CEMETR-2025-02
DECEMBER 2025

CEME Technical Report

The Center for Educational Measurement and Evaluation

Program Evaluation for Augustine Literacy Project-Charlotte

Qiao Liu

Carl D. Westine

Richard Lambert

A PUBLICATION OF THE CENTER FOR EDUCATIONAL MEASUREMENT AND EVALUATION

Program Evaluation for Augustine Literacy Project-Charlotte

Qiao Liu, M.A.

Carl D. Westine, Ph.D.

Richard Lambert, Ph.D.

Center for Educational Measurement and Evaluation

University of North Carolina at Charlotte

December 2025

Table of Contents

Executive Summary

The evaluation examined the effectiveness of the Augustine Literacy Project-Charlotte (ALP) tutoring in improving early literacy skills among students in early elementary grades. This tutoring program is based on the Orton-Gillingham approach and utilizes volunteer tutors meeting twice per week for 45 minutes with students from under resourced schools in Charlotte, NC. Specific attention was placed upon assessing effects associated with cumulative technology and data-driven program enhancement efforts to improve training and promote standardization during tutor sessions, which were realized in year three. Administrative datasets were obtained from 19 Charlotte-Mecklenburg elementary schools over three academic years (2021-22, 2022-23, and 2023-24). At baseline, students who received ALP tutoring (treatment group) differed from non-participating peers (untreated group) in both demographic characteristics and initial literacy (i.e., DIBELS) scores. To address these pre-existing differences, propensity score matching (PSM) was employed to create statistically comparable groups for outcome analysis.

Overall, students who received ALP tutoring demonstrated greater gains in literacy skills as measured by DIBELS composite scores than their matched peers. Positive effects, though not always statistically significant, were consistently observed among first-grade students across all three years. For second graders, positive treatment effects were evident in two of the three years. Furthermore, during the 2023-2024 academic year when ALP operated with enhanced tutor training, implementation guidelines, and progress monitoring, participation in ALP was associated with enhanced outcomes for students: both first and second graders in the ALP group showed statistically significant literacy gains compared to their matched counterparts, and there were substantial increases in the percentage of students who were performing at or above grade level in reading skills by the end of the academic year.

Summary of Key Findings

- On average and as intended, students in the ALP group began the school year with substantially lower literacy scores compared to their peers. Specifically, ALP treated first graders scored approximately 10 points lower, and ALP second graders scored 15 points lower than their non-treated counterparts.
- Across the three years combined, participation in ALP was associated with statistically significant higher DIBELS scores for matched first-grade students ($\beta = 4.82$, $p = .0096$).
- The largest gains occurred in the 2023-24 academic year, the first year of enhanced delivery. In this year, participation in ALP was associated with statistically significant higher DIBELS scores for matched first-grade students ($\beta = 5.77$, $p = .039$) and second grade students ($\beta = 6.40$, $p = .040$).
- During the 2023-24 academic year, first-grade ALP students demonstrated a 46.1 percentage-point increase in reaching target benchmark levels (i.e., core support) on the literacy assessment. This gain was nearly double that of the matched comparison group, who improved by 23.7 percentage points.

List of Tables and Figures

Project Overview

Augustine Literacy Project - Charlotte (ALP) contracted with the University of North Carolina at Charlotte (UNC Charlotte) to evaluate its literacy tutoring program. The evaluation involved two distinct phases. In the first phase the evaluation team examined implementation of the tutoring experience. The evaluators sought to describe and understand the benefits students experience from working with the ALP tutors. Additionally, the evaluators wanted to identify opportunities for growth with respect to the training of tutors and the tutors' experiences working with students. The phase included a survey of tutors along with follow-up interviews involving a subset of respondents. It also included a focused analysis of test score gains for a small group of ALP recipients in 2021-22.

Results from phase one illustrated ALP as a strong, well-liked tutoring experience from the perspective of tutors that filled an important need in the community by producing promising growth within the targeted student participants. A full description of the study and summary of phase one findings was captured in a published peer-reviewed article in *Literacy Research and Instruction* (Herrera & Lambert, 2024). Although narrow in scope, the study provided valuable insight into the strengths and challenges of consistently implementing the tutoring process as told by 135 ALP tutors through survey responses and 12 follow-up interviews. Important avenues for improvement were noted and included that some newer tutors desired additional training and there was a general need for greater monitoring and follow-up to ensure consistent implementation of the model. These results were used to help inform several technology and data-driven program enhancement efforts to improve training and promote standardization during tutor sessions. This set the stage for a second phase of the evaluation examining overall effectiveness with respect to student achievement across several cohort years and evaluating the program improvement efforts.

In the second phase, evaluators sought to evaluate the effectiveness of the ALP program using a more comprehensive quasi-experimental design that assessed student learning gains as measured by end-of-year examinations. The effort involved a quantitative analysis of standardized DIBELS test scores across three cohorts (2021-2022, 2022-2023, and 2023-2024). Evaluators compared literacy outcomes for two groups: students that received ALP tutoring (treatment), and an equivalent matched set of students that did not receive ALP tutoring (comparison group). The quasi-experimental evaluation is the focus of this report and is described below.

The Evaluation Team

The Center for Educational Measurement and Evaluation at UNC Charlotte

The Center for Educational Measurement and Evaluation (CEME) led the evaluation effort for the ALP evaluation. CEME is a vibrant and scholarly organization through which educational practitioners, policy makers, and UNC Charlotte faculty and students engage in mutually beneficial projects that lead to evidence-based practices and improved educational outcomes for children and families in the region. The central goal of the Center is to connect the expertise of the faculty of the UNC Charlotte College of Education to schools and related agencies to help educational policy makers and administrators make informed decisions about educational practice. The Center provides a vehicle through which the university faculty can establish new and enhance existing research and evaluation collaborations with educational practitioners in the region.

Richard Lambert, Ph.D., CEME director, and a team of faculty affiliates have extensive experience evaluating educational interventions and developing and validating educational and psychological measurement tools. The CEME staff of over 30 employees includes experienced researchers, mentors, evaluators, project coordinators, regional managers, and graduate assistants. Collectively, CEME provides the following services to school systems and related agencies in the region: teacher performance evaluation, mentoring for teachers, program evaluation services, statistical consulting and research methodology advice, and measurement expertise and technical assistance. Recent projects have included a formative evaluation of Charlotte-Mecklenburg Schools' Strategic Plan, conducting evaluation and research studies for the North Carolina Department of Public Instruction to support the implementation of the Early Learning Inventory, and providing mentoring, performance evaluation, and professional development services to the North Carolina Pre-Kindergarten program.

The CATO College of Education

CEME is housed within the CATO College of Education at UNC Charlotte. The college vision is to be a national leader in educational equity through excellence and engagement. The college is organized into five academic departments: Counseling, Educational Leadership, Middle, Secondary and K-12 Education, Reading and Elementary Education, and Special Education and Child Development. The college serves over 2,000 students and offers a wide range of undergraduate degrees, Graduate Certificate programs, Master's degrees, and Doctoral degrees. Specifically, the college offers programs across a variety of disciplines:

- Bachelor's degrees that lead to initial teaching licensure (5),
- Graduate Certificate programs that lead to initial or add-on licensure (21),

- Master of Arts in Teaching programs that lead to initial teaching licensure (6),
- Master of Education programs that lead to advanced teaching licensure (6),
- Additional Master's degree programs (M.S.A., M.A., and M.Ed.) (8),
- Doctoral degrees (5).

