for kindergartners (d = 0.26), "moderate" for 1st and 3rd graders (d = .09 and .05, respectively), but not significantly different from zero for 2nd graders. Second grade users did see significant but small effects of usage (d = .04) at 200% of vendor recommendations.
2. **EISP helps struggling students catch up to their peers**. Early literacy software usage had a significantly greater impact for multilingual learners, students receiving special education, and students with the lowest beginning-of-year scores on Acadience Reading. This pattern was not always significant at every grade level but was observed in kindergarten for multilingual learners; in kindergarten, 2nd grade, and 3rd grade for students receiving special education services; and for all four grade levels for students with the lowest scores on the beginning of the year Acadience Reading assessment. Higher levels of use were associated with smaller gaps between student groups.
3. **Implementation of EISP is broad.** 73% of all Utah public school students in kindergarten through 3rd grade participated in EISP.

4. **Students are highly concentrated in a small number of vendors**. One vendor (Lexia) accounted for 59% of all student users and the top two vendors accounted for 83%.
5. **Implementation varied dramatically across vendors**. Three vendors (including the two largest) had average student usage above 1,400 minutes (23 hours) for the year. Three vendors had average student usage below 300 minutes (5 hours) for the year.
6. **Implementation varied across student groups**. Students receiving special education services, Native American students, economically disadvantaged students, Pacific Islander students, Hispanic students, and multilingual learners tended to use literacy software less than their comparison groups.[1] Groups with significantly higher levels of usage than non-members were students with higher beginning-of-the-year Acadience Reading scores, Asian students, and students in schools with a higher percentage of the student body who were multilingual learners.

1.4 Recommendations

1. **Continue to support EISP**. Given the goal of achieving 70% on-grade-level reading for 3rd graders across the state by 2027, current reading-on-grade-level rates for 3rd graders around 50%, strong evidence that using early literacy software leads to gains in reading proficiency, and an already wide deployment of early literacy software, continuing to support EISP has a high probability of making progress toward the goal. If early literacy software use were reduced to zero from its current level, we estimate that rates of reading on grade level for 3rd graders would fall by 4.3 percentage points.
2. **Focus on improving EISP participation among lower-performing students to reduce reading gaps**. The EISP program showed the strongest impact for multilingual learners, special education students, and lower-performing readers. These groups historically show smaller learning gains, suggesting the potential to close achievement gaps. However, multilingual learners and special education students used the program less, so without targeted engagement efforts, the benefits may not reach those who need them most.
3. **Integrate EISP within Utah's broader strategy for improving early literacy**. Coordinating EISP with literacy coaching and professional development is likely to increase the efficacy of each initiative by aligning and reinforcing instructional goals and methods.
4. **Provide ongoing implementation support and monitoring to ensure high-quality use of reading software.** Given the importance of high usage levels for raising early literacy scores, teachers should be provided with clear guidelines for satisfactory student use, tools to track student engagement, training on how to use the software, and technical support.
5. **Study how impact could be increased**. Many questions remain about the relationship between software use and early literacy. Research is needed to uncover why 2nd graders do not show the same gains as students in other grades, why some vendors show greater implementation than others, and how teacher implementation practices and contextual factors affect impact.

[1] For racial and ethnic groups, the comparison group was White students. For other groups, it was students who did not possess the trait in question (e.g., receipt of special education services, economic disadvantage).

2 | Introduction

In this report, the Utah Education Policy Center (UEPC) presents findings from the 2024-2025 evaluation of the *Early Interactive Reading Software Program (EISP).* The UEPC was contracted by the Utah State Board of Education to evaluate the implementation and impact of EISP annually between 2025 and 2030.

Summary: Utah's Early Interactive Software Program (EISP) provides funding for K-3 literacy software across the state. This evaluation uses rigorous methods to assess both how the program is being implemented and whether it improves student reading outcomes. The evaluation is intended to inform decision-making about the effectiveness of the EISP intervention with particular attention to whether benefits extend equally across all student groups.

2.1 Program Overview

Utah has a history of investment in digital learning initiatives to improve early literacy. In 2012, House Bill 513 provided funding for interactive computer software focusing on literacy and numeracy for students in kindergarten and 1st grade. In 2022, Utah Senate Bill 127 established a goal for 70% of students to achieve 3rd grade-level proficiency on the state-administered reading assessment by July 1, 2027. The use of interactive software to improve literacy is a part of the initiatives supported by SB 127, along with a comprehensive professional development program for teachers on the Science of Reading and the assignment of additional literacy coaches to schools.

EISP is authorized under Utah Code 53F-4-203, which provides that the Utah State Board of Education (USBE) shall distribute funds to public schools for the purchase of personalized interactive early literacy software licenses. It also requires the USBE to contract with an independent evaluator to "determine the extent to which a public school uses the early interactive early literacy software" and to "evaluate a student's learning gains as a result of using early interactive early literacy software" using an assessment that a provider of early interactive early literacy software does not develop. No more than 6% of the funds may be used for administrative and evaluation costs; the remaining 94% are distributed directly to Local Education Agencies (LEAs).

For several years, the USBE maintained a curated list of approved software vendors from which LEAs selected products. However, Senate Bill 44 (2023 General Session) amended Utah Code 53F-4-203 by clarifying that the previously applied requirement to demonstrate an effect size of at least 0.40 does *not* apply to early literacy software.[2] As a result of SB 44, beginning with the 2023–2024 academic year (AY), LEAs were granted greater autonomy to independently select and procure interactive early literacy software for K–3 students. To qualify for reimbursement under EISP, LEAs must submit applications to the USBE for their chosen product. In addition, vendors must track student-level usage with State Student IDs (SSIDs) and participate in external evaluation activities to remain qualified for reimbursement.

[2] The 0.40 effect size requirement is described in Utah Code 53G-11-302.

In AY 2024-2025, 12 vendors participated in EISP—an increase from 11 vendors in AY 2023-2024. The programs adopted by schools included: (1) Amira Learning, (2) Core5 Reading (Lexia), (3) Renaissance Learning, (4) i-Ready (Curriculum Associates), (5) Imagine Language and Literacy (Imagine Learning), (6) IXL Language Arts, (7) MobyMax, (8) My Reading Academy (Age of Learning), (9) Read Naturally, (10) Reading Horizons, (11) Really Great Reading, and (12) Waterford Reading.[3]

2.2 Evaluation Overview

For this year's EISP evaluation, the UEPC focused on two sets of evaluation questions to examine 1) how EISP was implemented during the 2024-2025 school year and 2) the impact of the program on student scores on statewide reading assessments. Specifically, the UEPC examined the following implementation and impact evaluation questions.

2.2.1 Implementation Evaluation Questions

- **RQ1.** How many students are using early literacy software at schools that have received funding through the program? How is this distributed across vendor and grade level? How many unique schools and LEAs are served by each vendor?
- **RQ2.** How engaged are users (e.g., as measured by minutes of usage or number of units completed), and how does this level of engagement compare to recommendations set by vendors?
- **RQ3.** How does usage vary by student demographic characteristics, prior student achievement, or school demographic characteristics?

2.2.2 Impact Evaluation Questions

- **RQ4.** Do students who are using early literacy software at higher levels (and also specifically at levels recommended by vendors) show stronger achievement outcomes as measured by scores on Acadience Reading assessments than a matched comparison group of students who do not use the software at levels recommended by vendors?
- **RQ5.** Does the impact of early literacy software usage vary by student demographic characteristics, prior student achievement, or school demographic characteristics?

2.3 Data

To answer these implementation and impact evaluation questions, the UEPC utilized software usage data provided by participating vendors and demographic and achievement data provided by the USBE. These data sources are described below.

2.3.1 Software Usage Data

As part of their participation in EISP, software vendors are responsible for supplying student usage data to the UEPC as the program's independent evaluator. These data include State Student Identifiers (SSIDs) and the number of minutes of usage for each week that a student used the software. Vendors are also responsible for providing their product's thresholds for fidelity of student use in terms of minutes per week and weeks per year.

[3] Data from *Renaissance* was not available at the time of this report. An addendum with their results will be attached to this report when data are available.

2.3.2 Demographic and Achievement Data

SSIDs from software usage data were matched to student enrollment records using a Master Data Sharing Agreement (MDSA) between the UEPC and the USBE.[4] The student enrollment records used for this study included student demographic information (race and ethnicity, gender, eligibility for free or reduced-price lunch, multilingual learner status, and receipt of special education services), school information (percent of students at the school with each of the student-level characteristics, e.g., percent of students who were multilingual learners), and beginning-of-year and end-of-year student scores on the Acadience Reading assessment. Acadience Reading is a formative assessment used in Utah for benchmarking and monitoring progress of students in K-3.

2.4 Report Organization

This evaluation report is divided into seven sections. In Section 1, we provided an executive summary, which includes an overview of evaluation goals, methods, key findings, and recommendations. In Section 2, we provide an introduction to the *Early Interactive Reading Software Program* (EISP) and to the UEPC's evaluation of this program. In Section 3, we offer background for the current report by providing a brief review of the research and evaluation literatures that have sought to understand the role that educational technology – including learning software programs – might play in improving student outcomes in reading. In Section 4, we summarize findings related to program implementation. In Section 5, we summarize findings related to program impact. In Section 6, we offer conclusions and future directions for research that would lead to improvements in the program. Finally, in Section 7, we provide references for all publications and reports cited in this annual report.

2.5 Intended Audience

The primary audiences for this report include personnel from the USBE with expertise and interest in early literacy, outcomes in English/Language Arts, technology-enabled instruction, and personalized competency-based learning; learning software providers; and legislators. In addition, this evaluation report provides insights that educational leaders and educators from LEAs participating in the program will find useful for guiding EISP implementation and literacy improvement efforts. Overall, the evaluation report provides useful information to inform state- and LEA-level decision-making about the role of early literacy software in advancing early reading proficiency, including how best to target resources, strengthen professional development, and ensure that all of Utah's students have the skills they need to be successful in fourth grade and beyond. More broadly, the results can offer insights to other states considering similar investments in technology-enabled literacy interventions.

2.6 Contributions of the Current Evaluation

Evaluating the implementation and impact of early literacy software is important given the commitment of the state and Utah State Board of Education to improving early literacy outcomes (USBE, 2024). The current evaluation contributes to the existing evidence base on interventions to improve early literacy outcomes in several ways. First, unlike many studies that are limited to small samples, this evaluation examines statewide implementation of EISP. Second, the study uses a rigorous quasi-experimental

[4] The Utah Education Policy Center has a Master Data Sharing Agreement (MDSA) with the Utah State Board of Education permitting the use of education data for evaluation and research purposes. The UEPC adheres to the terms of the MDSA, including confidentiality, non-disclosure, data security, monitoring, and applicable state and federal laws and regulations. The UEPC also adheres to University of Utah Institutional Review Board provisions to protect student privacy and does not report any personally identifiable information.

design to examine associations between early literacy software use and student scores on statewide reading assessments. Unlike approaches that simply compare student software users to the state average or to non-users, a quasi-experimental approach attempts to provide an apples-to-apples comparison, isolating the effect of software use from other factors that tend to co-vary with it. Third, the evaluation investigates whether software use and its benefits extend equally across student groups, with particular attention to students from low-income households, students with disabilities, and multilingual learners.

3 | Relevant Background Research

Summary: Achieving reading proficiency by 3rd grade is predictive of positive outcomes, including later academic performance across subject areas. In 2024, only 48.2% of 3rd graders in Utah were reading on grade level. Early literacy software programs are one strategy that can help improve rates of students reading on grade level.

3.1 The Importance of Early Literacy Skills

Early literacy skills are widely recognized as a critical foundation for later academic success. Students who struggle with reading by the end of 3rd grade are more likely to fall behind across subject areas as instruction shifts from "learning to read" to "reading to learn" (Harlaar, Dale, & Plomin, 2007). These early difficulties can create a cascade of challenges that reduce the likelihood that students will graduate from high school and pursue postsecondary education (Annie E. Casey Foundation, 2010; Lesnick, Goerge, & Smithgall, 2010). Research also links lower literacy levels with longer-term outcomes such as fewer employment opportunities and lower community and civic participation (Kutner, Greenberg, Jin, Boyle, Hsu, & Dunleavy, 2007).

Despite the importance of early literacy, many students struggle to read at grade level. In Utah, proficiency rates currently remain well below targets on the Acadience Reading assessment, with only 48.2% of students reading on grade level in 3rd grade in 2024 (Utah State Board of Education (USBE), 2024).[5] State trends mirror national patterns where, in 2024, 4th grade reading scores were lower than scores in both 2022 and 2019, with proficiency rates below 20% for economically disadvantaged students, students with disabilities, and students identified as English learners (National Center for Education Statistics, 2024).

