Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1332

Creating Coherence: Does Instructional Alignment Affect the Impact of Tutoring?

Cara Jackson
The Center for Outcomes Based Contracting

Ayman Shakeel
Abt Global

This study examines the impact of using instructionally aligned literacy tutoring with students in kindergarten through third grade under a Response to Intervention framework. We conducted a randomized controlled trial to evaluate the impact on literacy assessment scores for 296 students in four schools in a large suburban school district in the southeastern United States. Students in the treatment group received tutoring where strategies and materials were aligned with core instruction, while those in the control group received tutoring that used supplemental strategies and materials that were distinct from core instruction. We find that students in the treatment group score an average of 0.12 standard deviations higher than the students in the control group. Exploratory analyses reveal that instructional alignment appears to have a greater impact on boys and lower-performing students. Additional exploratory analyses suggest the treatment effect is stronger when delivered in groups of four and by tutors who do not hold a master's degree.

VERSION: November 2025

Suggested citation: Jackson, Cara, and Ayman Shakeel. (2025). Creating Coherence: Does Instructional Alignment Affect the Impact of Tutoring?. (EdWorkingPaper: 25-1332). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/en9j-xp53

Creating Coherence: Does Instructional Alignment Affect the Impact of Tutoring?

Cara Jackson, The Center for Outcomes Based Contracting at SEF

Ayman Shakeel, Abt Global

This study examines the impact of using instructionally aligned literacy tutoring with students in kindergarten through third grade under a Response to Intervention framework. We conducted a randomized controlled trial to evaluate the impact on literacy assessment scores for 296 students in four schools in a large suburban school district in the southeastern United States. Students in the treatment group received tutoring where strategies and materials were aligned with core instruction, while those in the control group received tutoring that used supplemental strategies and materials that were distinct from core instruction. We find that students in the treatment group score an average of 0.12 standard deviations higher than the students in the control group. Exploratory analyses reveal that instructional alignment appears to have a greater impact on boys and lower-performing students. Additional exploratory analyses suggest the treatment effect is stronger when delivered in groups of four and by tutors who do not hold a master's degree.

Introduction

Early literacy is instrumental to students' success, both in terms of academic outcomes and longer-term economic outcomes (Gaab & Petscher, 2022), yet two-thirds of fourth grade students score below proficient in reading on the National Assessment for Educational Progress (NAEP, 2022). Additionally, research consistently shows boys lagging behind girls in early literacy assessments; differences emerge prior to kindergarten and are observed throughout elementary school (Chatterji, 2006). Early interventions to address academic challenges have had mixed results and can create less coherent educational experiences for students. A key question for the field is whether greater instructional coherence between interventions and regular classroom instruction generates better outcomes.

In recent years, tutoring has been advocated as a remedy to low student performance and inequities in educational outcomes. Evidence generally supports tutoring as an effective practice. A meta-analysis of tutoring programs found that tutoring yields consistently substantial positive impacts on learning, with an overall pooled effect size of 0.29 standard deviations and larger effects in early grades (Nickow, Oreopoulos, & Quan, 2023). In a recent large-scale randomized controlled trial, Bhatt et al. (2025) found that tutoring delivered during the 2023-24 school year was effective and impacts were robust across a variety of tutoring models (both in-person and virtual). Yet, effect sizes were considerably smaller than the Nickow et al. meta-analysis, ranging from 0.06-0.09 standard deviations (approximately 1-2 months of additional learning). Furthermore, these overall effects masked considerable variability across sites.

Why is such a promising intervention to improve student learning falling short of its promise? Is it possible to design tutoring interventions in ways that maximize impact? The Framework for High-Impact Tutoring from the National Student Support Accelerator (NSSA) provides recommendations for designing effective tutoring programs yet districts have struggled to implement high-dosage tutoring with fidelity (Carbonari et al., 2024). Since high-dosage

tutoring is often delivered out of school time, making student attendance a challenge, NSSA recommends embedded tutoring during the school day as a solution.