The Cato College of Education includes over 115 full-time professors who provide instruction, advise students, conduct research, and serve both the university community and the broader education profession through service at the local, state, national, and international levels. These faculty collectively possess expertise across a wide variety of educational disciplines and stand ready to contribute to the success of local educational partners. In addition to CEME, the Cato College of Education includes centers that focus on specific areas of expertise such as the Mebane Early Literacy Center, the Center for Adolescent Literacies, the Center for Science, Technology, Engineering, and Mathematics (STEM) Education, the Multicultural Play Therapy Center, and the Urban Education Collaborative.

The CEME Evaluation Team for ALP

Several CEME faculty and doctoral students in the Cato College of Education's Educational Research, Measurement, and Evaluation program provided sustained contributions to the evaluation effort. Dr. Richard Lambert oversaw the full evaluation effort across both phases. The initial phase of the evaluation involved ERME doctoral student Leonardo Herrera. The second phase included ERME student Qiao Liu and CEME-affiliated faculty Dr. Carl Westine. Various other CEME-affiliated faculty and students also contributed technical support to the efforts and provided constructive feedback during the two project phases.

Project Scope

The purpose of this study is to evaluate the effectiveness of the ALP tutoring program in improving early literacy outcomes among early-grade elementary students. This evaluation was conducted using student-level data collected across three academic years (2021-22, 2022-23, and 2023-24). The multi-year structure enables the estimation of both year-specific and aggregated treatment effects. A primary challenge in evaluating ALP is that students were not randomly assigned to tutoring services; rather, they were identified by teachers based on perceived need. As a result, students in the ALP program typically differ from non-tutored peers on pre-existing characteristics, including prior literacy performance, demographic factors, and risk indicators.

To address this selection bias, this study used propensity score matching (PSM) by including a set of demographic characteristics and accounting for baseline differences in literacy performance between students in the tutoring program and those in the comparison group. This analysis provides a more robust estimate of the association between exposure to the ALP treatment and the development of literacy skills as measured by differences in growth rates between the groups.

The methodological literature emphasizes that matching on a broad set of covariates, while important, may not fully address bias from variables that are especially strong predictors of the outcome (Stuart, 2010). Rubin and Thomas (2000) suggested that when such strong predictors exist, for example when including baseline achievement measures, it is advantageous to combine PSM with additional adjustment for these key prognostic covariates.

To estimate the effect of ALP participation on end-of-year literacy outcomes, the analysis strategy was applied separately for each school year. The approach incorporated BOY scores into both the propensity score model and the regression model after matching. This "doubly robust" strategy has been shown to reduce residual confounding and improve estimation efficiency. This is especially when covariates (e.g., BOY scores) are strongly predictive of end-of-year outcomes.

For this study, the evaluation team first assessed the extent to which ALP was reaching its intended audience, then examined the association between ALP tutoring and academic achievement. The latter analyses considered beginning to end of year gains as well as reduced risk for below grade level achievement. Of particular interest was the relative impact for the 2023-24 cohort year compared to prior years. In this year, ALP first implemented a strategic redesign of its delivery model. The redesign emphasized standardization and technology-enhanced tutoring to promote consistency and scalability of the ALP tutoring model.

Research Questions

This evaluation is guided by the following research questions:

1. Did children who were recommended for ALP support services by their teacher score lower on measures of basic early literacy skills relative to peers not identified by their teachers?
2. Did students who received ALP tutoring services demonstrate higher growth in basic early literacy skills when compared to a demographically similar group of students who did not receive those services?
3. Did students who received ALP tutoring services demonstrate reduced risk for below grade level academic achievement as compared to a demographically similar group of students who did not receive those services?
4. Was the growth in basic literacy skills made by ALP students relative to the comparison group greater in AY 23-24 (when tutors received enhanced implementation guidelines, training, and monitoring) than in previous years?

Methods

This study employed a quasi-experimental design using PSM to estimate the treatment effect of the ALP tutoring program on first- and second-grade students' literacy outcomes. The analysis compared students who received ALP tutoring (treatment group) with similar students who did not (comparison group). By matching students on relevant baseline characteristics, the study aimed to reduce selection bias and provide a more rigorous estimate of the program's effectiveness.

The Treatment

Since 2005, ALP has delivered high-impact, one-to-one literacy tutoring across Charlotte-Mecklenburg Schools. It is based upon the Orton-Gillingham Approach (Orton-Gillingham Academy, n.d.). This approach involves a structured and focused one-to-one delivery model involving both incremental and cumulative training (Sayeski et al., 2019). Early-grade (typically kindergarten, first grade, and second grade) students and volunteer tutors work collectively through cognitive explanations, using tailored diagnostic and prescriptive assessments, and personalized linguistics-based instruction. Each tutoring interaction lasts for 45 minutes twice per week, and the entire process is designed to continuously activate learning through multisensory engagement strategies.

Over the years, ALP has helped over 2700 students from under-resourced communities achieve reading proficiency. In a recent survey of tutors (conducted in the first phase of the evaluation effort), it was frequently acknowledged that tutors were driven by the social and educational inequities that exist in the community and are motivated to serve to help reduce opportunity gaps (Herrera & Lambert, 2024). The emphasis on individualized instruction was noted as highly rewarding to both the tutors and students and often cited as a central strength of the process of equipping kids with the foundational reading skills needed to succeed in school and in life.

Enhanced Treatment Delivery

In 2024, ALP completed a two-year transformation to standardize and scale its K–2 literacy tutoring model. Through scripted lessons, a remote tutor delivery platform, and continuous improvement systems, ALP achieved instructional fidelity and measurable quality across settings. These reforms have reduced tutor burden, increased engagement, and enabled more students to be served without compromising outcomes.

The process to revamp ALP delivery involved two incremental changes. Beginning in spring 2023, ALP implemented a series of strategic milestones designed to strengthen instructional quality, enhance consistency, and enable growth without compromising effectiveness. The school year ending 2024 (SYE 2024) marked a seismic shift—the culmination of two years of focused work on standardization and technology-enhanced tutoring to achieve scalability, fidelity, and reduced costs. This transformation combined

data-driven design, automation, and quality assurance practices to build a more efficient, measurable, and replicable tutoring system. Key developmental milestones for SYE 2024 included:

- In **May 2023**, ALP launched *scripted lessons through a summer reading program*, marking the organization's first structured effort to ensure lesson fidelity and consistent instructional delivery.
- By **August 2023**, the *Standard Curriculum* became the foundation for all new tutor training classes, signaling a move toward a uniform instructional framework across the organization.
- Between **August and October 2023**, ALP retrained 72 existing tutors, transitioning them from the legacy ALP-branded curriculum to the new scripted model, achieving near-total alignment across active tutors.
- In **January 2024**, ALP launched remote tutoring through its *remote tutor delivery platform, Pencil Spaces*—integrating scripted lessons, remote instruction, and data capture into a single, seamless environment. ALP's initial evaluation of the program found *no measurable difference in performance between in-person and remote tutoring*, reinforcing the strength and reliability of its standardized approach.

Major Shifts in SYE 2024

In addition to the changes in the delivery model, ALP was also able to enhance several aspects of its programmatic infrastructure. This helped to implement better data-driven decision-making for the onboarding and training of tutors, promote greater use of technology, and enhance standardization within tutoring sessions. This transformation also allowed tutors to focus more fully on the child and the power of relationship—deepening engagement and increasing impact by removing tasks that can be systematized and managed through technology. Accordingly, several major shifts were realized in the SYE 2024, prompting further evaluation into their impacts.