3.2 Literacy Software as an Intervention Strategy

In an effort to improve early literacy outcomes and address achievement gaps that were exacerbated by the pandemic (Kuhfeld & Lewis, 2022, 2024; Utah State Board of Education and the National Center for the Improvement of Educational Assessment, 2021), schools across the country are increasingly investing in learning software to provide individualized English Language Arts (ELA) support in the early grades (Dahl-Leonard, Hall, & Peacott, 2024). Proponents of learning software cite several potential advantages. First, learning software can deliver personalized practice that is aligned to each child's current skill level, offering structured progression through and immediate feedback on key reading skills. Second, software generates real-time data that teachers can use to monitor student growth, identify areas of need, and adjust instruction accordingly. Finally, interactive and gamified elements of learning software are designed to increase student motivation and engagement (Bernacki, Greene, & Lobczowski, 2021; Brizard, 2023; Huebner & Burstein, 2023; Pane, Steiner, Baird, & Hamilton, 2017; Van Schoors, Elen, Raes, Vanbecelaere, & Depaepe, 2023; Zheng, Long, & Gyasi, 2022).

Several recent meta-analyses report small-to-moderate positive effects of software use on a range of early literacy skills, including phonological awareness, word reading, reading comprehension, and

[5] See state report card: https://reportcard.schools.utah.gov/State/EarlyLiteracy/?StateID=99&SchoolLevel=K8&schoolyearendyear=2024

reading fluency (Jamshidifarsani, Garbaya, Lim, & Blazevic, 2019; McTigue, Solheim, Zimmer, & Uppstad, 2019; Verhoeven, Voeten, & Segers, 2022). However, these effects are not uniform. The effectiveness of software interventions—like other interventions—appears to be moderated by a variety of factors, including implementation fidelity and quality of teacher training (e.g., Archer, Savage, Sanghera-Sidhu, Wood, Gottardo, & Chen, 2014). Moreover, concerns have emerged about differences across groups of students in levels of access to and engagement with learning software, raising important concerns about whether the digital learning ecosystem is providing consistent benefits to all students or unintentionally contributing to the widening of achievement gaps (Altermatt, Altermatt, Rorrer, & Moore, 2022; Altermatt, Yildiz, & Rorrer, 2025; Reich & Ito, 2017).

4 | Implementation Evaluation

This section addresses the implementation evaluation. The following research questions (RQs) are answered:

> **RQ1**. How many students are using early literacy software at schools that have received funding through the program? How is this distributed across vendor and grade level? How many unique schools and LEAs are served by each vendor?
>
> **RQ2**. How engaged are users (e.g., as measured by minutes of usage or number of units completed), and how does this level of engagement compare to recommendations set by vendors?
>
> **RQ3.** How does usage vary by student demographic characteristics, prior student achievement, or school demographic characteristics?

4.1 RQ1: Enrollment

- ***RQ1:*** *How many students are using early literacy software at schools that have received funding through the program? How is this distributed across vendor and grade level? How many unique schools and LEAs are served by each vendor?*

Summary: A total of 141,854 unique students in kindergarten through 3rd grade during the 2024-2025 school year used early literacy software for one minute or more. This represents 73% of all public-school students in Utah in K through 3rd grade. Students are highly concentrated in a small number of vendors: One vendor (Lexia) accounted for 59% of all student users and the top two vendors accounted for 83%. Most vendors have students evenly spread across grade levels but some concentrate on kindergarten (Waterford) or on grades 2 and 3 (Read Naturally). The number of schools and LEAs served by the vendors ranged from 412 unique schools in 79 LEAs for the largest vendor, to one school in one LEA for the smallest participating vendor.

Table 1 reports the number of unique students reported by vendors, the number whose SSIDs could be matched to USBE records, the number with software usage for the year that did not exceed one minute ("zero usage"), the number who appeared in the user lists of more than one vendor ("shared students"), reflecting the degree of student exposure to multiple platforms), the number of unique schools served by the vendor, and the number of unique LEAs served by the vendor.

Table 1. 2024-2025 Program Enrollment Overview

Vendor	Reported Students	Matched SSID[a]	Zero Usage[b]	Shared Students[c]	Unique Schools	Unique LEAs
Lexia	89,642	88,620	0	6,579	412	79
i-Ready	36,399	36,018	0	2,358	118	20

Vendor	Reported Students	Matched SSID[a]	Zero Usage[b]	Shared Students[c]	Unique Schools	Unique LEAs
Amira	8,768	8,222	1	2,815	97	7
Really Great Reading	5,105	4,945	1,653	1,970	17	1
Imagine Language and Literacy	5,032	4,916	1,964	2,353	22	9
Waterford Reading	4,287	4,260	0	723	44	11
Age of Learning	1,496	1,485	803	846	5	1
IXL Language Arts	598	592	0	530	3	3
Reading Horizons	274	179	1	101	3	3
MobyMax	137	136	4	9	1	1
Read Naturally	55	55	0	55	4	1
Renaissance	-	-	-	-	-	-

Note: [a]Number of students whose State Student Identifiers (SSIDs) could be matched to USBE student records. [b]Number of students whose total software usage for the year was zero and had an SSID that matched USBE records. [c]Number of students who appeared in the user lists of more than one vendor and had an SSID that matched USBE records. Data for *Renaissance* was not available at the time of this report and will be reported in an addendum when their data become available.

Table 1 shows large differences across vendors in the number of students served. *Lexia* alone accounts for 59% of all SSID-matched students in EISP and, together with *i-Ready,* comprises 83%. As a result of their control over such a large share of the student market in Utah, reports of the overall implementation and impact of EISP will be dominated by these two vendors.

Only three vendors reported students who had less than one minute of usage for the year: *Really Great Reading*, *Imagine Language and Literacy*, and *Age of Learning* (see Table 1). Having students with zero usage is not by itself an indication that anything is wrong with a program and may reflect decisions by the school or teacher over which the vendors have no control. In addition, some vendors may simply not report students with zero usage, so comparisons between vendors on this measure are not advised. However, the vendors with zero usage in Table 1 may want to investigate why such a large percentage of their reported student users (between 33% and 54%) never engaged with their product because it could reflect challenges with implementation.

Table 2 reports the number of students served by each vendor by grade level.

Table 2. 2024-2025 Program Enrollment by Grade

Vendor	K	1st	2nd	3rd
Lexia	18,434	22,354	23,751	24,081
i-Ready	7,776	8,897	9,377	9,968
Amira	956	2,542	2,528	2,196
Really Great Reading	966	1,055	1,366	1,558
Imagine Language and Literacy	1,099	1,219	1,293	1,305
Waterford Reading	2,865	823	549	23
Age of Learning	370	350	390	37
IXL Language Arts	140	153	143	156
Reading Horizons	38	60	41	40

Vendor	K	1st	2nd	3rd
MobyMax	31	36	34	35
Read Naturally	0	0	41	14

Note: Counts reflect the number of unique student users with non-zero usage for the year whose SSIDs could be matched to USBE records. Counts include students who appeared in multiple vendor lists because otherwise, *Read Naturally* would have zero students. Data for *Renaissance* was not available at the time of this report and will be reported in an addendum when their data become available.

Table 2 shows that the number of users by vendor sometimes varies considerably across grade levels. *Read Naturally*, for example, has no student users in kindergarten or 1st grade. *Waterford Reading* and *Age of Learning*, in contrast, have very few students in 3rd grade. Other vendors such as *Lexia, i-Ready*, and *Imagine Language and Literacy* have usage that is fairly balanced across grade levels.

4.2 RQ2: Implementation

- ***RQ2:*** *How engaged are users (e.g., as measured by minutes of usage or number of units completed), and how does this level of engagement compare to recommendations set by vendors?*

Summary: The total number of minutes that students used literacy software for the school year varied dramatically across vendors. Three vendors (including the two largest) had average student usage above 1,400 minutes (23 hours) for the year. Three vendors had average student usage below 300 minutes (5 hours) for the year. Two vendors did not record minutes and instead recorded the number of questions answered or stories completed. Three vendors (including the two largest) had 47% or more of students using software at or above their minimum usage recommendations. Five vendors had 5% or fewer of students using software at or above minimum recommendations.

4.2.1 Vendor Use Recommendations

Implementation is important to assess because for a product to improve student literacy, it must not only be effective but also used appropriately. However, comparing usage across vendors presents significant challenges. Each vendor recommends different minimum usage levels for "fidelity," which is the degree to which the product is used as intended. Table 3 shows the recommended minutes per week and weeks per school year that each vendor considers necessary to impact literacy achievement. *Read Naturally* is not reported in Table 3 because they do not have minimum use expressed in minutes but rather recommend that students complete at least 1.5 to 2 stories per week, with a minimum of 24 stories completed during the year. *IXL Language Arts* does not report minutes, but instead reports number of questions answered. They recommend students answer at least 15 questions per week. *Lexia* is excluded from Table 3 because they do not have a constant recommendation but instead calculate user-specific recommendations that vary over time based on the student's interactions with the platform. We accounted for this variation in recommendations when calculating percent meeting recommendations (see Table 6). *Really Great Reading* recommends a range of 25 to 30 weeks of instruction, so the midpoint of 27.5 is used here. They also recommend a range of 30 to 40 minutes of student practice per week, so the midpoint of 35 is used here.

Table 3. Vendor Use Recommendations

Vendor	K	1st	2nd	3rd	Weeks
i-Ready	30	30	30	30	20
Amira	20	20	20	20	25
*Really Great Reading**	35	35	35	35	27.5
Imagine Language and Literacy	40	50	50	50	18
Waterford Reading	75	75	75	75	28
Age of Learning	45	45	45	45	16
Reading Horizons	60	60	60	60	8
MobyMax	30	30	30	30	36

Note: Data for *Renaissance* was not available at the time of this report and will be reported in an addendum when their data become available.

Complications arising from the differences among vendors in recommended use are discussed in Appendix B.

4.2.2 Program Use

Table 4 provides mean levels of usage (in minutes per week, total minutes, and number of weeks) by vendor and grade level. Table 4 shows a large degree of variation in usage across vendors. For example, *i-Ready* users had an average total usage of 2,067 minutes compared to only 31 average total minutes for *Reading Horizons.* Our impact evaluation (*Section 5*) is sensitive to this variation and considers how much students increase in reading scores *per unit of usage,* which allows us to include even students with very low levels of usage in our analysis.

Table 4. 2024-2025 Program Use by Vendor and Grade for Vendors Reporting Minutes

Vendor	Grade	*N*	Avg Weekly Minutes	Avg Total Minutes	Avg Weeks of Use
Lexia	K	18,434	45.91	1,342.82	26.98
	1	22,354	51.04	1,607.43	29.84
	2	23,751	46.87	1,455.3	29.31
	3	24,081	44.09	1,308.88	27.04
	Total	88,620	46.97	1,430.49	28.34
i-Ready	K	7,776	48.21	1,414.05	26.79
	1	8,897	66.6	2,252.04	31.53
	2	9,377	69.43	2,353.09	31.5
	3	9,968	68.02	2,140.94	29.59
	Total	36,018	63.76	2,066.69	29.96
Amira	K	956	14.3	292.66	17.36
	1	2,542	16.84	332.31	17.34
	2	2,528	20.17	466.95	19.37
	3	2,196	18.65	454.45	19.29
	Total	8,222	18.05	401.72	18.49
Really Great Reading	K	966	18.95	109.07	4.85

Vendor	Grade	*N*	Avg Weekly Minutes	Avg Total Minutes	Avg Weeks of Use
	1	1,055	17.54	243.72	12.64
	2	1,366	20.37	317.46	13.28
	3	1,558	23.05	351.87	12.44
	Total	4,945	20.46	271.86	11.23
Imagine Language and Literacy	K	1,099	38.11	835.7	19.27
	1	1,219	40.2	792.14	17.28
	2	1,293	37.27	603.04	12.55
	3	1,305	36.57	352.83	7.4
	Total	4,916	38.27	635.52	13.86
Waterford Reading	K	2,865	60.6	1,915.9	29.77
	1	823	50.62	1,298.24	22.43
	2	549	46.01	1,015.77	16.61
	3	23	32.74	376.43	10.35
	Total	4,260	56.64	1,672.26	26.55
Age of Learning	K	370	24.91	325.41	10.01
	1	350	25.19	412.22	12.52
	2	390	10.28	66.45	3.29
	3	375	8.93	18.44	0.81
	Total	1,485	19.66	200.34	6.51
Reading Horizons	K	38	4.22	23.08	3.29
	1	60	8.96	50.81	4.75
	2	41	3.29	19.95	4.34
	3	40	3.76	20.1	2.72
	Total	179	5.5	30.99	3.89
MobyMax	K	31	46.25	1,264.61	20.42
	1	36	32.86	561.56	11.56
	2	34	43.78	1,028.85	16.18
	3	35	33.68	689.29	16.77
	Total	136	38.95	871.51	16.07

Note: *Read Naturally* is excluded from the table because it does not track minutes but instead tracks number of stories that a student has read. *IXL Language Arts* does not track minutes but instead tracks number of questions answered. Data for *Renaissance* was not available at the time of this report and will be reported in an addendum when their data become available.