NSSA also recommends that "if classroom instruction is based on rigorous and high-quality materials, the tutoring program aligns to classroom curricula" (NSSA, n.d.). NSSA categorizes this recommendation as lacking a robust research base; however, practitioners and researchers agree on its likely importance for quality. Relatedly, observational data from four districts in Tennessee suggest that instructionally coherent small-group support through high-dosage tutoring could maximize student learning (State Collaborative on Reforming Education, 2023). The four districts captured in the observational data opted into different content and structure. Due to self-selection, the patterns observed in the data may reflect something other than the causal impact of instructional coherence. For example, it is possible that the two districts that used the same instructional materials during intervention as in core instruction also had more effective instructional coaching or school leadership relative to the districts that opted to use supplementary materials during intervention. Identifying the causal impact of coherence requires addressing this potential selection bias by identifying a comparison group that offers a valid counterfactual estimate of achievement trends in the absence of coherence.

Building on lessons learned from the observational data, the district in this study sought more rigorous empirical evidence to inform decisions about its approach to tutoring. This research aims to fill the gap in tutoring literature by providing empirical evidence on the effectiveness of instructionally aligned tutoring. We examine the potential for instructional coherence to bolster the efficacy of tutoring interventions by randomly assigning 328 students to either tutoring in which strategies and materials were aligned with core instruction, or to tutoring that used supplemental strategies and materials that were distinct from core instruction.[1]

[1] The final analytic sample comprises 296 students.

Tutoring was focused on early literacy skills for students in kindergarten through third grade so we study the impact of tutoring on literacy achievement. Additionally, we examine variation in impacts by students' prior achievement, gender, and characteristics of the tutoring groups including dosage and tutor qualifications.

Both the treatment and control groups were offered 40- or 45-minute sessions of tutoring, 5 days a week, by a teacher or a paraprofessional. We find that on average, students in the treatment group outperformed those in the control group by 0.12 standard deviations (p<0.10). We also observe meaningful heterogeneity in the treatment effects. The impact is greater for boys (0.22 standard deviations, p<0.10) and for students whose fall literacy scores fall below the median (0.18 standard deviations, p<0.1). Effectiveness also varies by tutoring group characteristics. Instructionally aligned tutoring yields higher gains for students that received 45-minute sessions (0.31 standard deviations, p<0.05), had a tutor who did not hold a master's degree (0.17 standard deviations, p<0.05), and were in a tutoring group of 4 students (0.49 standard deviations, p<0.01).

In recent years, a rich body of literature has emerged examining the impacts of tutoring on student outcomes. However, researchers and policymakers continue to explore which specific strategies and components make tutoring most effective. This study makes an important contribution to the field by demonstrating that tutoring with an instructionally coherent approach offers a promising pathway to improving student learning in the early grades.

Background

In general, Response to Intervention (RTI) involves educators placing students in reading groups and delivering services based, in part, on students' scores on screening assessments of skills such as word identification and letter sounds (Balu et al., 2015). While there is some evidence that RTI can have negative impacts on early literacy (Balu et al., 2015),

a recent study found that literacy tutoring delivered to at-risk kindergarten through third grade students using an RTI framework had positive impacts on students' reading fluency (Markovitz et al., 2022). Hence the effect of RTI on student achievement remains unclear.

Instructional Coherence and Alignment

One hypothesis for why RTI does not have a consistently positive impact on student achievement is that time spent in intervention decreases students' time in core instruction (Allensworth & Schwartz, 2020). Relatedly, RTI may be inconsistent with core instruction in terms of strategies, academic terminology, routines, and materials. Some literacy tutoring curricula include pedagogical approaches or content that vary from core instruction. Alternatively, an aligned curriculum might offer a deeper dive into the skills, texts, and tasks students encounter during core instruction with their classroom teacher. This alignment, known as instructional coherence, refers to the degree to which curriculum, instruction, and assessment act as a unified system for common learning goals.