- **Data Collection:** For the first time, tutors completed *Session Reflection Forms* after every lesson, creating a consistent feedback loop for program monitoring and continuous improvement.
- **Training:** Tutor certification was redesigned to focus on *lesson delivery and error correction*, while eliminating the need for tutors to create lessons or manage materials. The course was streamlined from four days (24 hours) to three (18 hours), expanding accessibility while maintaining rigor and instructional precision.
- **Assessment:** *DIBELS* was adopted as the organization's primary measure of student growth, aligning internal progress monitoring with district and national benchmarks.

- **Curriculum:** The adoption of *Scripted Lessons* standardized instruction across tutors, improving consistency, reliability, and scalability. This shift also marked the formal start of ALP's *continuous improvement cycle*—using the *Plan–Do–Check–Adjust (PDCA)* process to refine curriculum content, strengthen tutor support, and enhance program delivery through real-time data analysis.

Together, these changes marked a system-wide evolution for ALP—embedding fidelity, data-driven reflection, and instructional precision into every tutoring session. SYE 2024 was the point where these reforms converged, transforming ALP from a tutor centric and dependent model into a scalable, technology-enabled, and quality-assured literacy system. Prior to SYE2024, ALP had relied on training volunteers in the breadth and depth of early literacy instruction. Upon completion of these systemic improvement efforts ALP now consists of a unified model that integrates methodology and deployment—centrally monitored, measurable, and continuously improved through analytics and real-time feedback. As a result of the investment in more powerful systems and prioritization for greater efficiency, many ALP tutors have been inspired to take on a second student. This has raised important questions about effectiveness, and ultimately the scalability of the model as ALP looks to expand its service to a greater number of students in the region.

The Outcome Measure

The outcome measure for this evaluation is a composite Dynamic Indicators of Basic Early Literacy Skills (DIBELS) scale score. DIBELS is a reading skills assessment that ALP students must take at their respective schools at the beginning, middle, and end of their school year. This measure has three primary purposes: to identify students at risk of reading difficulties, to record student progress in reading skills because of intervention programs, and to establish minimum levels of performance based on benchmark goals defined by DIBELS developers. To analyze growth over time, the researchers examined student assessment scores from three time points: the beginning, middle, and end of the school year. Given sample size restrictions within each grade and year, statistical comparisons utilized only the beginning- and end-of-year observations.

Study Participants

Administrative datasets were obtained for 19 elementary schools within Charlotte-Mecklenburg Schools in which ALP provides tutoring. The combined datasets included 9,881 first and second graders enrolled across three academic years: 2021-2022, 2022-2023, and 2023-2024, approximately 3300 per year. The administrative datasets provided by the district included demographic information (e.g., gender, race, English learner status, special education status), school identifiers, and student literacy scores collected at the beginning (BOY), middle of year (MOY), and end (EOY) of each school year. A total of 8,522 records included complete information on all variables of interest. A summary of the dataset showing the percentage of students in each variable group by year is presented in Table 1.

Table 1
Demographic Characteristics of the Samples by Year

Characteristic	Academic Year 2021-22	2022-23	2023-24
n	2809	2848	2865
Treated (ALP)	2.5%	5.1%	6.8%
Untreated	97.5%	94.9%	93.2%
Female	50.6%	50.8%	51.8%
Male	49.4%	49.2%	48.2%
Students with Disabilities	9.7%	11.3%	12.1%
English Learners	36.1%	39.7%	41.6%
AIG	3.0%	3.5%	3.5%
White	8.2%	8.0%	9.1%
Black	40.6%	39.6%	37.2%
Hispanic	44.2%	46.7%	48.5%
Multi-Racial	2.6%	2.2%	2.4%
Native American	0.1%	0.1%	0.0%
Unknown Race	4.2%	3.2%	2.7%

The treatment group included any student that received the ALP tutoring during each year. Matched comparison groups were generated from the remaining students in the respective grades that did not receive the ALP tutoring. The treatment group constituted between 1.6% and 4.6% of the full dataset each year, reflecting the increased presence of ALP within each school. The vast majority of students had complete data on all the variables, but missing data was present in certain variables. Table 2 provides a summary of demographic information and other key variables for treated and untreated students in grades 1 and 2 across the three years combined. The datasets for matching included 4,262 first-grade and 4,260 second grade students. Thus, on average the treatment group represented 5.8 percent and 3.9 percent of data for the two grades, respectively.

Table 2

Descriptive Statistics for Treated and Untreated Groups by Grade - All Years Combined

	1st Grade		2nd Grade	
	Treated	Untreated	Treated	Untreated
n	247	4015	165	4095
Female	43.3%	51.6%	47.9%	51.1%
Male	56.7%	48.4%	52.1%	48.9%
Students with Disabilities	4.5%	10.9%	9.1%	11.7%
English Learners	24.3%	41.0%	28.5%	38.6%
White	6.7%	8.4%	3.6%	8.1%
Hispanic	34.8%	47.9%	45.4%	46.1%
Black	54.7%	37.8%	47.9%	39.4%
Multi Racial	2.8%	2.3%	1.8%	2.5%
Native American	0.3%	0.1%	0.0%	0.1%
Unknown Race	0.8%	3.4%	1.2%	3.6%
Composite Score BOY	316.54	326.48	305.55	320.24

Baseline comparisons on demographic variables highlighted several differences between the treated and untreated groups including gender, English learners, and within several racial subgroups. Additionally, although both treated and untreated groups included students in all reading support categories/risk as defined by University of Oregon DIBELS (2020; Intensive/at risk, Strategic/some risk, Core/minimal risk, and Core/negligible risk) at pretest, ALP students were more likely to be identified in the lowest risk range. The average pretest literacy score for the ALP group fell into the "at risk" group each year (M = 312.4, 315.5, and 319.3) while the average pretest literacy score of the untreated group was consistently higher, and in the final years fell into the "some risk" and "minimal risk" category (M = 313.9, 327.5, and 331.3). These pretreatment differences between the ALP students and untreated students necessitated the use of PSM to reduce selection bias and improve comparability.

Analyses

The Matching Procedure

Matching was conducted separately for each school year using the MatchIt package in R. Students in the treatment group were matched with those in the comparison group based on their estimated propensity scores, which represented the likelihood of receiving ALP support given their observed characteristics, including BOY literacy score, gender, race, English learner (EL) status, students with disabilities (SWD), and academically or intellectually gifted (AIG). Several matching strategies were assessed for accuracy, but the PSM method chosen for this study was the nearest neighbor with a 1:1 matching ratio given the relative quality of the matches and the ease of interpretation. All covariates included in the model were first tested for a significant correlation with the treatment or with the outcome in at least one of the years. Given small sample sizes, to maintain consistency across years, all indicator variables for racial groups were included in the model regardless of significance.

Once the propensity scores were calculated, the MatchIt algorithm was used to assign nearest neighbor matches based on propensity scores. Matching was restricted to exact matches based on school and disability status. Matching within schools strategically aimed to account for unmeasured factors that are likely shared by students attending the same school, such as free/reduced lunch, neighborhood socioeconomic status and school resources, which are not available in the dataset. Additionally, exact matching on the disability status indicator ensured that students with disabilities were also compared to peers with the same disability status. This addressed potentially large differences in educational needs and supports for the few students with a disability designation in the dataset. Together, these strategies enhance the validity of the estimates by further minimizing confounding from both observed and unobserved characteristics.