Mean levels of use for vendors who do not record minutes are presented in Table 5. Note that *Read Naturally* was not implemented in any kindergarten or 1st grade classrooms. The large differences in units completed between vendors reflects the complications of comparing across vendors when the meaning of a "unit" can be so different.

Table 5. 2024-2025 Program Use by Vendor and Grade for Vendors Not Reporting Minutes

Vendor	Grade	*N*	Avg Units Completed	Avg Total Units Completed	Avg Weeks of Use
Read Naturally (units: stories)	K	0	-	-	-
	1	0	-	-	-
	2	41	1.53	7.46	4.93
	3	14	1.47	5.43	3.93
	Total	55	1.52	6.95	4.67
IXL Language Arts (units: questions)	K	140	15.06	204.7	6.54
	1	153	36.98	543.57	10.5
	2	143	38.71	485.69	11.06
	3	156	39.26	701.56	13.55
	Total	592	32.82	491.08	10.5

4.2.3 Percent Meeting Recommendations

Because of the different standards and recommendations for weekly use across vendors (see Table 3), it is difficult to evaluate the average minutes of use reported in Table 4 with regard to whether students are using the software above or below expectations. To address this concern, we standardize student usage by considering it in the context of each vendor's level of recommended usage. Table 6 reports the percentage of students whose usage was at four levels relative to vendor recommendations: 80% (below recommendations), 100% (meeting recommendations), 200% (twice the recommended level), and 300% (three times the recommended level).

Table 6. Percentage of Students Meeting Vendor Recommendations for Use

Vendor	Grade	*N*[a]	% at 80% of Rec.[b]	% at 100% of Rec.[b]	% at 200% of Rec.[b]	% at 300% of Rec.[b]
Lexia	K	18,434	59%	50%	22%	7%
	1	22,354	65%	53%	22%	8%
	2	23,751	60%	48%	19%	7%
	3	24,081	52%	40%	13%	4%
	Total	88,620	59%	47%	19%	7%
i-Ready	K	7,776	76%	63%	5%	0%
	1	8,897	91%	84%	16%	0%
	2	9,377	91%	84%	20%	0%
	3	9,968	88%	82%	10%	0%
	Total	36,018	87%	79%	13%	0%
Amira	K	956	23%	16%	0%	0%
	1	2,542	24%	16%	0%	0%
	2	2,528	39%	30%	0%	0%
	3	2,196	35%	27%	0%	0%
	Total	8,221	31%	23%	0%	0%
Really Great Reading	K	966	2%	0%	0%	0%

Vendor	Grade	*N*[a]	% at 80% of Rec.[b]	% at 100% of Rec.[b]	% at 200% of Rec.[b]	% at 300% of Rec.[b]
	1	1,055	7%	2%	0%	0%
	2	1,366	10%	6%	0%	0%
	3	1,558	13%	8%	0%	0%
	Total	4,945	9%	5%	0%	0%
Imagine Language and Literacy	K	1,099	38%	23%	3%	0%
	1	1,219	28%	19%	0%	0%
	2	1,293	21%	14%	0%	0%
	3	1,305	12%	8%	0%	0%
	Total	4,916	24%	15%	1%	0%
Waterford Reading	K	2,865	40%	23%	0%	0%
	1	823	28%	13%	0%	0%
	2	549	21%	6%	0%	0%
	3	23	0%	0%	0%	0%
	Total	4,260	35%	19%	0%	0%
Age of Learning	K	370	12%	4%	0%	0%
	1	350	21%	4%	0%	0%
	2	390	1%	0%	0%	0%
	3	375	0%	0%	0%	0%
	Total	1,485	8%	2%	0%	0%
IXL Language Arts	K	140	46%	41%	19%	7%
	1	153	86%	83%	55%	26%
	2	143	95%	90%	60%	30%
	3	156	82%	80%	53%	31%
	Total	592	78%	74%	47%	24%
Reading Horizons	K	38	0%	0%	0%	0%
	1	60	0%	0%	0%	0%
	2	41	0%	0%	0%	0%
	3	40	0%	0%	0%	0%
	Total	179	0%	0%	0%	0%
Read Naturally	K	-	-	-	-	-
	1	-	-	-	-	-
	2	41	0%	0%	0%	0%
	3	14	0%	0%	0%	0%
	Total	55	0%	0%	0%	0%
MobyMax	K	31	42%	0%	0%	0%
	1	36	17%	0%	0%	0%
	2	34	29%	0%	0%	0%
	3	35	29%	0%	0%	0%
	Total	136	29%	0%	0%	0%

Note: [a]*N* is the count of unique student users, including those who had zero usage for the year and those who appear in multiple vendor lists but excluding those who could not be matched to USBE records by SSID.
[b]Percentages are the number of students with usage at different percentages of vendor recommendations, divided

by the number of students in the *N* column. Data for *Renaissance* was not available at the time of this report and will be reported in an addendum when their data become available.

Table 6 reinforces the observation from Table 4 that there is large variation across vendors in student usage. Whereas Table 4 expressed this variation in minutes, Table 6 does so in the percentage of students whose usage met vendor recommendations. For example, 79% of *i-Ready* users, 74% of IXL users, and 47% of *Lexia* users met 100% of vendor usage recommendations, but none of the students using *Reading Horizons*, *Read Naturally*, or *MobyMax* did so. In some cases, differences among vendors could be partly explained by differences in their recommendations (see Table 3). *Reading Horizons* recommended 60 minutes per week and *Waterford Reading* recommended 75 minutes, while *i-Ready, Amira*, and *MobyMax* recommended only 30 minutes or fewer.

Despite vendor differences in usage recommendations, it is still disconcerting that some vendors are showing low rates of fidelity in implementation, even in relation to their own usage recommendations. For an educational intervention to have an impact, it must not only be effective when used but must also be implemented so that all students (or as many students as possible) use it. Because of these differences in recording student usage (units completed and variations in how time is recorded) and in vendor recommendations, caution should be exercised in any comparisons of efficacy among vendors.

4.3 RQ3: Student and School Differences in Implementation

- ***RQ3:*** *How does usage vary by student demographic characteristics, prior student achievement, or school demographic characteristics?*

Summary: Members of some student groups showed significantly lower levels of usage than non-members of those groups: Students receiving special education services, Native American students, economically disadvantaged students, Pacific Islander students, Hispanic students, and multilingual learners. Groups with significantly higher levels of usage than non-members were students with higher beginning-of-the-year Acadience Reading scores, Asian students, and students in schools with a higher percentage of the student body who were multilingual learners.

To explore how usage varied across student and school characteristics, we used a statistical model with usage (as a percentage of vendor recommendation) as the outcome and student demographics (grade, gender, race and ethnicity, multilingual learner status, receipt of special education services, eligibility for free or reduced-price lunch), prior student achievement (score on the Acadience Reading assessment measure from the beginning of the year), and school demographic characteristics (percent of students who are multilingual learners, percent of students receiving special education services) as predictors.[6] The analysis was limited only to students with non-zero software usage because of uncertainty about

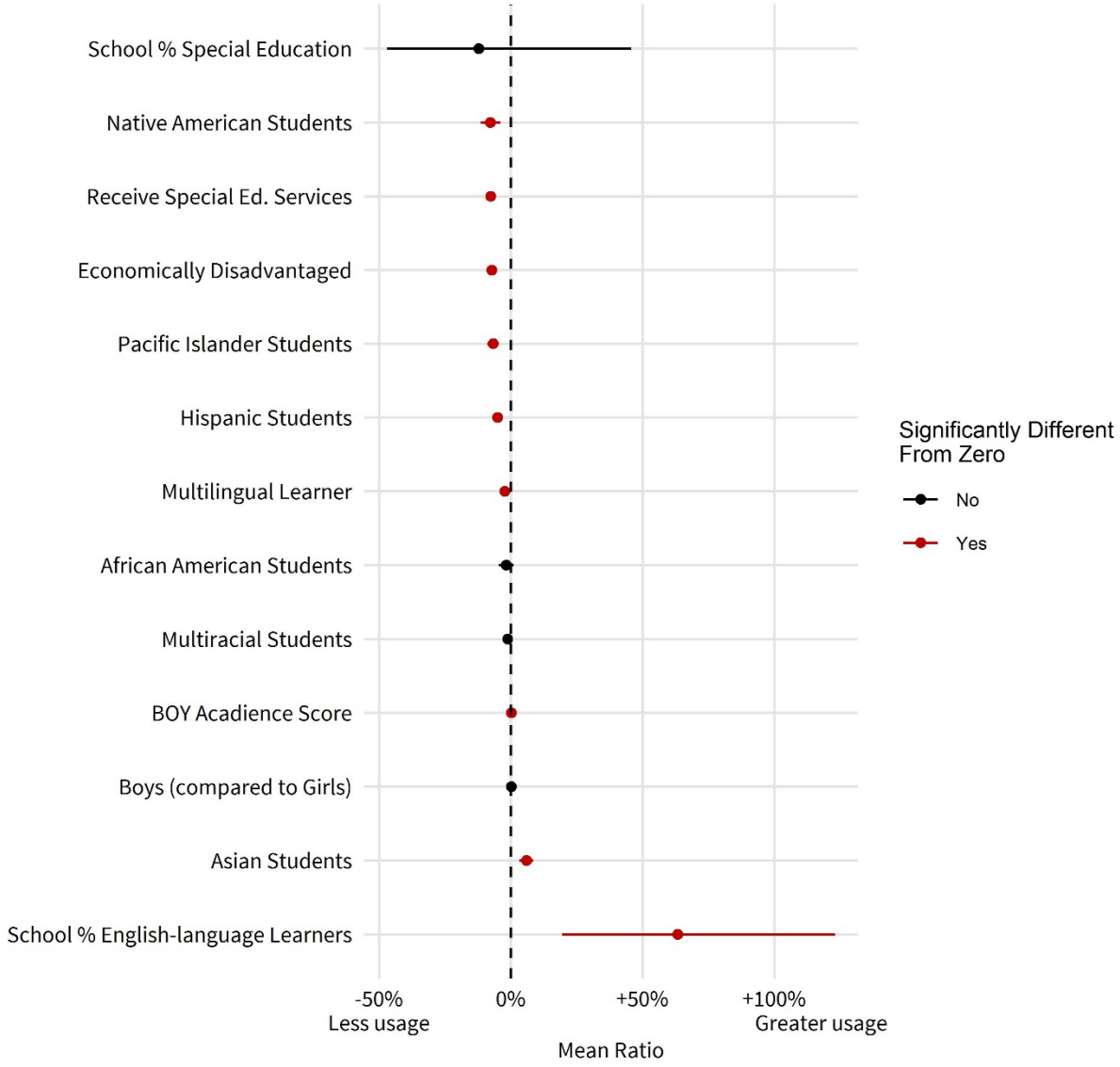
[6] The model was a multilevel gamma regression: Multilevel because of clustering within schools and gamma because the outcome variable is positive and right-skewed (a small percentage of students at 100%, many close to 0%). Although we initially also included the percent of students eligible for free or reduced-price lunch, we removed that variable because it showed multicollinearity with other school-level variables and produced estimated values with a very different pattern than the pattern obtained from raw means.

what zero usage represented (e.g., unsuccessful attempts at usage, administrative decisions to use a different vendor, or something else).

The relationship between each of the student and school characteristics and usage is presented in Figure 1. Figure 1 provides the regression coefficients from the analysis to show their relative strength and direction of association with usage. Because White students were in the majority, race and ethnicity was coded with White students as the reference group so that the relationship between race and ethnicity and usage is expressed as the contrast between White students and each of the other racial or ethnic groups.[7]

[7] This is a standard practice when one category represents the largest or baseline group and permits more stable estimates for comparisons among groups of students.