The idea of instructional coherence emerged from organizational theory, and researchers have studied its role in school improvement. Newmann, Smith, Allensworth, and Bryk (2001) conceptualized instructional coherence as having three essential components: a common instructional framework that guides curriculum and teaching across the school, staff working conditions that support implementation of the framework, and allocation of resources to schools that advance the instructional framework. In an instructionally coherent learning environment, students experience consistent expectations and pedagogical approaches, thus reducing cognitive load associated with navigating different instructional strategies or academic terminology. Studies from the Chicago Consortium on School Research demonstrate that improvements in instructional coherence are associated with increased student achievement gains in both reading and mathematics (Bryk, 2010; Newmann et al., 2001).

Related to coherence, instructional alignment focuses on the correspondence between intended learning outcomes, instructional activities, and assessment practices. Biggs' (1996) constructive alignment theory posits that optimal learning occurs when teaching methods and assessment directly support the intended learning outcomes. This alignment ensures that what is taught matches what is tested, and both correspond to the stated curriculum goals. Anderson (2002) expanded Biggs' theory to include alignment between written, taught, and tested curricula, noting that misalignment in any area can undermine educational effectiveness. Coherence and alignment may be particularly crucial for struggling learners who benefit from predictable instructional routines.

In the state where the current study was conducted, intervention tutor groups used the same instructional strategies and materials as those used in core instruction, while control tutor groups used different strategies and materials for the intervention period. Previous observational data suggest that students tutored with the same instructional approach as used in core instruction made greater growth on the literacy universal screener than their peers in tutor groups that used different materials for the intervention period (State Collaborative on Reforming Education, 2023). However, the districts examined may have varied along a number of dimensions, including prior achievement and other facets of the tutoring experience. As such, the observational data cannot support causal claims about the effectiveness of aligning tutoring instruction with core instruction.

To address these limitations, this study uses a randomized controlled trial to examine the effect of high-dosage literacy tutoring using an instructionally aligned approach compared to tutoring with different materials and strategies, holding other features of tutoring constant. All curricula are considered high-quality by the state. We are not testing whether one type of curriculum is more effective than another. The contrast in question is between using the same instructional materials and strategies in tutoring as in core instruction, versus using a different

curriculum during tutoring. To the best of our knowledge, this is the first study to examine the causal impact of instructional coherence on literacy achievement.

Tutor Qualifications

Research has consistently demonstrated that tutor expertise influences both the quality of instruction and subsequent student learning outcomes, though the relationship is nuanced (Kohlmoos & Steinberg, 2024; Nickow, Oreopoulos, & Quan, 2023; Robinson & Loeb, 2021). Wasik and Slavin's (1993) meta-analysis of five major tutoring programs found that programs using certified teachers as tutors produced larger effect sizes than those using paraprofessionals or volunteers. However, Elbaum, Vaughn, Hughes, and Moody (2000) found that well-trained paraprofessionals could achieve outcomes comparable to those of certified teachers when provided with structured curricula and ongoing supervision.

Since content knowledge and pedagogical content knowledge are important qualifications for early literacy instruction, training and ongoing support for tutors may explain the discrepancy in findings. Fitzgerald (2001) examined volunteer tutoring programs for first and second grade students in schools with low average reading scores and found that training quality was a stronger predictor of outcomes than tutors' formal educational credentials. An experimental study of a two-week summer institute followed by support during the school year to bolster kindergarten and first-grade teachers' knowledge of language structure and reading development improved students' reading achievement (McCutchen et al., 2002). Additionally, training and support may help tutors become adept at adjusting instruction based on student needs. Chi, Siler, Jeong, Yamauchi, and Hausmann (2001) analyzed tutoring dialogues and found that effective tutors engaged in sophisticated pedagogical reasoning, using student errors as opportunities for targeted instruction rather than simply correcting mistakes.

Program Context / Treatment and Control Conditions

This study is set in a large suburban school district in the southeastern United States that uses high-dosage tutoring as part of its RTI framework. State policy requires the use of assessment data to identify students furthest from grade-level literacy benchmarks. All students below the 40th percentile on the universal literacy screener (aimswebPlus) were eligible for tutoring during the intervention block. The tutoring model that we examine adheres to evidence-informed recommendations for effective scheduling practices: sessions between 30 and 60 minutes; three or more sessions per week for a minimum duration of ten weeks; and tutoring sessions built into each school's master schedule (National Student Support Accelerator, n.d.).