Standardized mean differences between the treatment and comparison groups are used to assess group equivalence. As shown in Table 3, prior to matching group differences were large, reflecting the baseline differences present between the two groups. After matching, most demographic covariates achieved adequate balance ($|SMD| < 0.10$), with a few exceptions. However, each of the BOY composite scores across the two grades was still slightly elevated. To account for the remaining differences, researchers used a doubly robust approach (i.e., incorporating covariates into both the matching process and the estimation model) because the matching was not perfect (Rubin & Thomas, 2000).

Table 3

Standardized Mean Differences (SMDs) Before and After Matching by Grade

	1st Grade		2nd Grade	
	Before	After	Before	After
Distance	0.719	0.097	0.677	0.106
Female	-0.167	0.008	-0.065	0.000
Students with Disabilities	-0.314	0.000	-0.090	0.000
English Learners	-0.389	0.009	-0.225	0.013
Hispanic	-0.271	0.043	-0.011	0.073
Black	0.341	0.008	0.171	-0.109
Multi Racial	0.039	-0.098	-0.054	0.045
Native American	0.048	0.000	-0.028	0.000
Unknown Race	-0.279	0.045	-0.222	0.111
Composite Score BOY	-0.948	-0.176	-1.179	-0.149

Results

Findings for each research question are presented below. The first research question examines group composition prior to treatment administration and broadly compares ALP students to all other students in the data set. The remaining three questions examine group difference between ALP students and a more focused comparison group of students that did not receive ALP. Each of the four research questions is repeated below for easy reference.

Research Question 1 (RQ1) – Differences at Baseline

Did children who were recommended for ALP support services by their teacher score lower on measures of basic early literacy skills relative to peers not identified by their teachers?

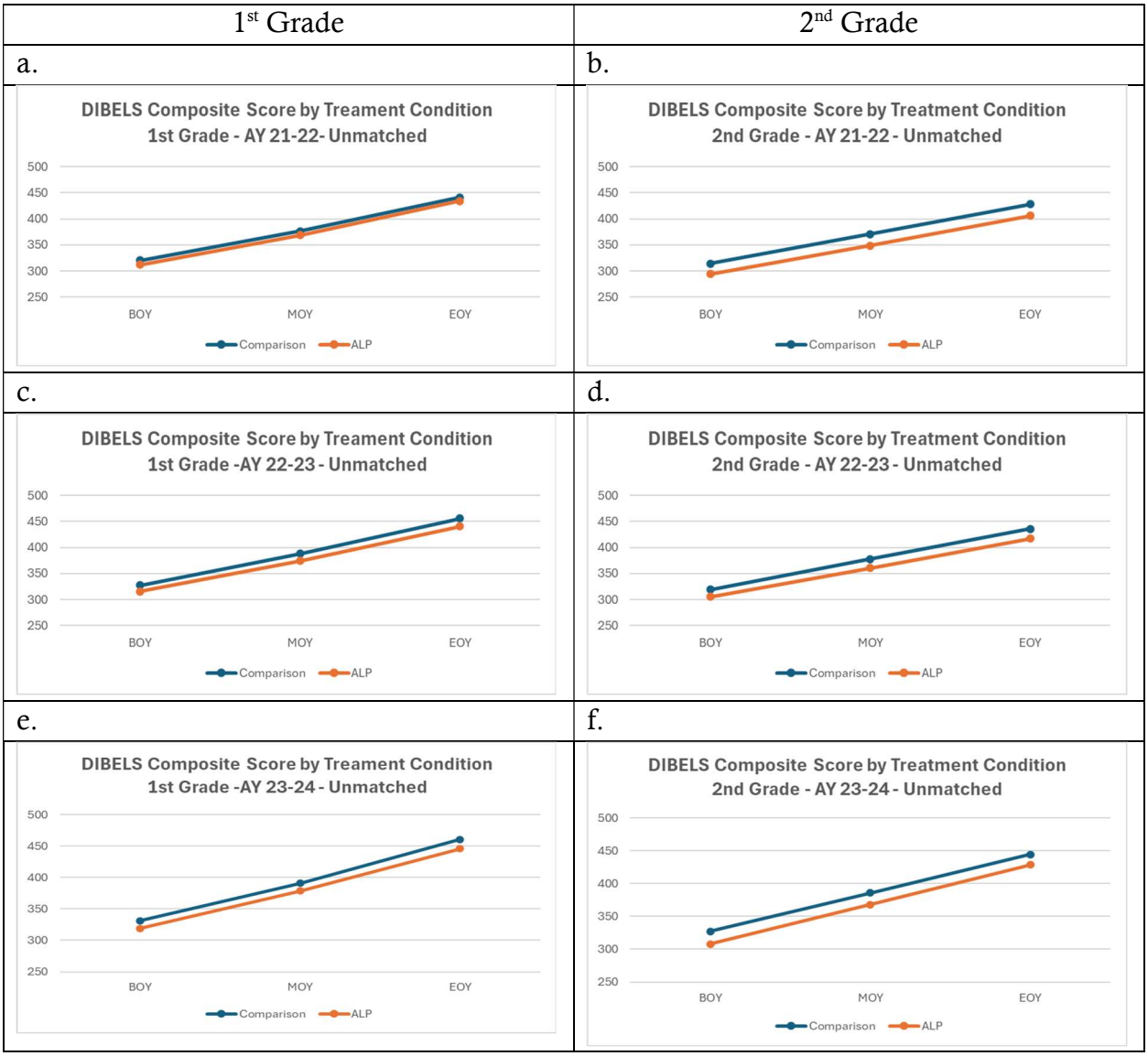
Students who received the treatment were compared to all other students in the same grade level, who attended the same schools during the same academic year, but did not receive the treatment. From Table 2, it is easy to see that across three years composite scores for the treated students are lower at BOY than for untreated students. However, to address research question 1, researchers used t-tests to compare the average BOY DIBELS composite scores for the treatment group (i.e., ALP) to those of the untreated students for each cohort year. Levene's tests of equality of variance showed the variances for the ALP group were statistically significantly smaller than the variances for the comparison group for both grade levels across all three academic years, so unequal variances were assumed.

For the 2021-22 academic year, first grade ALP students scored, on average, 8.38 points lower than comparison group students ($t_{(87.44)} = 5.29$, $p < .001$). Using Hedges unbiased standardized mean difference effect size, this difference translates into an effect size of -.328. Second grade ALP students scored, on average, 21.62 points lower than comparison group students ($t_{(16.54)} = 10.01$, $p < .001$). This difference translates into an effect size of -.798. For the 2022-23 academic year, first grade ALP students scored, on average, 11.96 points lower than comparison group students ($t_{(126.19)} = 7.76$, $p < .001$). This difference translates into an effect size of -.433. Second grade ALP students scored, on average, 13.62 points lower than comparison group students ($t_{(103.518)} = 7.61$, $p < .001$). This difference translates into an effect size of -.489. For the 2023-24 academic year, first grade ALP students scored, on average, 12.04 points lower than comparison group students ($t_{(425.95)} = 10.57$, $p < .001$). This difference translates into an effect size of -.429. Second grade ALP students scored, on average, 19.49 points lower than comparison group students ($t_{(177.42)} = 12.74$, $p < .001$). This difference translates into an effect size of -.612. Hence, at the beginning of the year, ALP students, on average, scored statistically significantly lower than the untreated students for both grade levels, and across all three academic year cohorts. Furthermore, the scores for ALP students were also consistently more homogeneous.