Figure 1. Student Predictors of Software Usage



Note. Grade level, vendor, and whether the beginning of the year Acadience Reading Score was missing were also included in the model, but were not included in this figure.

Figure 1 shows the predictor variables (e.g., Native American Students, Multilingual Learners) on the vertical axis and their relationship to early literacy software usage is indicated by the horizontal axis. The dashed vertical line represents no relationship between a predictor variable and software usage. When a point is to the left of the dashed line, the variable is negatively associated with usage. For example, on average, students receiving special education services use early literacy software 8% less than students who do not receive special education services. When a point is to the right of the dashed line, the variable is positively associated with participation. For example, on average, Asian students use early literacy software 6% more than White students.[8] The horizontal whiskers extending to the left and right of each

[8] Note that this is a 6% increase in an outcome variable that is itself a percentage (percentage of vendor recommendations met). If the average White student met 114% of vendor recommendations, a 6% increase would be 6.84 percentage points.

point are 95% confidence intervals and represent the degree of statistical precision in the estimate of the point. Wider whiskers represent more uncertainty in the value of the point. Dots that are red in Figure 1 represent predictor variables that are significantly associated with usage at $p < .01$, indicating that the relationship is stronger than would be expected by chance. Dots that are black in Figure 1 represent predictors whose relationship to participation is non-significant, indicating that their relationship to participation is uncertain and could be the result of chance. The predictor variables in Figure 1 are arranged from top to bottom based on their relationship to participation, with predictors having a negative relationship at the top and those having a positive relationship at the bottom.

Several student groups' early literacy software use was significantly lower than their comparison groups. For example, Pacific Islander, Hispanic, and Native American students used early literacy software less than White students. Students who were multilingual learners, economically disadvantaged, and who received special education services also showed significantly lower usage rates than their counterparts. Finally, beginning-of-year Acadience Reading score was positively correlated with early literacy software usage, indicating that students with worse beginning-of-year scores use early literacy software less than students who score higher on the beginning-of-year Acadience Reading assessment. While the differences are not large (all are less than 10% lower), they persist despite statistically controlling for beginning-of-year Acadience Reading scores and all other variables. Thus, they represent implementation gaps that are not easily explained by differences among groups in their past performance. These differences in implementation may further widen gaps between student groups over time.

From our analysis, the strongest predictor of higher early literacy software usage was being in a school with a higher percentage of multilingual learners. Compared to students in schools with a lower than average percent of multilingual learners, students in schools with a higher than average percent of multilingual learners tended to use the early literacy software 7 percentage points more.[9] This difference may be attributed to an intentional program of supplemental English language instruction using software at schools that have a high percentage of multilingual learners, although this evaluation doesn't assess how programs are being used instructionally.

[9] "Lower than average" and "higher than average" schools of multilingual learners were defined as half a standard deviation below and above the mean school percent multilingual learners, respectively.

5 | Impact Evaluation

This section addresses the impact evaluation. The following research questions (RQs) are answered:

> **RQ4**. Do students who are using early literacy software at higher levels (and also specifically at levels recommended by vendors) show stronger achievement outcomes as measured by scores on Acadience Reading assessments than students who do not use the software at levels recommended by vendors?
>
> **RQ5**. Does the impact of early literacy software usage vary by student demographic characteristics, prior student achievement, or school demographic characteristics?

5.1 Analytical Approach

In this section, we describe the "weighting" approach used for this evaluation, how it differs from the "matching" approach used in the previous evaluation (Evaluation and Training Institute, 2024), and the key reasons for the change. We also describe how usage and reading performance were measured.

5.1.1 Matching and Weighting to Draw Cause-and-Effect Conclusions

> **Summary:** Weighting achieves the same benefits of matching, such as being more confident about a *causal* relationship between early literacy software use and reading proficiency than a regression-with-controls approach, with the added benefit of estimating the likely gains in reading proficiency at *every* level of software use (not just 80% or 100% of recommended levels).

The central question of this evaluation is a *causal inference* question. That is, does participating in early literacy software programs *cause* students to perform better on reading assessments? The gold standard in causal inference is a randomized controlled experiment, where students are randomly assigned to either the control group (no software usage) or the treatment group (use early literacy software). By making assignment to a condition random, the risk of systematic pre-existing differences between the control group and treatment group is minimized.

When treatment is not randomly assigned, as is the case in the implementation of EISP, drawing causal inferences becomes more difficult because of the risk that the treatment and control groups are different in ways other than their level of software usage. If a difference in outcomes were found between the treatment group and the control group, it could be due either to the differences in treatment or to those pre-existing differences. For example, students identified as economically disadvantaged may differ from their economically advantaged peers in two ways: they may tend to use early literacy less often and their scores on standardized assessments may be lower. If we were to compare the outcomes of students who use early literacy software less and those who use it more, we would also be comparing students who are more versus less economically disadvantaged. In that case, we could not confidently say that any differences were due to the early literacy software because they could also be due to pre-existing differences in economic disadvantage. In this case, economic disadvantage would be a *confound* because

it prevents us from concluding that the software alone is *causing* increased performance among students.[10]

One way confounds can be addressed when random assignment is not possible is through *matching*. In matching, each student in the treatment group is paired with a student who did not receive the treatment but who is otherwise identical, or at least as similar as possible, on as many student characteristics as possible. For example, if a treatment student is male, has an average previous achievement test score, is a multilingual learner, etc., the matching process would try to find another student who is also male, has an average previous achievement test, and is also a multilingual learner who did not receive the treatment. By creating matched pairs and focusing comparisons within those pairs, matching minimizes the possible confounding effect of each of the matched variables. In the end, the key is that **matching attempts to minimize any systematic relationship between student characteristics and the treatment itself.**

One obstacle to matching is the case where the number of students who received the treatment far exceeds the number of students who did not receive the treatment. In this case, it is not possible to have one control student paired with each student in the treatment group.[11] The consequences of this imbalance between the number of students in the treatment and non-treatment groups are that 1) the quality of the match tends to be poor, and 2) because there will often be no reasonable match for some students in the treatment group, they may be unmatched, making the matched group unrepresentative of the treatment group and potentially biased.

Another complication for the matching approach is when treatment is not binary (i.e., received treatment vs. did not receive treatment), but is rather continuous, with different students receiving different "dosages" of the treatment, like in the present evaluation. In cases when the treatment can occur at multiple "dosages," the question for matching becomes: how large of a dose does the student have to receive before you count them as having "received" the treatment? In previous evaluations of this program, evaluators have used three different thresholds for treatment: (1) non-zero minutes of early literacy software use, (2) 80% or more of the vendor's usage recommendation, and (3) 100% or more of the vendor's usage recommendation. While this is a good approach, it has the disadvantage of not using all the information available. For example, while previous evaluations of this program have found that the "higher bar" analyses, where the threshold for treatment is stricter, demonstrate better outcomes compared to control students, this approach can't address certain questions. For example, how much usage is needed before students start seeing any positive effect at all? Or, is there a level of usage above which there are diminishing returns? These kinds of questions are simply not answerable using a matching approach unless evaluators create matched samples for every percent level of usage. Even then, the threshold approach conflates all "treatment" students together (e.g., when the threshold is set at "50% or more," a student who completed 50% of the usage recommendations and a student who completed 150% of the vendor's usage recommendation are both considered to have the same "level" of the treatment).

[10] The word "confound" is from the Latin "confundere" ("to pour together, mix up"), illustrating the mixing together of the treatment with another potential cause of the outcome.
[11] Unless you allow a control student to be matched with more than one treatment student, a procedure called matching with replacement.

An alternative to matching is *weighting*. In weighting, instead of finding a control student for every treatment student to minimize the relationship between student characteristics and the treatment itself, some students are given more "weight" in the analysis than others so that, when the whole weighted sample is analyzed, the relationship between the student characteristics and the treatment is minimized (Austin, 2011; Austin & Stuart, 2015). Weighting achieves the same goal of matching (i.e., both minimize the relationship between student characteristics and treatment). However, it does so in a different way, and it comes with multiple benefits. First, all treatment and control students can be included in the analysis. Having more observations increases statistical power: the probability that if there is a true relationship between the treatment and the outcome of interest, the model will detect it. Second, weighting allows us to use the treatment variable in a way that accounts for the continuous nature of "dosage," which allows for an estimation of the expected effects of usage at every level of usage. Finally, like matching, weighting has the advantage of supporting causal inferences about the relationship between the treatment and the outcome.[12]

Because there were so few non-users of EISP software (only 27% of Utah public school students in K-3 were not in the EISP vendor user lists) and because the treatment variable was continuous rather than binary, we opted to use weighting rather than matching to analyze the relationship between early literacy software usage and end-of-year Acadience Reading scores.[13] A detailed description of the analysis procedure is provided in Appendix A.

5.1.2 Measures

As indicated in the Section on *Percent Meeting Recommendations* and Table 6, we operationalized early literacy software usage as the percentage of vendor-recommended use (either minutes or units completed) that students completed. However, vendors have two recommendations for use: one for average weekly usage and the other for number of weeks. In the last evaluation (Evaluation and Training Institute, 2024), students were only considered to meet recommendations if they met both of those thresholds. To align our measure of student usage with the last evaluation, each student was assigned a percentage representing the lesser of (1) the percentage of vendor-recommended weekly usage and (2) the percentage of vendor-recommended number of weeks. For example, if a student's average use was 80% but they used the software for 100% of the number of recommended weeks, then both we and the previous evaluator would assign the student a value of 80%.

As our outcome measure, we used the end-of-year Acadience Reading composite score. As outlined by Good and colleagues (2011), the Acadience Reading assessment is designed to measure early literacy and reading ability for students in kindergarten through 6th grade. An example of one of the tasks on Acadience Reading is the "retell" task, which involves students reading passages and then demonstrating their comprehension by summarizing the passage to a teacher or teacher's designee, who scores their response based on whether particular words are mentioned. Scores on Acadience Reading increase with grade level, so average performance in 3rd grade will have a higher raw value than average performance in kindergarten.

[12] Like matching, this is contingent on the assumptions of weighting being met. The most important assumption is that the strongest confounding variables have been included, or are correlated with the confounds that have been included, in the model.

[13] Specifically, Covariate Balancing Propensity Score Weighting (CBPS). See Fong, Hazlett, and Imai (2018).

5.1.3 Analysis

For the weighting procedure and for our analysis, we controlled for the following student-level variables: eligibility for free or reduced-price lunch, student race and ethnicity, multilingual learner status, whether the student received special education services, beginning-of-the-year Acadience Reading Composite score, whether the student did not have a beginning of the year Acadience Reading Composite score[14], gender, and the student's early literacy software vendor.[15] We also included two school-level variables: the percent of the students at the school who are multilingual learners and the percent of students at the school receiving special education services. We then used these predictor variables in a weighted linear regression separately for each grade level with Acadience Reading score as the outcome variable. For more details, see Appendix A.

5.2 Impact Evaluation Results

5.2.1 Assessing Weighting Effectiveness

For details regarding how we assessed balance, see Appendix A. Overall, our weighting approach significantly reduced imbalance, bringing the average correlation between covariates and the treatment variable from .07 to .01. This means that, on average, our weighting approach was successful in reducing the correlation between baseline covariates and the treatment variable, achieving the same goal as matching by minimizing the relationship between student characteristics and software usage. This improvement increases confidence that any observed differences in outcomes are more likely due to differences in software usage rather than pre-existing differences among students.

5.2.2 RQ4: Impact

- ***RQ4****: Do students who are using early literacy software at higher levels (and also specifically at levels recommended by vendors) show stronger achievement outcomes as measured by scores on Acadience Reading Assessments?*

Summary: Compared to non-users, students who used early literacy software at 100% of vendor recommendations showed increases in reading that were "large" at kindergarten (*d* = 0.26), "moderate" at 1st and 3rd grades (*d* = .09 and .05, respectively), but not significantly different from zero at 2nd grade. Second grade users did see significant but small effects of usage (*d* = .04) at 200% of vendor recommendations.

The relationship between adherence to vendor recommendations and end-of-year Acadience reading composite score for each grade level is illustrated in Figure 2, which aggregates across all vendors.

[14] We did this because for students missing a beginning of the year Acadience reading score, we used median imputation (by grade level) to reduce the number of students lost from our sample due to missing data. We used this variable to indicate whether that score was median imputed. This occurred for less than 0.01% of all students, and thus, even when we just excluded these participants, the results were the same.

[15] If the student was not included on any vendor list, they were labelled as "no vendor" on this variable.