The program and scheduled dosage were intended to vary based on fall assessment results, as shown in Table 1. We define program dosage as the *total amount of tutoring a student would receive under ideal conditions and* scheduled dosage as *the amount of tutoring an individual student would receive if they attended all scheduled tutoring sessions (*Accelerate, 2025).

Program Dosage:

The district aimed to provide tutoring from October 2024 for all grades until the end of the academic year. District guidance established tutoring dosage expectations that vary based on the fall (baseline) score on aimswebsPlus.[2] The district expected 5 days of tutoring per week for students at or below the 25th percentile, and 3 days per week for those over 25th percentile. As illustrated in Table 1, the district expected tutoring group size to vary based on the baseline

[2] For students scoring between the 25th and 40th percentile or those legally required to participate based on promotion and retention criteria, state guidance specified a tutoring dosage of three sessions per week, each lasting 30 minutes. To qualify as tutoring under this guidance, student-to-tutor ratios were required to be 1:3 for three-day-per-week services or 1:4 for five-day-per-week services.

performance on aimswebsPlus. The district did not specify different dosages in terms of length of tutoring sessions and aimed to provide 40-minute sessions.

Scheduled Dosage:

The scheduled dosage varied slightly from the district-expected dosage. While tutoring started in October 2024 for students in grades 1-3, tutoring for kindergarten students started in January 2025 to allow them time to acclimate to school. In total, for both the treatment and control groups, students in grades 1-3 were offered a total of 121 sessions and students in kindergarten received 71 sessions. Scheduled dosage panel in Table 1 illustrates that all students were offered 5 days of tutoring per week, despite the expectation that students in the 26-40th percentile would be offered tutoring 3 days per week. Both Tier 2 students and students above the 25th percentile had smaller than expected group size (about 1 student fewer than district expectations in both cases). Session length varied, with some students receiving 45-minute sessions rather than 40-minute sessions. These longer sessions were mainly concentrated in one school, but some students in other schools also received 45-minute sessions. Session length did not vary across percentile groups (approximately 41.5 minutes per session, on average).

Notably, students were eligible to exit tutoring once they reached the 40th percentile on aimswebPlus. According to the district, the intended dosage (dosage threshold above which the district believes that there will be meaningful impacts) was 10 weeks of tutoring with a session length of 40 minutes, for at least 3 days per week – totaling to 30 sessions and 1,200 minutes.

Table 1: Program and scheduled dosage, by fall (baseline) achievement percentile

	Tier 3 (Baseline Achievement: 1st -10th Percentile)	Tier 2 (Baseline Achievement: 11th -25th Percentile)	Baseline Achievement: 26th – 40th Percentile
Program Dosage			
Frequency (days per week)	5	5	3
Length (minutes per session)	40	40	40
Tutoring Group Size	3	5	3
Scheduled Dosage			
Frequency (days per week)	5	5	5
Length (minutes per session)	41.55	41.35	41.47
Tutoring Group Size	3.39	3.91	3.95
Observations	*174*	*78*	*44*

Notes: Baseline achievement is standardized fall oral reading fluency (ORF) scores for students in grades 1, 2, and 3 and fall early literacy (EL) for students in kindergarten. This table is based on information for students that have non-missing baseline achievement.

The district in this study began using "instructionally aligned" tutoring: foundational literacy skills, texts, and tasks that align to core (Tier 1) instruction. The treatment consisted of small group tutoring aligned with core instruction. Instructionally aligned tutoring groups used Benchmark Advance materials that covered the same skills in the same sequence as core classroom instruction, with the same routines used by classroom teachers.