Next, the researchers extended these analyses by using repeated measures MANOVA to explore and describe the growth rates of the treated and untreated groups

across the academic year. These analyses indicated that the growth rates for the treatment group were not statistically significantly different from the growth rates for the untreated condition. These findings were consistent across all grade levels and academic years (see Figures 1a-f). Specifically, the group by time interaction term for first grade students was not statistically significant for any of the academic years: 2021-22 ($F_{(2)}$ = .219, p = .803), 2022-23 ($F_{(2)}$ = .631, p = .532), and 2023-24 ($F_{(2)}$ = .931, p = .395). Similarly, the group by time interaction term for second grade students was not statistically significant for any of the academic years: 2021-22 ($F_{(2)}$ = .199, p = .819), 2022-23 ($F_{(2)}$ = 1.916, p = .148), and 2023-24 ($F_{(2)}$ = 1.093, p = .336). These results demonstrate how students in both groups, on average, made similar gains across the academic year.

Figure 1. *DIBELS Composite Scores by Treatment Condition*



It is important to reiterate that these results were conducted using an untreated comparison condition that included all students in the same grades within the study schools who did not receive the treatment. Furthermore, the research design did not include random assignment of students to treatment conditions, and membership in the treatment condition involved censoring the population by intentionally selecting lower achieving students. Thus, to facilitate stronger inferences that would address the remaining research questions more directly, given the differences in group composition, it was necessary to introduce a method that would create a comparison group that was demographically and developmentally similar to the treatment group.

Propensity scores were calculated and utilized to generate statistically equivalent groups and improve comparability. Using observable pretreatment information (e.g., demographic characteristics and pretest scores), researchers used the PSM method to identify a focused comparison group of untreated students who had a similar probability of receiving the treatment but did not participate in the ALP tutoring service. For the remaining research questions, the researchers compared ALP students to the matched comparison group.

Research Question 2 (RQ2) – Program Effectiveness using Scale Scores

Did students who received ALP tutoring services demonstrate higher growth in basic early literacy skills when compared to a demographically similar group of students who did not receive those services?

To examine the association between ALP tutoring and students' literacy growth, a doubly robust estimation model was employed using post-matching samples from three school years (2021-22, 2022-23, and 2023-24). The outcome variable was students' End-of-Year (EOY) scale scores on DIBELS, with ALP indicating the treatment participation. Results of the doubly robust estimation model are presented in Table 4. There was not a statistically significant association found for the 2021-2022 or 2022-2023 school years. There was a statistically significant and positive effect of the ALP treatment in the 2023-2024 school year for both first graders and second graders. Specifically, in 2023-2024, first graders who received ALP tutoring made an average of 5.77 points more progress in early literacy compared to their matched peers who did not receive the ALP support. Similarly, second graders in the ALP group made 6.40 points more progress than their matched counterparts.

Given the targeted student population, it is important to acknowledge the size of the effect at EOY represents about half the range of the strategic support category that characterizes students as "some risk" (University of Oregon, 2020). The estimated effects in the previous two years, which contained smaller samples, were not statistically significant, though the effects were all in a positive direction.

Table 4

Estimated ALP Effects on EOY Literacy Scores for Students Across Three School Years

	2021-2022		2022-2023		2023-2024	
	β_Treat	*p*	*β_Treat*	*p*	*β_Treat*	*p*
1st Grade	3.215	.373	1.370	.716	5.767	.039*
2nd Grade	2.968	.657	-1.180	.689	6.397	.040*

Note. Count of matched pairs for successive years by grade: 1st grade - (57, 76, 114); 2nd grade - (14, 69, 82).

To assess overall effectiveness, Table 5 summarizes the aggregated effects of ALP across all three years by grade. Results indicate a statistically significant positive effect for first-grade students ($\beta = 4.82$, $p = .0096$). The aggregated effects for second-grade students were not statistically significant, though the direction was positive for second graders.

Table 5

Estimated ALP Effects on Aggregated Literacy Outcomes by Grade

	1st Graders (n=247)		2nd Graders (n=165)	
	β_Treat	*p*	*β_Treat*	*p*
Estimated ALP Effect	4.819	.010**	2.928	0.146

Research Question 3 (RQ3) – Program Effectiveness using Risk Levels

Did students who received ALP tutoring services demonstrate reduced risk for below grade level academic achievement as compared to a demographically similar group of students who did not receive those services?

To address this research question, student performance was categorized into the four DIBELS risk levels - Blue (Negligible Risk), Green (Minimal Risk), Yellow (Some Risk), and Red (At Risk) at the beginning, middle, and end of each academic year. The tables below summarize the percentage of students from both the ALP and matched comparison groups in each risk category by grade level and treatment status over three academic years

(2021-22, 2022-23, 2023-24). Tables 6-8 summarize the shifts in students' risk levels by grade, treatment condition, and time point across the 3 academic years.

Table 6

Percentage At Each Level by Grade, Treatment Condition, and Time of Year in 21-22

Grade Level	Treatment Condition	Level	BOY	MOY	EOY	% Point Change
First	ALP	Blue - Core Support, Negligible Risk	0.0%	1.8%	5.3%	5.3%
		Green - Core Support, Minimal Risk	3.5%	7.0%	31.6%	28.1%
		Yellow - Strategic Support, Some Risk	24.6%	17.5%	19.3%	-5.3%
		Red - Intensive Support, At Risk	71.9%	73.7%	43.9%	-28.0%
	Comparison	Blue - Core Support, Negligible Risk	1.8%	1.8%	5.3%	3.5%
		Green - Core Support, Minimal Risk	3.5%	7.3%	28.1%	24.6%
		Yellow - Strategic Support, Some Risk	12.3%	12.7%	19.3%	7.0%
		Red - Intensive Support, At Risk	82.5%	78.2%	47.4%	-35.1%
Second	ALP	Blue - Core Support, Negligible Risk	0.0%	0.0%	0.0%	0.0%
		Green - Core Support, Minimal Risk	0.0%	0.0%	7.7%	7.7%
		Yellow - Strategic Support, Some Risk	0.0%	7.1%	7.7%	7.7%
		Red - Intensive Support, At Risk	100.0%	92.9%	84.6%	-15.4%
	Comparison	Blue - Core Support, Negligible Risk	0.0%	0.0%	0.0%	0.0%
		Green - Core Support, Minimal Risk	0.0%	0.0%	21.4%	21.4%
		Yellow - Strategic Support, Some Risk	7.1%	7.1%	21.4%	14.3%
		Red - Intensive Support, At Risk	92.9%	92.9%	57.1%	-35.8%

The 2021-2022 Cohort

Across the first year of ALP implementation (21-22), both ALP and matched comparison students demonstrated meaningful movement out of the highest risk category "red" and into lower risk benchmark levels over the course of the school year. Among first graders in this cohort, both the treatment and comparison groups showed substantial reductions in the proportion of students classified as "At Risk" from BOY to EOY. When focusing on students who reached the "Blue" or "Green" benchmark levels - indicating minimal to negligible risk - both groups demonstrated meaningful improvement. However, the ALP group showed a larger overall gain, with a 33.4 percentage point increase in students achieving these higher benchmark levels compared to a 28.1 point increase in the comparison group.

In second grade, students in the ALP group showed modest improvement in reaching lower-risk benchmark levels, while the comparison group exhibited a more substantial gain. However, interpretation of these results should be cautious due to the small sample size (only 14 matched pairs) and the fact that all ALP students began the year at a more disadvantaged starting point compared to their matched peers - all 14 of them were in the highest risk category "Red".