Figure 2. Predicted EOY Acadience Score by Grade Level and Percent of Vendor Recommendation Met

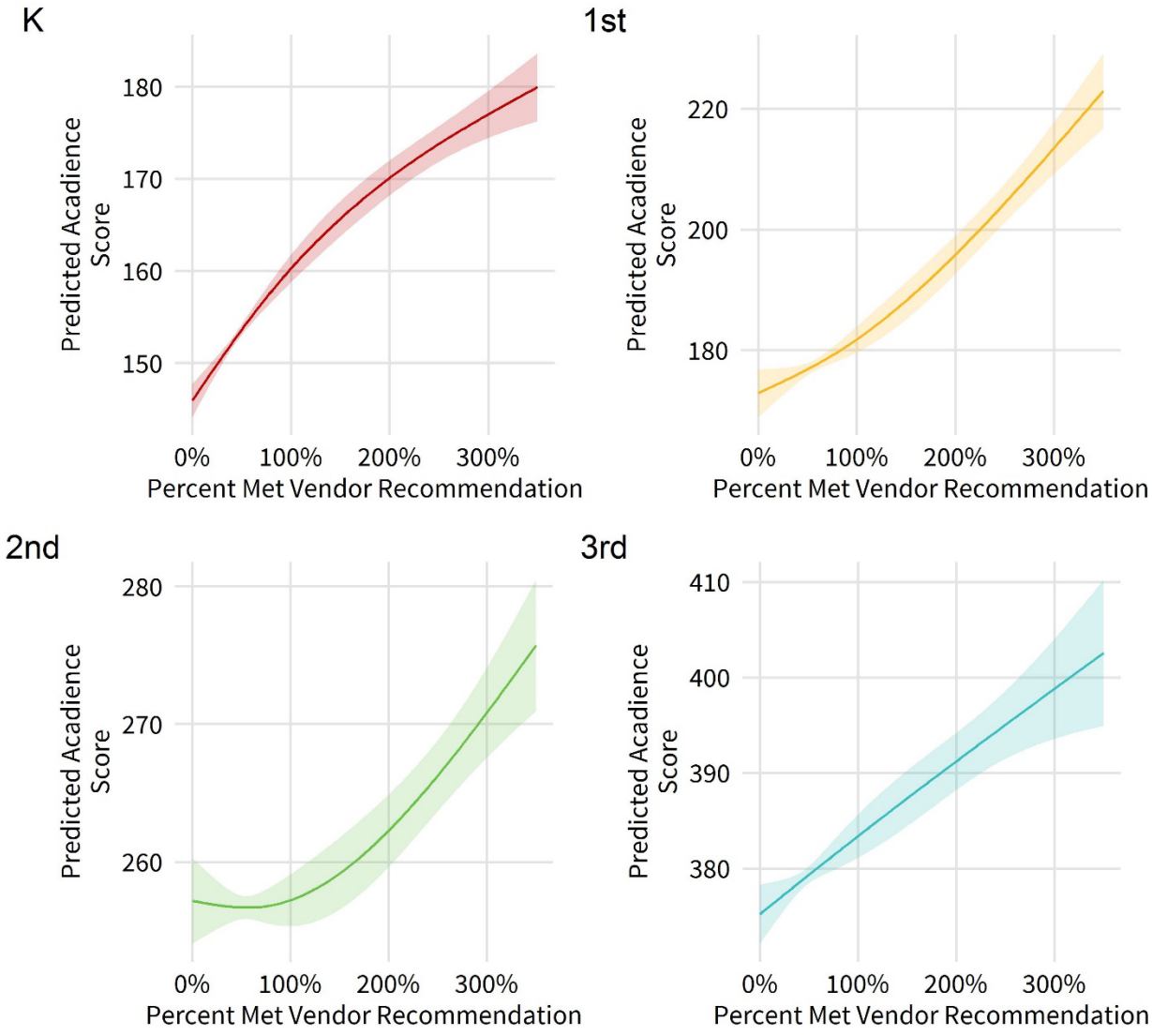


Figure 2 shows the estimated dose-response curve for each grade. In this case, "dose" is defined as the percentage of a vendor's software usage recommendation a student met. "Response," in this case, is the end-of-year Acadience Reading Composite score associated with each dosage level. These dose-response curves use weighting and regression to describe the relationship between software usage and reading scores in a way that minimizes the effects of confounding variables. The solid lines represent predicted mean values and the shaded areas are 95% confidence intervals around those predicted means. As mentioned above, Acadience Reading scores are designed to be higher at later grades.

For kindergarten, 1st grade, and 3rd grade, increases in software usage from 0% to 100% of recommendations were significantly related to increases in students' end-of-year Acadience Reading score ($p < .001$). However, for 2nd grade, increases in Acadience Reading scores do not occur until levels of student usage exceed 100% of vendors' recommendations. This pattern is consistent with the report by

the previous evaluator (Evaluation and Training Institute, 2024), which failed to find a significant effect of early literacy software use on students' scores among students in 2nd grade.

To illustrate the differences in Acadience Reading scores associated with different levels of usage, Table 7 presents the effect sizes (expressed in standard deviation units or Cohen's *d*) for contrasts of the predicted Acadience Reading scores when usage is at 0% of recommendations vs. at 80%, 100%, 200%, or 300% of recommendations.

Table 7. Effect Sizes for Contrasts in Usage

Contrast	Kindergarten	1st Grade	2nd Grade	3rd Grade
80% vs. 0%	0.21	0.07	< 0.01	0.04
100% vs. 0%	0.26	0.09	< 0.01	0.05
200% vs. 0%	0.44	0.23	0.04	0.11
300% vs. 0%	0.56	0.41	0.12	0.16

According to the effect sizes presented in Table 7, students in kindergarten who used the software at 100% of vendor recommendations performed 0.26 standard deviations higher than students who used the software at 0% of recommendations. This difference represents a "large" effect size in the context of educational interventions.[16] At all four grade levels, the expected gains in Acadience Reading scores increase as students meet and exceed 100% of vendor recommendations for usage. However, the magnitude of the benefits of greater use diminishes at higher grades and are strangely suppressed in 2nd grade, where gains are near zero through 100% of vendor recommendations, small (below 0.05) at 200% of recommendations, and only moderate (between .05 and .20) at 300% of recommendations. This discontinuity in efficacy at 2nd grade was also noted in the previous EISP evaluation report (Evaluation and Training Institute, 2024). By 3rd grade, however, the effect of using software at 100% of vendor recommendations has rebounded and is just over the threshold for a "moderate" sized effect (Kraft, 2020).

Although it is beyond the scope of this report, we did observe significant variability in effectiveness across vendors. Effectiveness at the vendor level will be provided to individual vendors in supplemental reports.

[16] According to Kraft (2020), an effect size above 0.20, if it were obtained in a randomized experiment, should be considered a "large" sized effect in the context of educational interventions.

5.2.3 RQ5: Student and School Differences in Impact

- ***RQ5:*** *Does the impact of early literacy software usage vary by student demographic characteristics, prior student achievement, or school demographic characteristics?*

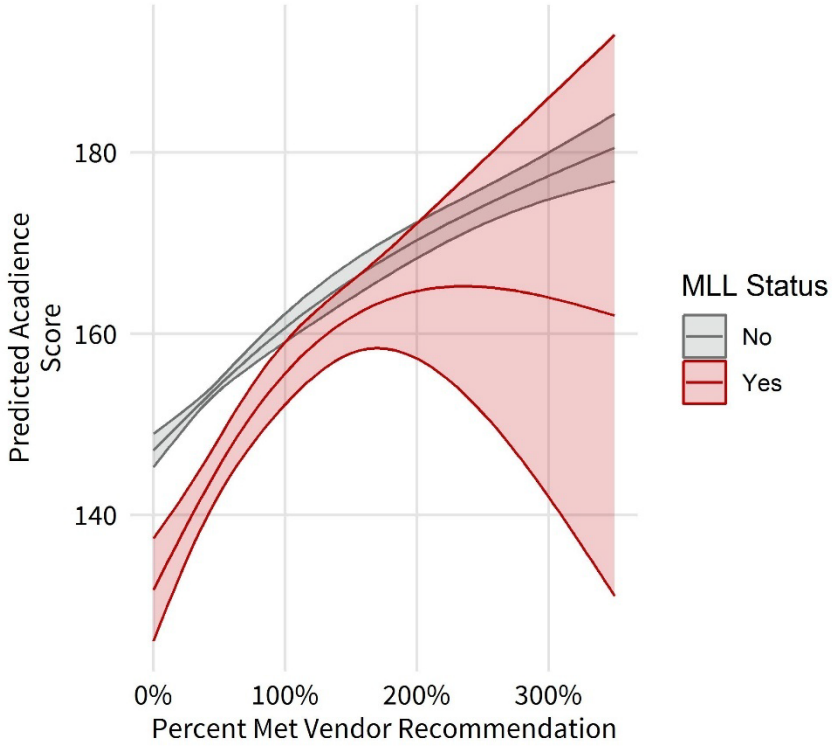
Summary: Early literacy software usage had a greater impact among multilingual learners, students receiving special education, and students with the lowest beginning-of-year scores on Acadience Reading. This pattern was not always significant at every grade level but was observed in kindergarten for multilingual learners; in kindergarten, 2nd grade, and 3rd grade for students receiving special education services; and in kindergarten, 1st grade, and 3rd grade for students with the lowest scores on the beginning of the year Acadience Reading assessment. This greater impact would result in shrinking gaps between student groups over time.

In this section, we examine whether the relationship between early literacy software usage and Acadience Reading scores (see Figure 2) differ by the following student characteristics: gender, race or ethnicity, free or reduced-price lunch status, multilingual learner status, special education status, and beginning of the year Acadience Reading composite score. We also examine whether the effect of early literacy software differs by two school-level characteristics: the percent of students receiving special education services and the percent of students who are multilingual learners. Testing whether the impact of early literacy software usage varies by other characteristics was performed by testing the statistical significance of the interaction between usage and each characteristic. Because this resulted in a large number of comparisons, we only focus on the comparisons that were statistically significant. Note that in this section, we are concerned with how *impact* (i.e., the strength of the relationship between usage and learning gains) varies across student and school characteristics. The relationship between student and school characteristics and *usage* was examined in Research Question 3 and Figure 1.

5.2.3.1 Early Literacy Software Reduces Gap in Acadience Scores Between Multilingual Learner and non-Multilingual Learner Students in Kindergarten

The relationship between software usage and Acadience Reading scores for kindergartners who were and were not multilingual learners is presented in Figure 3. As usage increased from 0% to 100%, the rate of gains in Acadience Reading was greater for multilingual learners than it was for non-multilingual learners. This is illustrated in Figure 3 by a steeper slope for multilingual learners (red line) than for non-multilingual learners (gray line) between 0% and 100% usage. The effect of this difference in slope for the two groups is that the gap between multilingual learners and non-multilingual learners that is observed at 0% usage has diminished by 100% usage. The dramatic widening of the confidence interval for multilingual learners as usage exceeds 200% is due to the small number of multilingual kindergartners who used the software at that level. Although the red line representing the mean for multilingual learners appears to curve downward as usage exceeds 200%, the wide confidence interval indicates that the true trend could be anywhere within the pink shaded area, making any interpretation of a downward trend highly uncertain. The interaction between multilingual learner status and usage was not statistically significant among students in other grade levels.