The business-as-usual group received small group tutoring with curricula distinct from core instruction. Under the district's traditional intervention model, students identified as struggling readers received small-group instruction using specialized intervention programs that operated independently from their core classroom curriculum. Students performing below grade-level benchmarks were pulled out for targeted support using research-based programs like SPIRE, a comprehensive, multisensory reading intervention with systematic, scripted 10-step lessons covering phonemic awareness through comprehension. Others received, for instance, Lexia Core5, which aims to develop fundamental literacy skills through explicit, systematic instruction with personalized learning paths, or SPIRE Foundations: Sounds Sensible, which provides instruction in phonological awareness, alphabet knowledge, and understanding letter-

sound relationships. While all of the traditional intervention programs used with the control group are grounded in reading science and designed to support struggling learners, they employed different instructional materials, vocabularies, and methodological approaches than students encountered during their regular literacy block. This created a bifurcated learning experience where students toggled between their grade-level core instruction and these supplemental intervention systems, requiring them to navigate distinct pedagogical frameworks throughout their school day. This approach represented standard practice across the district for years, reflecting the widespread educational assumption that students significantly behind grade level require specialized, standalone programs separate from the materials used for on-level instruction.

Resources required for implementation included tutors (teachers or paraprofessionals), space for up to five students and one adult, and the curriculum materials. Tutoring took place either in the classroom or in a separate room in the school building. Some control tutor groups required the use of the computer; instructionally aligned tutor groups did not involve any computer-based lessons. For the control group, any tutor type (teacher or paraprofessional) could deliver tutoring. For the instructionally aligned treatment group, district leadership indicated that it was beneficial for the teacher to deliver tutoring since it was consistent with the lessons delivered during core instruction.

Study Design

The study included four research questions.

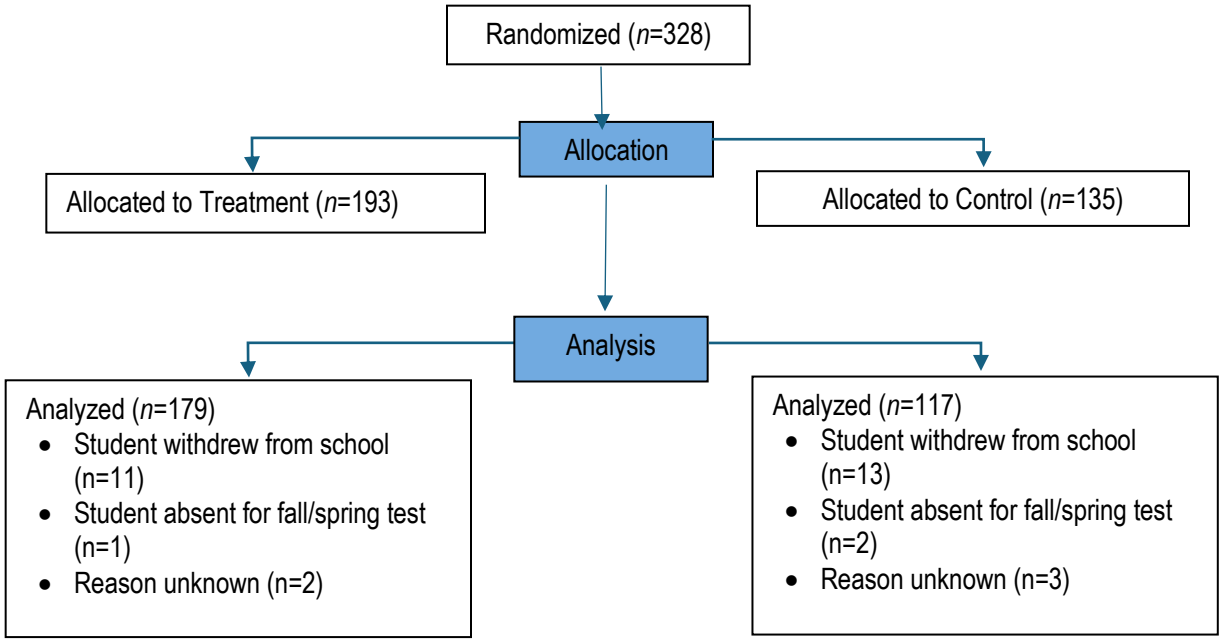
1) *Implementation Question:* To what extent do tutoring groups vary in dosage and tutor qualifications?
2) *Main Impact Question:* For students who start the year performing below the 40th percentile using aimswebPlus, what is the effect of participating in instructionally aligned tutoring grades K-3 on literacy outcomes compared to business-as-usual tutoring?