Table 7
Percentage At Each Level by Grade, Treatment Condition, and Time of Year in 22-23

Grade Level	Treatment Condition	Level	BOY	MOY	EOY	% Point Change
First	ALP	Blue - Core Support, Negligible Risk	1.3%	1.3%	15.8%	14.5%
		Green - Core Support, Minimal Risk	2.6%	18.4%	18.4%	15.8%
		Yellow - Strategic Support, Some Risk	19.7%	15.8%	25.0%	5.3%
		Red - Intensive Support, At Risk	76.3%	64.5%	40.8%	-35.5%
	Comparison	Blue - Core Support, Negligible Risk	1.3%	2.6%	11.8%	10.5%
		Green - Core Support, Minimal Risk	7.9%	18.4%	31.6%	23.7%
		Yellow - Strategic Support, Some Risk	27.6%	18.4%	15.8%	-11.8%
		Red - Intensive Support, At Risk	63.2%	60.5%	40.8%	-22.4%
Second	ALP	Blue - Core Support, Negligible Risk	0.0%	0.0%	2.9%	2.9%
		Green - Core Support, Minimal Risk	8.7%	10.1%	22.1%	13.4%
		Yellow - Strategic Support, Some Risk	7.2%	18.8%	14.7%	7.5%
		Red - Intensive Support, At Risk	84.1%	71.0%	60.3%	-23.8%
	Comparison	Blue - Core Support, Negligible Risk	1.4%	4.3%	4.4%	3.0%
		Green - Core Support, Minimal Risk	11.6%	15.9%	20.6%	9.0%
		Yellow - Strategic Support, Some Risk	7.2%	8.7%	16.2%	9.0%
		Red - Intensive Support, At Risk	79.7%	71.0%	58.8%	-20.9%

The 2022-2023 Cohort

In the 22-23 school year, among both first graders and second graders, ALP and comparison groups continue to show effect in reducing the proportion of students identified as "At Risk". First grade ALP students reduced their "At Risk" status by 35.5 points, compared to a 22.4 point reduction in the comparison group. However, a larger percentage of comparison group students ended in the Blue/Green categories. Among second graders, the ALP group made slightly larger gains 16.3 percent in the Blue/Green categories, compared to 12 percent gain for the comparison group. Both groups had similar reductions in "Red" risk classification.

Table 8
Percentage At Each Level by Grade, Treatment Condition, and Time of Year in 23-24

Grade Level	Treatment Condition	Level	BOY	MOY	EOY	% Point Change
First	ALP	Blue - Core Support, Negligible Risk	0.9%	0.9%	10.6%	9.7%
		Green - Core Support, Minimal Risk	6.1%	19.5%	42.5%	36.4%
		Yellow - Strategic Support, Some Risk	38.6%	31.9%	23.9%	-14.7%
		Red - Intensive Support, At Risk	54.4%	47.8%	23.0%	-31.4%
	Comparison	Blue - Core Support, Negligible Risk	4.4%	4.4%	13.2%	8.8%
		Green - Core Support, Minimal Risk	21.1%	20.4%	36.0%	14.9%
		Yellow - Strategic Support, Some Risk	20.2%	25.7%	17.5%	-2.7%
		Red - Intensive Support, At Risk	54.4%	49.6%	33.3%	-21.1%
Second	ALP	Blue - Core Support, Negligible Risk	0.0%	1.2%	7.4%	7.4%
		Green - Core Support, Minimal Risk	1.2%	11.0%	27.2%	26.0%
		Yellow - Strategic Support, Some Risk	22.0%	23.2%	25.9%	3.9%
		Red - Intensive Support, At Risk	76.8%	64.6%	39.5%	-37.3%
	Comparison	Blue - Core Support, Negligible Risk	0.0%	4.9%	12.2%	12.2%
		Green - Core Support, Minimal Risk	15.9%	19.5%	24.4%	8.5%
		Yellow - Strategic Support, Some Risk	15.9%	19.5%	18.3%	2.4%
		Red - Intensive Support, At Risk	68.3%	56.1%	45.1%	-23.2%

The 2023-2024 Cohort

By the 2023-2024 school year, as ALP implementation strengthened and sample sizes grew, the program effects became more pronounced. First-grade ALP students saw a 46.1 percentage-point gain into Blue/Green benchmark levels. That gain nearly doubled the comparison group's 23.7 percentage-point improvement. Similarly, second-grade ALP students advanced by 33.4 points into lower risk categories, compared to a 20.7-point gain in the comparison group. Reductions in the "At Risk" category were also stronger for ALP students in both grades.

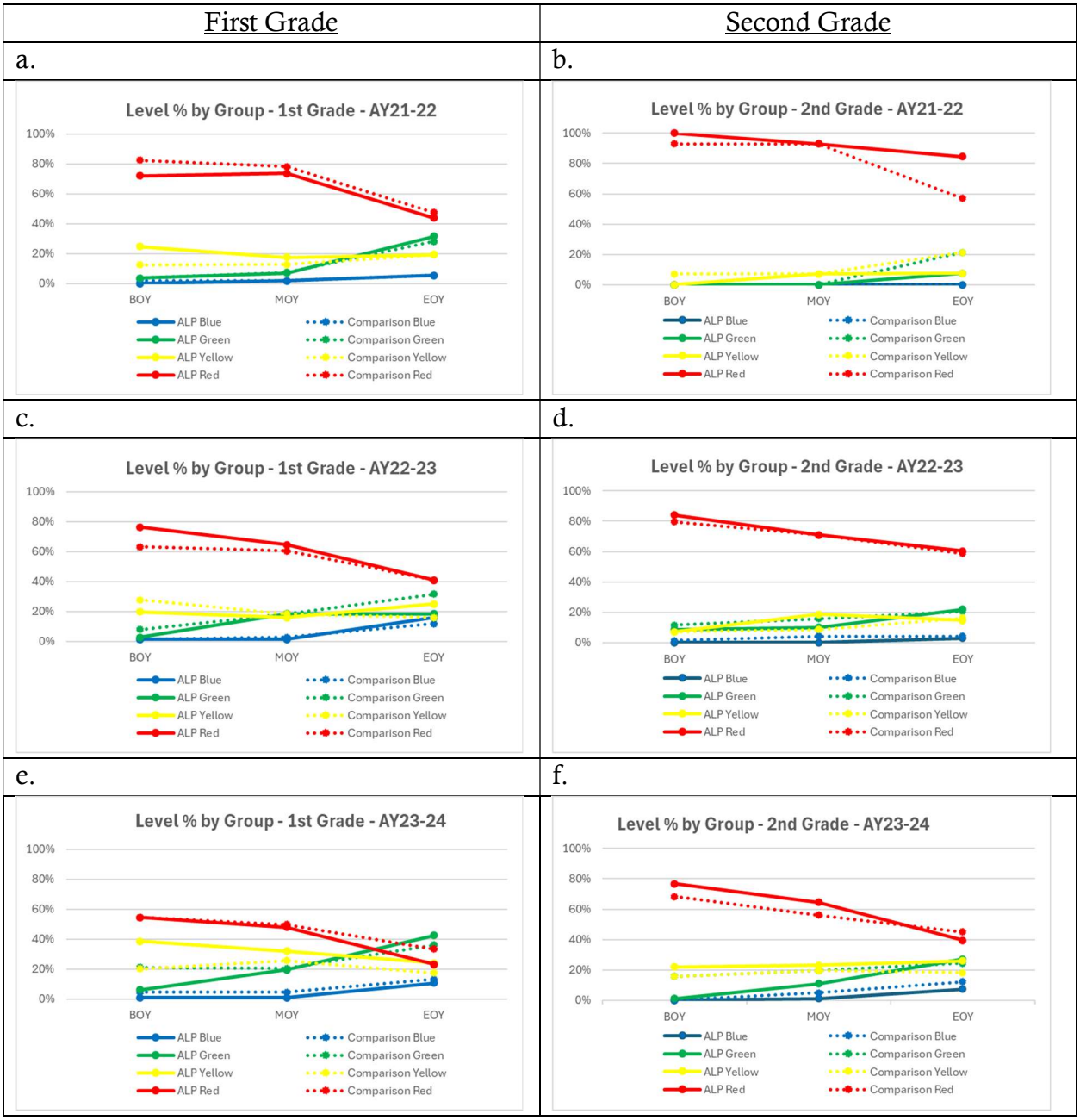
Research Question 4 (RQ4) – Effects of Enhanced Implementation