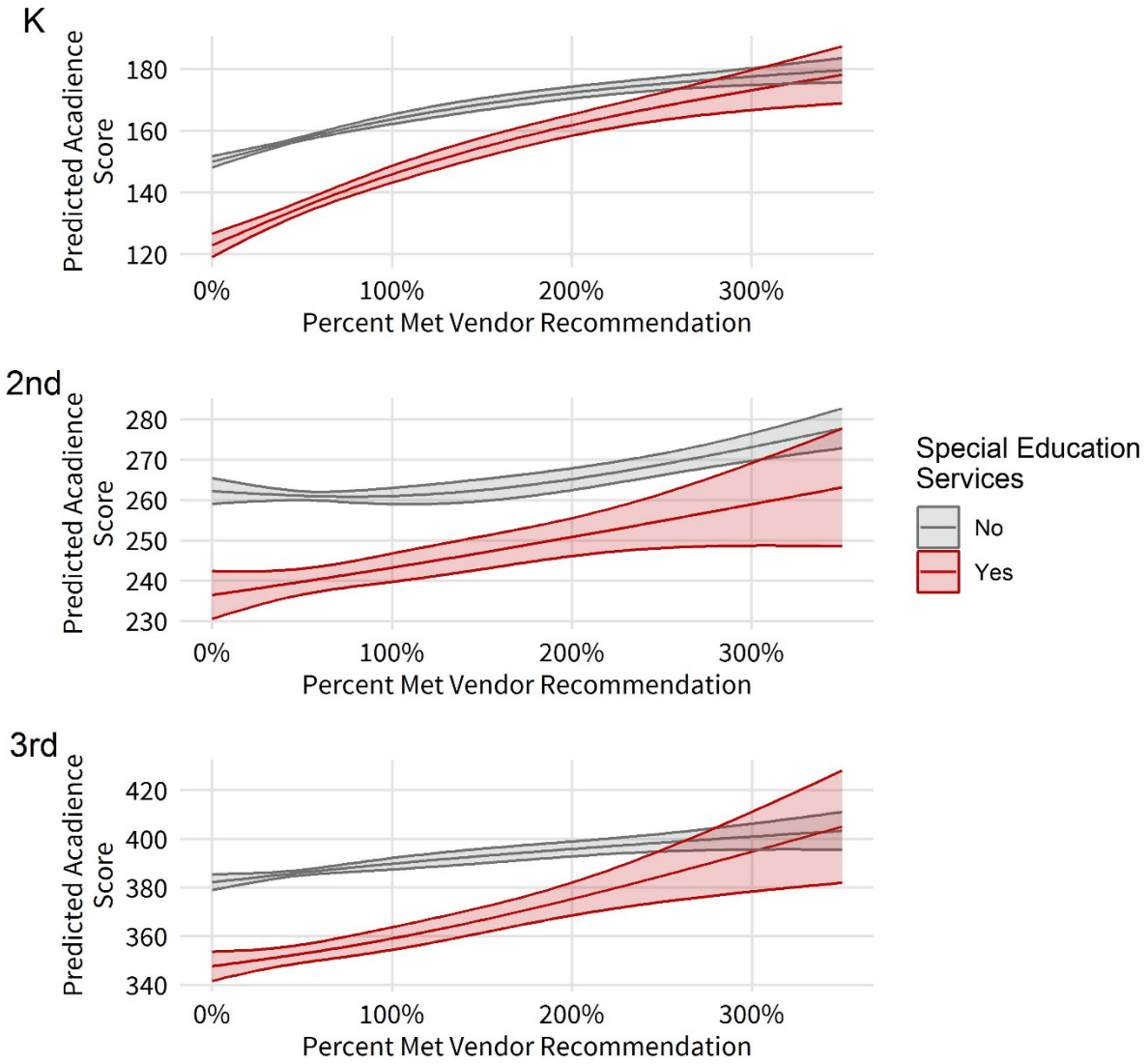
Figure 3. Usage Interacts with Multilingual Learner Status for Kindergartners



5.2.3.2 Early literacy software Reduces Gap Between Students Receiving Special Education Services and Students Who Do Not in Kindergarten, 2nd Grade, and 3rd Grade

Figure 4 shows that in every grade except 1st grade, increases in reading scores with usage were significantly stronger among students receiving special education services than among students who did not receive special education services. As it was in Figure 3, this can be seen in Figure 4 by the steeper slope for the red line (this time indicating students receiving special education services) than for the gray line (this time indicating students not receiving special education services). A lower starting point combined with a steeper positive slope for students receiving special education services results in a shrinking gap between students who did and did not receive special education services at higher levels of usage.

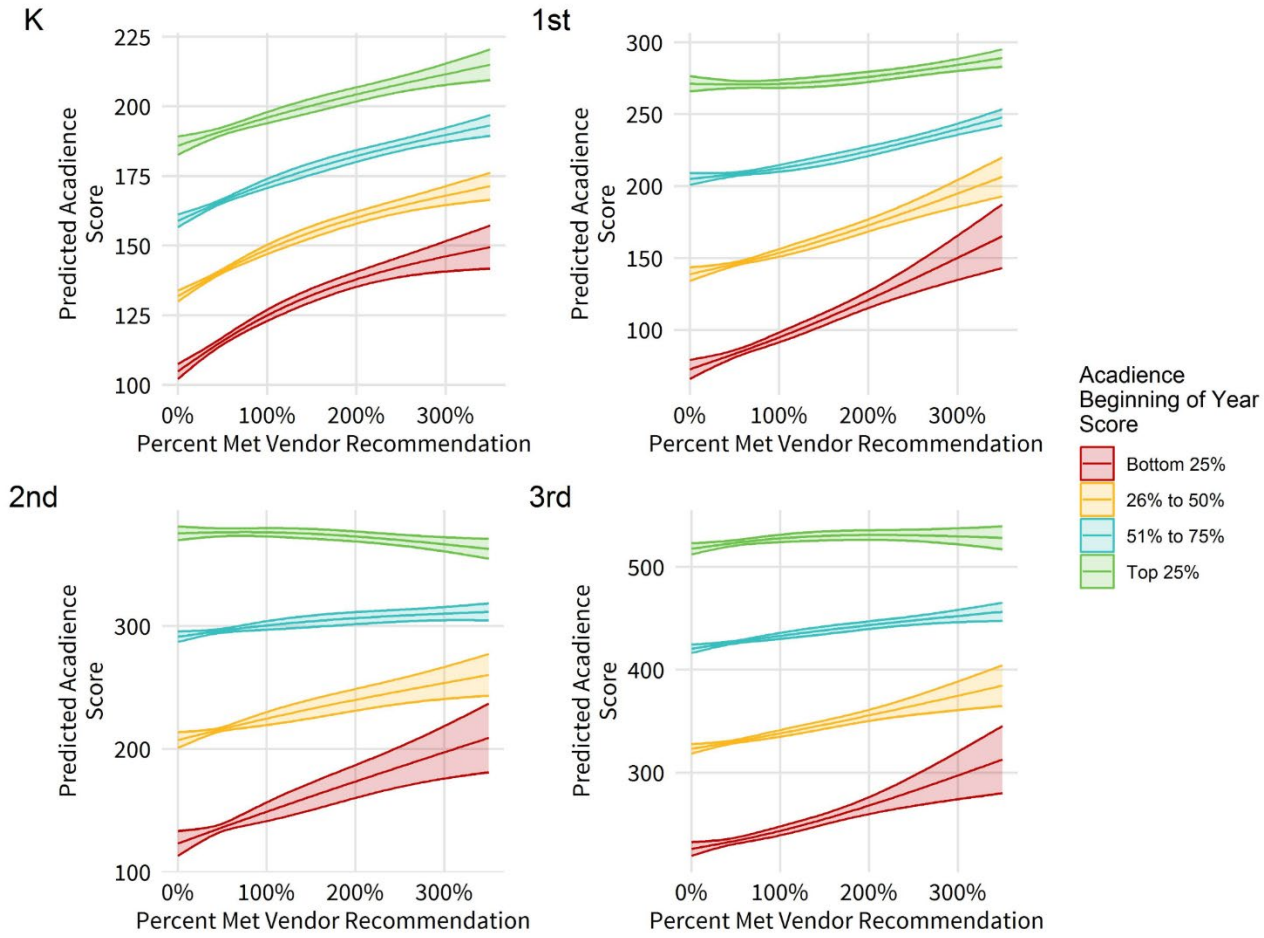
Figure 4. Usage Interacts with Receipt of Special Education Services



5.2.3.3 Across All Grades, Early Literacy Software Was Most Effective for Students Who Performed Worse on the Beginning of Year Acadience Reading Test

Across all four grades, the strongest increase in end-of-year Acadience Reading scores was observed among the students whose performance was in the lowest quartile (bottom 25%) at the beginning of the year. This can be seen in Figure 5 by the steeper slope in the line for students in the bottom 25% of beginning-of-year scores (red line) than in the lines for students at higher beginning-of-year quartiles.[17]

Figure 5. Usage Interacts with Beginning-of-Year Acadience Score



This result was statistically significant for every comparison: the bottom 25% saw greater gains with increased software usage than all three other groups, the 26% to 50% group saw greater gains than the two groups above, and so on. The cumulative effect of this pattern is that greater usage shrinks the gaps between groups of students defined by different levels of reading performance at the beginning of the year.

[17] It should be noted that when students were missing beginning of the year scores, their scores were imputed with the median within their grade level. However, this occurred for less than 0.01% of the students in the data set, and thus did not have a significant effect on the findings.

6 | Conclusions and Recommendations

Summary: The Early Interactive Software Program (EISP) shows robust evidence of effectiveness as an early literacy intervention. Drawing on prior research and findings from this evaluation, the results indicate that consistent use of early literacy software is associated with meaningful gains in reading proficiency, particularly for multilingual learners, students receiving special education services, and those who began the year with lower reading scores. Based on our models and analysis, Utah's statewide share of third graders reading on grade level in 2024–25 would have been approximately four percentage points lower if the software had not been used. These findings position EISP as a key strategy among the interventions identified in Utah SB 127 to support the state's goal of having at least 70 percent of third graders reading on grade level by July 2027.

Utah has the ambitious goal of ensuring that 70% or more of 3rd grade students are reading on grade level by July 2027 (Utah SB 127, 2022). With less than two years remaining, and fewer than 50% of 3rd graders reading on grade level in 2024, there is an urgent need for interventions that are both scalable and demonstrably effective.[18] The present evaluation reveals that EISP could play a key role in supporting early literacy success for students and helping the state move toward its 70% goal.

Senate Bill 127 (2022) authorized several interventions designed to improve early literacy, including digital learning literacy platforms such as those supported by EISP, literacy coaching for teachers in schools with low scores on reading assessments, professional learning on the science of reading for teachers, and changes to educator preparation programs and licensing requirements. Among these interventions, EISP stands out as having strong and robust evidence of effectiveness in raising students' levels of reading competency. The present study used rigorous quasi-experimental methods to measure and remove the influence of variables that might differ between users and non-users and provide relatively unbiased estimates of the relationship between level of software use ("dosage) and gains on Acadience Reading assessments ("response"). Moreover, the large sample (i.e., over 149,000 users) ensures that the findings are statistically robust and representative of broad implementation.

Below, we draw conclusions based on the present evaluation, make recommendations for implementation of EISP, note limitations of the present evaluation, and suggest some future directions for research.

6.1 Conclusions

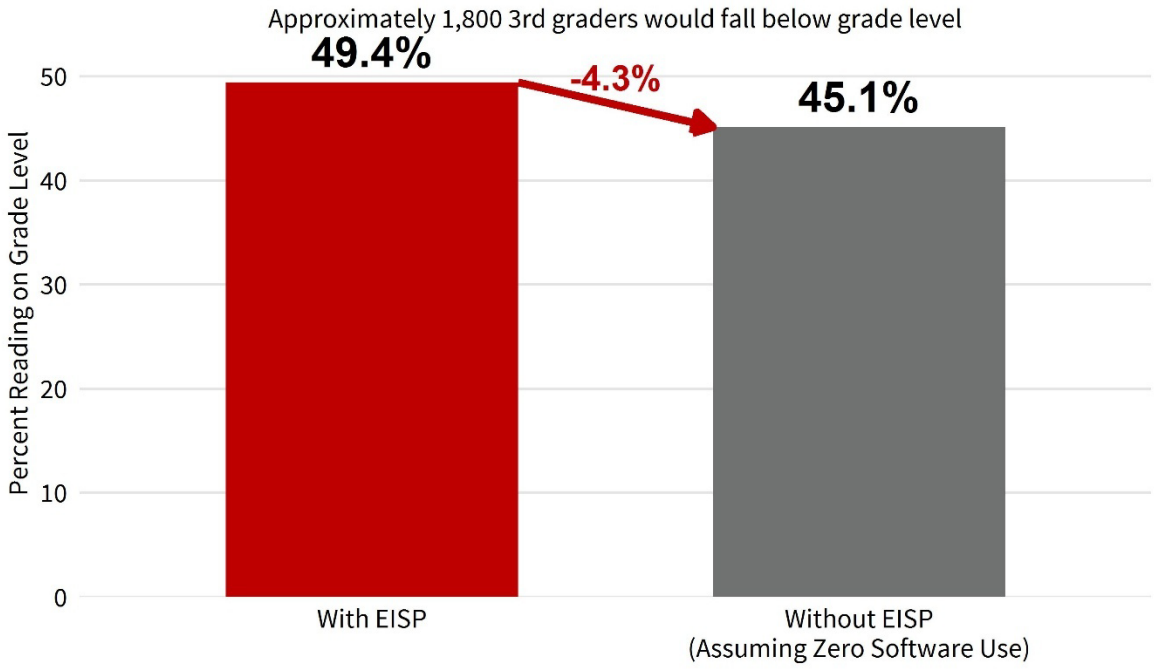
- **EISP shows strong evidence of effectiveness in improving reading outcomes for kindergarten, 1st grade, and 3rd grade students.** EISP appears to be especially effective in kindergarten, where the effect size of 0.26 is considered "large" in the context of educational interventions (Kraft, 2020) and is similar in size to the effect of intensive individualized tutoring (Cook et al., 2015). Effectiveness in kindergarten is notable given that reading on grade level at