3) *Exploratory Question:* How do the effects of instructionally aligned high-dosage tutoring on academic achievement differ by students' prior achievement levels and gender?
4) *Exploratory Question:* How do the effects of instructionally aligned high-dosage tutoring vary by tutoring group characteristics?

Methods

This study used a student level randomized controlled trial to address research questions 2-4. We conducted randomization in partnership with the school district and the four partner schools. In each of the four schools, building-level staff first identified students eligible for the study based on fall aimswebPlus assessments. Prior to random assignment, we excluded students who were eligible for other services during the intervention block from the study sample (such as English learners or students with individualized education plans) and from all analyses. We conducted randomization in two phases. In the first phase, carried out in September 2024, we randomly assigned 245 first, second, and third grade students to the treatment and control groups. Students were stratified by school, grade, and skill group, and then randomly assigned within those groups – we refer to this as the randomization block. The next phase of randomization took place in January 2025, where we randomly assigned 83 kindergarten students to the treatment and control groups.

Figure 1. Consolidated Standards of Reporting Trails (CONSORT) diagram



Overall, 193 students were allocated to the treatment group (59%) and 135 (41%) to the control group. The proportion of students assigned to treatment versus control was based on the availability of instructional materials. Two hundred and ninety-six students had complete data for the analytic sample: 179 (60%) in the treatment group and 117 (40%) in the control group (Figure 1).[3] The study was preregistered at OSF Registries (see Appendix C).

Data

The school district partner provided student-level literacy achievement test scores and demographic data for the study sample who were eligible for tutoring based on the district's literacy screener, aimswebPlus. The test score data included results from aimswebPlus Early Literacy and Reading for students in all grades and results from the English Language Arts (ELA) state test for student in grade three. The specific measures covered by aimswebPlus vary across grades and time of year (Pearson, n.d.). In this study, we focus on the Oral Reading

[3] Table A1 shows the breakdown of the number of students randomized intro treatment and control groups by school and grade.

Fluency (ORF) measure for students in grades 1 through 3 and on Early Literacy (EL) for students in kindergarten, the only measures for which all students in the study had data. Internal consistency for ORF is 0.96 and for Early Literacy is 0.91 (Pearson, 2017). The demographic data included gender and race/ethnicity.

The school district provided data on tutoring dosage, as proxied by student school attendance. While the original plan was to track attendance at the level of individual tutoring sessions, technical issues with the tracking system prevented this. As a result, we relied on administrative data shared by the district. The district provided information on tutor and student school attendance, days tutoring occurred, and tutoring exit date (if applicable). The district expressed confidence that when both students and tutors were present at school on a given day, it was highly likely that tutoring took place. Accordingly, we used these data to construct a measure of tutoring dosage.

School partners also provided data on the skills intended to be addressed during tutoring, number of sessions per week, number of minutes per session, the tutoring group size, and tutor type (certified teacher or paraprofessional).

Analysis

We use descriptive statistics on tutoring dosage and tutor qualifications to examine the qualitative features of the support offered, documenting the differences in the characteristics of the support placement settings between treatment and control groups. We then examine average growth in the Oral Reading Fluency score from aimswebPlus from the beginning to the middle of the 2024-2025 school year. The basic model is as follows:

$$Y_{ij} = \beta_0 + \beta_1(T_{ij}) + \beta_2(Y_{ij}^{*}) + \sum_{m=1}^{M} \beta_{3.m} X_{mij} + \sum_{j=1}^{J-1} \beta_{4.j} Block_j + \varepsilon_i$$

Where:

Y_{ij} = the *outcome* for the *i*th student in the *j*th block.

β_0 = the intercept (i.e., the covariate adjusted mean outcome for students in the comparison group in the reference block).

β_1 = the treatment effect (instructionally aligned tutoring).

T_{ij} = 1 if student *i* is assigned to treatment within block *j*, and = 0 if assigned to comparison within block *j*.

β_2 = the effect of pretest.

Y_{ij}^* = a pre-test measure for the *i*th student in the *j*th block.