Was the growth in basic literacy skills made by ALP students relative to the comparison group greater in AY 23-24 (when tutors received enhanced implementation guidelines, training, and monitoring) than in previous years?

While Tables 6-8 document the changes in the percentage of students within each risk category across time, it is helpful to visually compare the change in the DIBELS reading

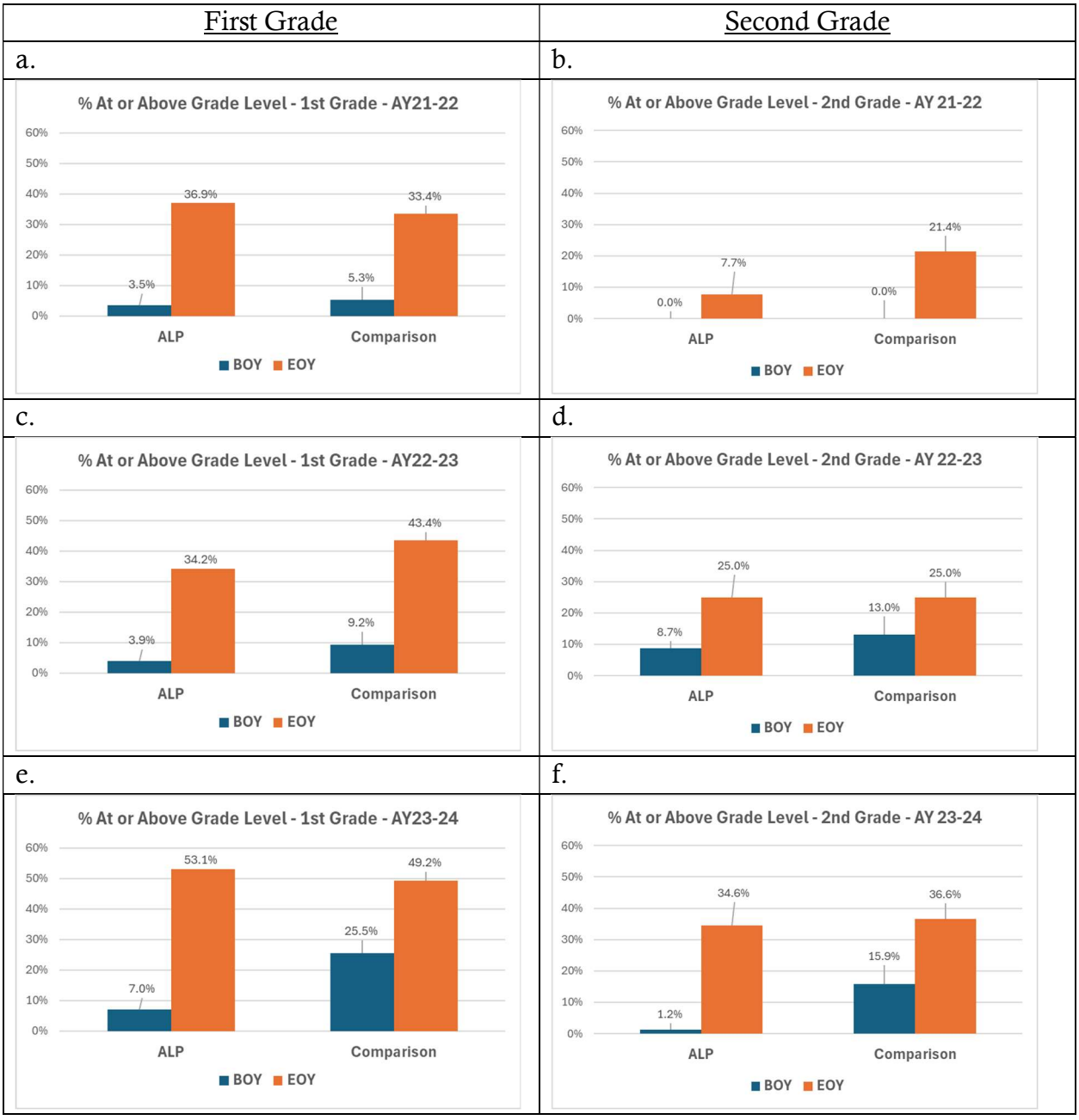
risk levels across the academic year for each cohort group to highlight important differences between the first two cohort years and the last cohort year. Figure 2 illustrates these changes by grade across the three years. Panels a, c, and e of Figure 2 depict these changes for the three first grade cohorts, while Panels b, d, and f illustrate second grade. The ALP treatment group is shown with solid lines, and the comparison group is noted by dotted lines.

Figure 2. Comparison of Trends in Reading Risk Categories by Treatment Condition



While the composition of each group is unique to the cohort year and grade, there is generally a reduction throughout the year in the percentage of students in the highest two categories of risk of reading difficulty and increases in the lowest two risk categories in both grades. These changes tend to be more pronounced in first grade. Of particular interest is the differential amount of change occurring between the solid and dotted lines in any one given year. Movement appears more consistent during the 2021-22 and 2022-23 years, but pronounced in 2023-24, where ALP lines showed the greatest movement. Aggregated changes for the top two risk categories from BOY to EOY are further depicted in Figure 3.

Figure 3. Percent At or Above Grade Level at BOY and EOY by Treatment Condition



In the 2023-24 academic year, during which ALP tutors received enhanced implementation guidelines, training, and monitoring, the data reflect more substantial gains in student literacy performance compared to prior years (see panel e of Figure 3). First-grade ALP students began the 2023-24 year with notably lower proficiency - only 7% were performing at or above grade level, compared to 25.5% of their matched peers. By the end of the year, however, 53.1% of ALP students reached grade-level benchmarks, surpassing the comparison group's 49.2%. This represents a 46.1 percentage point gain for ALP students, which was nearly double the percentage gain among comparison students (23.7%).

Similarly, second-grade ALP students started the year with just 1.2% at or above grade level, which is a much smaller percentage compared to the comparison group's 15.9%. By the end of the academic year, 34.6% of ALP students reached grade-level performance, a 33.4 percentage point gain, compared to the comparison group which had a 20.7 percentage point gain in those categories (see panel f of Figure 3). These gains are mirrored in the DIBELS risk-level plots, which show sharper reductions in the proportion of ALP students in the "Red" category, alongside greater increases in "Green" and "Blue" categories across the year.