[18] Defined as having a composite score that places the student in the "above benchmark" category. https://reportcard.schools.utah.gov/State/OverallPerformance?StateID=99&SchoolID=&DistrictID=&SchoolNbr=&SchoolLevel=K8&IsSplitSchool=0&schoolyearendyear=2024

the end of the year in kindergarten is indicative of a 66% chance of reading on grade level at the end of the year in 3rd grade.[19]

- **The impact of EISP usage was stronger among multilingual learners, students receiving special education services, and students with the lowest scores on beginning-of-the-year measures of reading proficiency**. These groups have historically experienced fewer opportunities and gains over time, meaning that consistent use of early literacy software programs has the potential to improve literacy for all students and reduce longstanding achievement gaps in reading proficiency.
- **EISP meaningfully increased 3rd grade reading proficiency statewide.** As shown in Figure 6, our 3rd grade model estimates that if EISP were discontinued and early literacy software usage fell from its current level to zero (assuming schools did not purchase the software using their own budgets), the percent of 3rd graders who were reading on grade level at the end of the year in school year 2024-2025 would have been 45.1% instead of the actual percentage of about 49.4%.[20] This translates to an estimated 1,883 fewer 3rd graders statewide achieving grade level reading proficiency in the absence of EISP software.[21]

Figure 6. Projected Impact of Eliminating EISP Software Usage



[19] We calculated this estimate for the 2021-2022 school year kindergarten cohort (who were in 3rd grade in 2024-2025). It reflects the positive predictive value (i.e., the percent of those who were reading on grade level in kindergarten who went on to be reading on grade level in 3rd grade).

[20] This 49.4% value is not the official rate and is based on our analysis of the data. Official estimates have yet to be released by USBE and may differ from this estimate based on differences in data cleaning and data quality procedures.

[21] These estimates were generated by using our 3rd grade model and predicting the estimated composite score values if all students in the data set had 0% "met vendor recommendation" usage. Then, we took these estimated values and categorized them as "above benchmark" if the estimate was at or above 405. This estimate assumes students would not have access to any literacy software if EISP did not exist.

6.2 Recommendations for Implementation

- **Raise levels of early literacy software use.** Figure 2 shows that higher levels of early literacy software use are associated with higher scores on the Acadience Reading assessment even after controlling for student and school characteristics (including beginning-of-year Acadience Reading scores). However, Table 6 shows that many students are using the software at levels below vendor recommendations. Even effective interventions cannot improve student outcomes when they are not used as intended.
- **Focus on improving EISP participation among lower-performing students to reduce reading gaps**. The impact of EISP usage was strongest among multilingual learners, students receiving special education services, and students with lower beginning-of-year reading proficiency scores. These groups have historically shown smaller learning gains over time, indicating that the program has the potential to support the achievement of all students and help close achievement gaps. In current practice, usage of EISP was lower in all three groups. This suggests that without targeted efforts to ensure similar opportunities for engagement across groups, the program's benefits may not reach those who could gain the most. In fact, limited access and usage could risk widening existing gaps. Importantly, the UEPC has observed similar patterns for users of math learning software in Utah, where students who stand to benefit most often engage at lower rates (Altermatt, Altermatt, Rorrer, & Moore, 2022; Altermatt, Yildiz, & Rorrer, 2025). Ensuring that historically under-performing students use early literacy software at recommended levels is therefore critical to realizing their potential to reach ambitious goals for reading proficiency by 2027.
- **Further Integrate EISP within Utah's broader strategy and instructional practices for improving early literacy**. Coordinating EISP implementation with literacy coaching, ongoing professional learning, and other evidence-based interventions can help ensure that classroom instruction and digital learning reinforce one another. When teachers are supported in using EISP effectively, students are more likely to engage meaningfully with the program and make measurable reading gains. This integrated approach can strengthen the overall impact of Utah's early literacy initiatives and support more students in reaching grade-level proficiency.

6.3 Limitations

Although this study involved a large number of students and employed methods designed to maximize confidence in cause-and-effect conclusions, there were several limitations that should be acknowledged. First, the methods we employed to maximize confidence in cause-and-effect conclusions (i.e., covariate balancing propensity score weighting and multilevel regression with covariates) reduce the risk that any of the measured covariates (e.g., gender, race or ethnicity, beginning-of-year reading scores) are the true cause of the observed relationship between software usage and reading scores. However, those methods do not reduce that risk to zero, nor do they rule out the possibility that unmeasured differences among students (e.g., self-confidence, curiosity) might explain the relationship. Second, our analysis relied on a measure of student software use (the degree to which a student's use met the vendor's recommendations) that had a different interpretation for different vendors. For example, some vendors recommended only 30 minutes per week while others recommended 60. Furthermore, some vendors'

recommendations did not involve time but rather the completion of questions or stories. Finally, as noted in Appendix B, time may be measured differently between vendors. These differences across vendors could introduce error (unaccountable variation), resulting in effect size estimates that are smaller than their true value. The differences could also result in estimates of the dose-response relationship that do not generalize across all vendors. The latter problem is more likely to occur for vendors with fewer students and with recommended and typical usage that depart from the mean.

6.4 Future Directions for Research and Evaluation

To optimize the reach and effectiveness of EISP across Utah schools and support the state in meeting its 70% literacy goal, the UEPC suggests that research on EISP be extended:

- **Investigate outcomes for 2nd graders.** The UEPC found that early literacy software was not as effective in 2nd grade as it was in other grade levels, with positive outcomes for 2nd graders occurring only at usage levels that were well above vendor recommendations. This finding is consistent with the results from last year's evaluation (Evaluation and Training Institute, 2024). One potential explanation for this discrepancy is the measure itself. As noted in the previous evaluation, the Acadience Reading measure in 2nd grade places a greater emphasis on accuracy than it does on other literacy measures. However, the hypothesis that the different outcome for 2nd graders is due to differences in the measure of literacy is not supported by examination of the pattern of correlation across grade levels. If the 2nd grade measure was capturing something fundamentally different from the measures at other grade levels, then a student's score in 2nd grade would be expected to show a lower correlation with their scores at other grade levels than the same student's scores in kindergarten, 1st, and 3rd grades. Examination of the correlations indicates no difference for 2nd grade, indicating that measurement alone is unlikely to be the explanation for the diminished impact observed for 2nd graders. Further research is needed to investigate why EISP does not appear to be as effective in 2nd grade.
- **Identify barriers to implementation.** Further research is needed to investigate the reasons why recommended usage was so often out of reach for students. Some hypotheses include bottlenecks in accessing the necessary computing resources and teacher struggles with integrating regular usage into their curriculum. These questions cannot be answered with the data currently available but could be explored through surveys or interviews with the teachers who are using the products.
- **Identify factors that maximize the impact of EISP.** Building on UEPC's prior research linking teacher implementation practices to changes in student attitudes and achievement (e.g., Altermatt & Rorrer, 2024a, 2024b; Altermatt, Rorrer, Altermatt, Doane, & Timmer, 2022), future research could explore how learning gains can be further increased. Key questions include: Which teacher practices increase efficiency of learning gains? How do classroom conditions, such as class size, affect outcomes? Are community-level factors, such as housing costs or average wages, influencing participation? Addressing these questions could provide actionable strategies to increase program impact, especially for students who stand to benefit the most.
- **Examine program instructional integration and sustainability.** Future research could investigate how EISP interacts with other literacy supports, including classroom instruction, literacy coaching, ongoing professional development, and other evidence-based interventions, including high-dosage tutoring. Studies could explore which combinations of supports maximize student engagement with the software and accelerate reading gains, as well as how teacher practices influence the effectiveness of integrated approaches. UEPC research could also

examine how instructional integration affects the sustainability of student learning outcomes over time, including whether gains from EISP are reinforced across classrooms, grades, and school years. Understanding these dynamics would provide actionable guidance for districts and schools on how to align digital learning tools with broader literacy initiatives to strengthen overall impact and help more students reach grade-level proficiency.

7 | References

Abrami, P., Borohkovski, E., & Lysenko, L. (2015). The effects of ABRACADABRA on reading outcomes: A meta-analysis of applied field research. *Journal of Interactive Learning Research, 26*(4), 337-367.

Altermatt, T. W., Altermatt, E. R., Rorrer, A. K., & Moore, B. (2022). *Math personalized learning software: Examining usage and associations with achievement in Utah during the Covid-19 pandemic.* Salt Lake City, UT: Utah Education Policy Center.

Altermatt, E. R., & Rorrer, A. K. (2024a). Teacher goal setting for personalized learning software: Associations with perceptions of the value of software and growth mindsets. *Journal of Research on Technology in Education,* 1-16. https://doi.org/10.1080/15391523.2024.2411697

Altermatt, E. R. & Rorrer, A. K. (2024b). *Promising Practices for Creating Strong Technology-Enabled Learning Environments: Associations Between Teachers' Math Learning Software Implementation Practices and Students' Mathematics Achievement.* Salt Lake City, UT: Utah Education Policy Center.

Altermatt, E. R., Rorrer, A. K., Altermatt, T. W., Doane, M., & Timmer, M. (2023a). *Blended Learning Research Brief No. 1. Associations Between Math Personalized Learning Software Use and Personalized, Competency-Based Instructional Strategies.* Salt Lake City, UT: Utah Education Policy Center.

Altermatt, E. R., Rorrer, A. K., Altermatt, T. W., Doane, M., & Timmer, M. (2023b). *Blended Learning Research Brief No. 2. Associations Between Student Attitudes Toward Math and Math Personalized Learning Software Use.* Salt Lake City, UT: Utah Education Policy Center.

Altermatt, E. R., Rorrer, A. K., Altermatt, T. W., Doane, M., & Timmer, M. (2023c). *Blended Learning Research Brief No. 3. Associations Between Teachers' Goals for Students' Math Software Use and Teachers' Perceptions of the Value of Software.* Salt Lake City, UT: Utah Education Policy Center.

Altermatt, E. R., Yildiz, M., & Rorrer, A. K. (2025). *STEM Action Center's K-12 Math Personalized Learning Software Grant Program: 2024-2025 Evaluation Report.* Salt Lake City, UT: Utah Education Policy Center.

Annie E. Casey Foundation. (2010). Early Warning: Why Reading by the End of Third Grade Matters. KIDS COUNT Special Report. Baltimore, MD. Retrieved from https://assets.aecf.org/m/resourcedoc/AECF-Early_Warning_Full_Report-2010.pdf

Archer, K., Savage, R., Sanghera-Sidhu, S., Wood, E., Gottardo, A., & Chen, V. (2014). Examining the effectiveness of technology use in classrooms: A tertiary meta-analysis. *Computers & Education, 78,* 140–149.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research,46* (3), 399–424. https://doi.org/10.1080/00273171.2011.568786.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34* (28). 3661–79. https://doi.org/10.1002/sim.6607.

Bernacki, M.L., Greene, M.J. & Lobczowski, N.G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose(s)? *Educational Psychology Review, 33,* 1675–1715.

Brizard, J.-C., (2023, April). Breaking with the past: Embracing digital transformation in education. *Digital Promise.*

Carbonari, M. V., DeArmond, M., Dewey, D., Dizon-Ross, E., Goldhaber, D., Kane, T. J., McDonald, A., McEachin, A., Morton, E., Muroga, A., Salazar, A., & Staiger, D. O. (2024). *Impacts of academic recovery interventions on student achievement in 2022-23*. Research Report. Cambridge, MA: Center for Education Policy Research, Harvard University/CALDER Working Paper No. 303-0724.

Dahl-Leonard, K., Hall, C. & Peacott, D. (2024). A meta-analysis of technology-delivered literacy instruction for elementary students. *Educational Technology Research and Development,* 72, 1507–1538.

Evaluation and Training Institute (ETI). (2024). *Utah's early intervention reading software program: 2023-2024 program evaluation findings.* Los Angeles, CA.

Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Annual Review of Applied Statistics, 12*(1), 156-177.