$\beta_{3.m}$ = the effects of student covariates.

X_{mij} = the *m*th of M additional covariates representing demographic characteristics of student *i* in block *j* (dummy variables for gender and race/ethnicity).

$\beta_{4.j}$ = the effect of randomization block (i.e., the difference in the intercept between block *j* and the reference block).

$Block_j$ = 1 if student is in randomization block *j* (*j*=1,2, …, *J*), otherwise = 0.

ε_i = a residual error term for student *i*.

In the model, the main coefficient of interest is β_1 that reflects the average difference in achievement score between the treatment and control groups, conditional on baseline fall literacy achievement score, which was administered between August 29th and September 6th of the 2024 2025 school year; student demographics; and school, grade, and skill group blocks. β_1 shows the effect of the instructionally aligned tutoring on student achievement. We also test for robustness of the results by including models that exclude the baseline ORF and the student demographic variables. We use robust standard errors which are the most appropriate for individual-level blocked random assignment designs (Abadie et al., 2023).

Findings / Results

Descriptive Statistics

Table 2 presents the baseline equivalence table. The first three columns of the table present baseline analysis for the randomized sample and the last three columns present baseline analysis for the final analytic sample. As the randomization is conducted within

randomization blocks, the difference in column (3) and (6) is residualized by the randomization blocks. As expected, the first three columns of Table 2 show that the two groups are very similar on baseline characteristics. None of the covariate differences between the treatment and control groups are statistically significant, with the exception of the proportion of Asian students, which differs slightly between the two groups (4% in the control group versus 1% in the treatment group). This difference is small and likely attributable to random variation.

As the analytic sample excludes some students in the randomized sample, it is also important to test for balance between the treatment and control group students that contribute to the final analysis. The concern is that if missing data is nonrandom, we cannot confidently claim that the observed effects are a result of instructional coherence. If the missing data is random and not systematically related to the treatment, we would hope to see balance between the treatment and control group students in the analytic sample. The last three columns of Table 2 closely resemble the first three, indicating that the treatment and control groups remain largely comparable (*F=0.73, P = 0.666)*. This similarity suggests that attrition is likely random and does not introduce systematic bias.

Table 2. Descriptive Statistics and Baseline Equivalence

	Randomized Sample			Final Analytic Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Variable	Control	Treatment	Model Adjusted Difference	Control	Treatment	Model Adjusted Difference
Baseline Literacy Achievement*	-	-	-	-1.372	-1.326	0.103
	-	-	-	(0.696)	(0.654)	(0.06)
Female	0.546	0.494	-0.040	0.538	0.497	-0.032
	(0.500)	(0.501)	(0.06)	(0.501)	(0.501)	(0.06)
Black	0.176	0.183	0.015	0.179	0.184	0.014
	(0.383)	(0.388)	(0.04)	(0.385)	(0.389)	(0.04)
Hispanic	0.126	0.178	0.063	0.128	0.179	0.062
	(0.333)	(0.383)	(0.04)	(0.336)	(0.384)	(0.04)
Asian	0.042	0.011	-0.034**	0.043	0.011	-0.034**
	(0.201)	(0.105)	(0.02)	(0.203)	(0.105)	(0.02)
American Indian or Pacific Islander	0.008	0.011	0.003	0.009	0.011	0.003
	(0.092)	(0.105)	(0.01)	(0.092)	(0.105)	(0.01)
Observations	135	193	328	117	179	296

Note: The "Randomized Sample" comprises all students who are randomly assigned to the treatment of control groups, while the "Final Analytic Sample" comprises students with non-missing fall *and* spring scores. Columns (1), (2), (4), and (5) report the means and standard deviations (in parentheses) for each variable and columns (3) and (6) present the residualized difference between the control and treatment groups for each variable along with the standard error (in parentheses) for the difference. The residuals are obtained by regressing the variables on the randomization block fixed effect. *p<0.10, **p<0.05, ***p<0.01
*We use fall ORF score as the *Baseline Literacy Achievement* for students in grades 1, 2, and 3, and fall EL score as *Baseline Literacy Achievement* for students in kindergarten. This variable is missing in the descriptive statistics for the randomized sample as some students in the randomized sample are missing *Baseline Literacy Achievement* score.