Taken together, these findings show that the enhanced implementation of ALP in the academic year 2023-2024 contributed to an accelerated growth in foundational literacy skills among ALP students compared to the other two years, especially for those starting the year at greater risk of future reading difficulties.

Discussion

The present study investigated the effectiveness of the ALP tutoring program on students' literacy growth by employing a doubly robust analytic strategy. The decision was motivated by significant baseline differences observed between the treated and untreated groups: preliminary analyses indicated that the children who receive the treatment began the year with lower initial scores than their peers who did not receive treatment. This highlights that the program targeted students with higher academic needs, which is a strength of the intervention design, but also a source of bias in simple group comparisons.

Findings from RQ1 indicate that the ALP program is successfully targeting students who are performing lower than their peers on targeted reading skills. Specifically, during AY 23-24, first grade ALP students scored .43 standard deviation units below their untreated peers on the BOY assessment. Similarly, second grade ALP students scored .61 standard deviation units below their untreated peers on the BOY assessment during AY 23-24. Effect sizes of this magnitude would be considered moderately large. Hence, without random assignment, this baseline inequivalence necessitated analytic steps to achieve greater comparability between groups. To address the baseline imbalance, the researchers used a doubly robust approach which combines propensity score matching with regression adjustment for baseline literacy scores to estimate program effect. This approach is supported in methodological literature as an effective way to reduce bias when strong predictors of the outcome are available (Rubin & Thomas, 2000).

Specifically, the results from the investigation of RQ2 suggest, for both first and second grade, a statistically significant advantage of approximately 6 points at EOY for the ALP group compared to matched comparisons during AY 23-24. This is particularly compelling given the program's reliance on volunteer tutors meeting only twice per week with students for 45 minutes each time. To evaluate the potential practical meaning of an advantage of this size, this finding can be interpreted relative to the size of the Yellow band (Strategic Support), the band immediately below grade level performance. At EOY, the Yellow band (Strategic Support) is 13 points wide for first grade and 17 points wide for second grade. Therefore, an advantage of 6 points represents approximately 46% of the first grade band and 35% of the second-grade band. When viewed this way, educators could expect first grade ALP students who scored in the upper 46% of Yellow band at BOY to have made enough additional gains compared to matched comparisons to be placed in the Green band at EOY. Similarly, second grade ALP students scoring in the upper 35% of Yellow band at BOY will have made enough additional gains compared to matched comparisons to be placed in the Green band at EOY.

The boundary between Yellow (Strategic Support) and Green (Core Support) is also a useful point of comparison for interpreting these findings as it represents a way to separate students who are at or above grade level performance from those below. The boundary

between the Yellow band and the Green band, for first grade students, is 330 at BOY and 441 at EOY. First grade ALP students scored at BOY, on average, approximately 317 which is below the boundary. At EOY, first grade students scored, on average, 441, which is right at the boundary. The boundary between the Yellow band and the Green band for second grade students is 329 at BOY and 439 at EOY. Second grade ALP students scored, on average at BOY, approximately 320 which is below the boundary. At EOY, second grade students scored, on average, 422, which represents substantial movement toward the boundary.

The results from the investigation of RQ3 build on these concepts by illustrating that for first grade ALP students during AY 23-24, only 7.0% were at or above the Green band at BOY while 53.1% were at or above the Green band at EOY. These findings illustrate a statistically significant 46.1 percentage point gain in at or above grade level performance by ALP students, compared to a 23.7 percentage point gain for first grade matched comparisons. For second grade ALP students during AY 23-24, only 1.2% were at or above the Green band at BOY and 34.6% were at or above the Green band at EOY. These findings illustrate a statistically significant 33.4 percentage point gain in at or above grade level performance by ALP students, compared to a 20.7 percentage point gain for second grade matched comparisons.

Furthermore, with respect to RQ4, this study offered evidence that the ALP program instituted a successfully revised system of tutor training, monitoring, and resource distribution starting in AY 23-24. These improvements to the program not only reduced the delivery costs (broader organizational benchmarks indicate the cost-per-child has reduced 37% since 2021-22) and helped to expand services (broader organizational benchmarks indicate the growth in the number of students served is more than 60% during this timeframe), but also led to enhanced outcomes for the ALP group of students within the study sample.

Limitations and Directions for Future Research

It is important to acknowledge the limitations inherent in the data and methodology. Only one outcome measure, one in use by the partner school system, was used to assess literacy skills. It is possible that the benefits of the ALP tutoring intervention are not measured completely by this one measure. Future research can benefit from a range of literacy measures intentionally tied to the focus of the treatment. Next, although the matching process achieved good balance, the set of covariates available for matching and adjustment was limited: key demographic variables such as free/reduced lunch status or other direct measures of socioeconomic status, student age or grade retention history were not available.

The researchers were also unable to monitor the consistency of treatment implementation across tutors and schools. The lack of these variables means that some

potential confounding factors could not be directly controlled, which may leave residual bias even after matching and adjustment. Future research should investigate the influence of these additional demographic and contextual factors and explore ALP impacts in other grades and for various dosages.

Future research is also needed to demonstrate that the ALP program has continued to implement their tutoring program with sufficient fidelity to yield similar ongoing substantial benefits for students who are struggling with reading. This may help to better explain the role of the enhanced implementation guidelines, training, and monitoring given the positive correlations found, despite the program's relatively unique delivery model of having volunteers working with students only twice per week. Future research designs will also need to extend the existing evidence beyond quasi-experimental designs by randomly assigning students with reading challenges to treatment and comparison conditions.

Furthermore, future research designs will need to include a wider range of student demographic variables, multiple outcome measures, larger sample sizes of both schools and students, and multilevel designs that incorporate the nesting of students within their respective school settings. For the current study, not only were sample sizes per academic year relatively small, but the research team did not have access to data that allowed us to track student performance across years. Therefore, future longitudinal designs will also strengthen the evidence to support the ALP tutoring program by exploring the enduring effects of the program over multiple academic years.

Conclusion

This study offered evidence that indicates ALP is targeting and serving students with demonstrated needs in the area of literacy development. This study also offered evidence to support the conclusion that ALP has improved the implementation and delivery of a systematic, theoretically grounded, and evidence-based tutoring system for first and second grade students who are struggling with acquiring reading skills. The enhanced implementation of this tutoring program yielded substantial increases in the percentage of students who were performing at or above grade level in reading skills by the end of the academic year.

References

Herrera, L., & Lambert, R. G. (2024). Helping children achieve literacy proficiency: A case study. *Literacy Research and Instruction*, *64*(4), 387–410. https://doi.org/10.1080/19388071.2024.2341862

Orton-Gillingham Academy. (n.d.). *Principles of the Orton-Gillingham Approach*. Retrieved from https://www.ortonacademy.org/resources/og-approach-principles-2/

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573–585. https://doi.org/10.1080/01621459.2000.10474233

Sayeski, K. L., Earle, G. A., Davis, R., & Calamari, J. (2019). Orton-Gillingham: Who, What, and How. *Teaching Exceptional Children*, *51*(3), 240-249. https://www.doi.org/10.1177/0040059918816996

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward." *Statist. Sci. 25*(1), 1 - 21. https://doi.org/10.1214/09-STS313

University of Oregon (2020, July). *DIBELS 8th Edition Benchmark Goals*. University of Oregon Center on Teaching and Learning. https://dibels.uoregon.edu/sites/default/files/2024-01/dibels8_benchmark_goals.pdf.