Good, R. H., III, Kaminski, R. A., Cummings, K. D., Dufour-Martel, C., Petersen, K., Powell-Smith, K. A., Stollar, S., & Wallin, J. (2011). *Acadience Reading K-6 Assessment Manual*.

Harlaar, N., Dale, P. S., & Plomin, R. (2007). From learning to read to reading to learn: Substantial and stable genetic influence. *Child Development*, 78(1), 116–131.

Huebner, T. A., & Burstein, R. (2023). *Strategies for encouraging effective technology-enabled instructional practices in K–12 education: A thought piece drawing on research and practice.* WestEd.

Jamshidifarsani, H., Garbaya, S., Lim, T, Blazevic, P. & Ritchie, J. M. (2019). Technology-based reading intervention programs for elementary grades: An analytical review. *Computers & Education, 128,* 427-451,

Kuhfeld, M., & Lewis, K. (2022). *Student Achievement in 2021-2022: Cause for Hope and Continued Urgency.* NWEA Research Brief, July 2022.

Kuhfeld, M. & Lewis, K. (2024). *Recovery Still Elusive: 2023-24 Student Achievement Highlights Persistent Achievement Gaps and a Long Road Ahead.* NWEA Research Brief.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y., and Dunleavy, E. (2007). Literacy in Everyday Life: Results From the 2003 National Assessment of

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49* (4), 241-253.

Lesnick, J., Goerge, R.M., & Smithgall, C. (2010). Reading on grade level in third grade: How is it related to high school performance and college enrollment? Chicago, IL: Chapin Hall at the University of Chicago.

McTigue, E., Solheim, O., Zimmer, W., & Uppstad, P. (2019). Critically reviewing GraphoGame across the world: Recommendations and cautions for research and implementation of computer-assisted instruction for word reading acquisition. *Reading Research Quarterly, 55*(1), 45-73.

National Center for Education Statistics. (2024). *The Nation's Report Card: Reading—National Achievement—Grade 4.* U.S. Department of Education, Institute of Education Sciences. https://www.nationsreportcard.gov/reading/nation/achievement/?grade=4

Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2017). Informing Progress: Insights on Personalized Learning Implementation and Effects. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2042.html

Reich, J. & Ito, M. (2017). *From Good Intentions to Real Outcomes: Equity by Design in Learning Technologies.* Irvine, CA: Digital Media and Learning Research Hub.

Utah State Board of Education (2024). *Strategic Plan. 2024 Implementation Update.* Retrieved from https://schools.utah.gov/board/utah/_strategies_/USBEStrategicPlan2024ImplementationUpdate.pdf

Utah State Board of Education and the National Center for the Improvement of Educational Assessment, Inc. [USBE & NCIEA] (2021). *Exploring the effects of the Covid-19 pandemic on student achievement in Utah.* https://le.utah.gov/interim/2021/pdf/00003999.pdf.

Van Schoors, R., Elen, J., Raes, A., Vanbecelaere, S., & Depaepe, F. (2023). The charm or chasm of digital personalized learning in education: Teachers' reported use, perceptions and expectations. *TechTrends, 67*(2), 315–330.

Verhoeven, L., Voeten, M., & Segers, E. (2022). Computer-assisted word reading intervention effects throughout the primary grades: A meta-analysis 2022. *Educational Research Review, 37,* 100486,

Zheng, L., Long, M., Zhong, L., Gyasi, J. F. (2022). The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: A meta-analysis. *Education and Information Technologies, 27*(8), 11807-1183.

Appendix A: Technical Details of Analysis

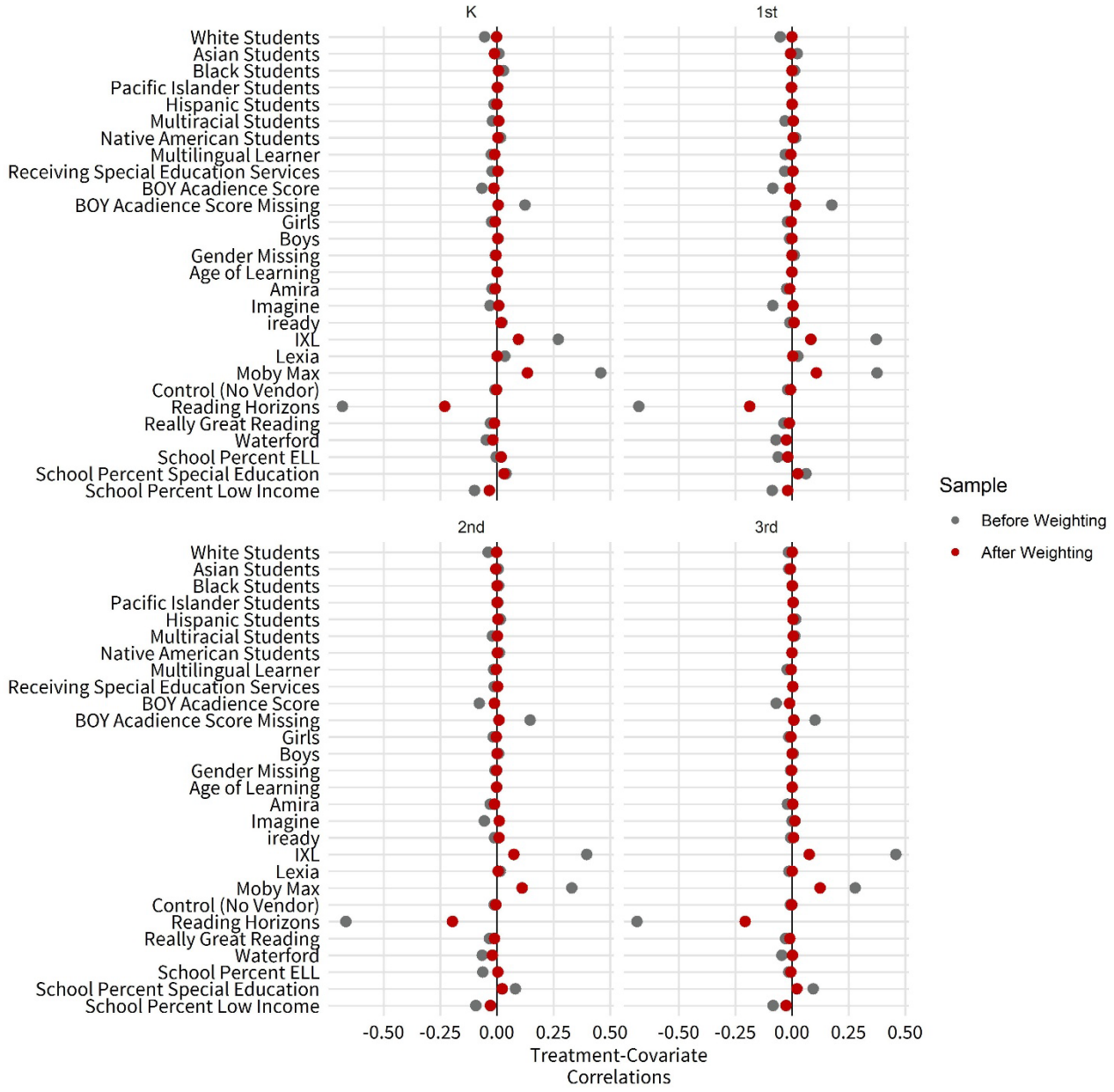
The goal of weighting is to minimize the correlation between the level of treatment received by a student (i.e., their level of software usage) and student and school characteristics. This is the same goal pursued by matching and by random assignment to conditions, the gold standard of research designs for causal inference. We used the *WeightIt* R package to conduct Covariate Balancing Propensity Score weighting to generate weights that reduced the correlation between the treatment variable (i.e., percent met vendor recommendation), and the following variables: free/reduced-price lunch status, student race, multilingual learner status, receipt of special education services status, beginning of the year Acadience Reading composite score, whether the beginning of the year Acadience Reading composite score was missing for that student, student gender, which vendor of early literacy software the student used (or if they didn't use any vendor, they were labelled as "No vendor"), school level percent multilingual learners, and school level percent of students receiving special education services. Weighting and analysis were done separately by grade level. Note, for beginning of the year Acadience Reading composite score, we used median imputation within grade level to eliminate missingness.

After estimating the weights and assessing covariate-treatment correlations, we used a propensity score-weighted linear regression with natural splines ($df = 2$) to estimate the causal relationship between *percent met vendor recommendation* and end-of-year Acadience Reading composite score. To estimate the average dose-response function, we used *g*-computation. Specifically, we predicted 100 evenly-spaced outcomes across the complete range of values of *percent met vendor recommendation*. We estimated whether the slope was statistically significant. We then computed the average marginal effect and tested whether it was significantly different from zero.

For moderation analyses, weights were computed using the same variables that were used in the overall analysis. For binary moderators, weights were computed within each level of the moderator, and covariate-treatment correlations were assessed across all other covariates within levels of the moderator. For continuous moderators (i.e., beginning of the year composite score, school level percent multilingual learners, and school level percent receiving special education services), we split individuals' values on these variables into quantiles, and used this quantile version of the variable to assess moderation. We did this because it was not feasible to estimate or assess balance across all covariates across all values of a continuous moderator.

Figure 7 below shows the correlation between each student or school characteristic and the treatment (i.e., the percent the student adhered to vendor recommendations) before weighting and after weighting at each grade level. The closer the dot is to the dark vertical 0-correlation line in the center of each plot, the smaller the relationship between that variable and the level of treatment. As you can see, after weighting, some variables demonstrate no relationship at all with the treatment variable, whereas others still demonstrate a relationship (albeit diminished) with treatment.

Figure 7. Weighting Balance Plot



Appendix B: The Problem of Variation in Recommended Usage

Summary: Differences in how vendors measure student use of the learning platform can cause some products to appear to be more efficient at causing student learning gains than other vendors when there is no true difference. As a result, caution should be exercised in any comparisons between vendors.

Three factors contribute to challenges in making comparisons across vendors: differences in usage recommendations, differences in whether time is reported, and differences in how time is reported.

First, vendors vary in their recommendations for student software usage (see Table 3). If two vendors have products that are equally effective and one vendor has recommendations that require more time, then that vendor will appear to be more effective if only those students who meet minimum recommendations are considered. By raising the bar for recommendations high enough, a vendor could increase the likelihood of finding a relationship between meeting recommendations and showing learning gains. This potential for bias is a "selection effect": a situation where a treatment group performs differently not because of the treatment itself but rather because the treatment has selected (filtered out) a distinctive group of participants from the sample. In this case, setting recommendations to a high level would likely select for students with higher average past performance (because those students would be more likely to use the product and to have the skills necessary to meet the recommendations) and this group would be more likely to receive high scores on end of year assessments of learning. As a result, meeting vendor usage recommendations would be positively correlated with learning gains because it is a mechanism for selecting students with higher average past performance and not because the product is effective. One method for examining the potential bias introduced from different usage recommendations is to inspect the rate of meeting recommendations for each vendor. If a vendor is setting too high a bar, then the percentage of students meeting recommendations will be low.

Second, two vendors (*Read Naturally* and *IXL*) do not report the time that students used the software. Instead, these two vendors report the number of units (e.g., stories, questions) that students completed. One complication for the assessment of programs that report units rather than minutes is that unit completion is likely to require some degree of mastery to accomplish. As a result, it is likely that students who complete more units will show more growth than students who do not. As noted above, the level of bias introduced by this possibility could be examined by inspecting the percentage of students who meet recommendations for unit completion. When the percentage of students who meet recommendations for unit completion is high, there is less risk for bias due to selection effects.

Third, vendors do not all record student usage time in the same way. Some vendors may start timing usage when a student logs in and stop timing when the student logs out. Other vendors may stop the timer when a student is inactive for more than a few seconds, restarting it when they re-engage. These two approaches would produce different time logs for an identical student experience. Two vendors with identical products but different timing systems would show different relationships between time and

learning gains. Specifically, the vendor that stopped time during inactivity would appear to be more "efficient" because the same learning gains would occur over a smaller period of [recorded] time.

Because of these differences, caution should be exercised in any comparisons of efficacy among vendors.

Project Staff

The following Utah Education Policy Center (UEPC) team members contributed to this project.

T. William Altermatt, Ph.D., Lead Data Scientist

James Gallyer, Ph.D., Data Scientist

Muhammed Yildiz, Ph. D., Data Scientist

Ellen Altermatt, Ph.D., Assistant Director for Research and Evaluation

Andrea Rorrer, Ph.D., Director