Implementation Analysis

RQ1: To what extent do tutoring groups vary in dosage and tutoring group characteristics?

As noted before, both treatment and control students were assigned to high dosage tutoring. To assess the overall impact of the treatment, we examine differences in actual dosage—the amount of tutoring participating students receive (on average) across the implementation period *(*Accelerate, 2025)—and differences in tutoring characteristics between the treatment and control groups. If such differences exist, the estimated treatment effect likely reflects a combination of assignment to instructionally aligned tutoring as well as variations in dosage and tutoring characteristics.

Table 3. Tutoring Dosage, by Group Assignment

	Randomized Sample			Final Analytic Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Variable	Control	Treatment	Model Adjusted Difference	Control	Treatment	Model Adjusted Difference
Scheduled Dosage						
Frequency (days per week)	5.000	5.000	0.000	5.000	5.000	0.000
	(0.000)	(0.000)	(0.00)	(0.000)	(0.000)	(0.00)
Length (minutes per session)	41.407	41.606	0.000	41.368	41.564	0.000
	(2.257)	(2.341)	(0.00)	(2.238)	(2.325)	(0.00)
Group Size	3.548	3.596	0.115	3.615	3.603	0.049
	(1.170)	(1.081)	(0.07)	(1.144)	(1.083)	(0.07)
Actual Dosage						
Sessions Attended	87.526	90.245	1.812	91.966	92.107	-0.664
	(27.31)	(24.897)	(1.87)	(21.519)	(23.680)	(1.06)
Minutes Attended	3,625.889	3,759.089	76.609	3,805.342	3,834.635	-22.488
	(1,1512)	(1,074.929)	(77.51)	(918.950)	(1,026.647)	(43.68)
Observations	135	193	328	117	179	296

Note: The "Randomized Sample" comprises all students who are randomly assigned to the treatment of control groups, while the "Final Analytic Sample" comprises students with non-missing fall *and* spring scores. Columns (1), (2), (4), and (5) report the means and standard deviations (in parentheses) for each variable and columns (3) and (6) present the residualized difference between the control and treatment groups for each variable along with the standard error (in parentheses) for the difference. The residuals are obtained by regressing the variables on the randomization block fixed effect. *p<0.10, **p<0.05, ***p<0.01

Table 3 presents the differences in tutoring dosage for treatment and control groups. Scheduled dosage appears to be similar across the control and treatment in frequency (5 sessions per week), length (about 41.5 minutes per session), and group size (between 3 and 4 students per tutor). On average, students in the analytic sample attended 92 sessions, which amounts to approximately 3,800 minutes. Figure B1 and Table B1 show the detailed actual dosage for the treatment and control groups.

While we do not observe statistically significant differences in scheduled dosage between the two groups, Table 4 reveals some differences in tutoring characteristics. Specifically, the treatment group includes a higher proportion of teachers (69% versus 34%), and this difference is statistically significant. Conversely, the control group has a higher proportion of paraprofessional tutors. Compared to when teachers serve as tutors, when

paraprofessionals serve as tutors it is more likely that the classroom has two adults present during tutoring. As expected, the control group therefore has on average a greater number of adults in the classroom compared to the treatment group.

Table 4. Tutor Qualifications, by Group

	Randomized Sample			Final Analytic Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Variable	Control	Treatment	Model Adjusted Difference	Control	Treatment	Model Adjusted Difference
Teacher tutor	0.341	0.694	0.344***	0.342	0.704	0.344***
	(0.476)	(0.462)	(0.04)	(0.476)	(0.458)	(0.04)
Tutor years of district experience	8.624	8.016	-0.520	8.530	8.062	-0.532
	(6.518)	(7.293)	(0.64)	(6.729)	(7.366)	(0.68)
Tutor with a master's degree	0.252	0.241	-0.012	0.256	0.254	-0.010
	(0.436)	(0.429)	(0.04)	(0.439)	(0.437)	(0.04)
Number of adults in classroom	1.556	1.199	-0.328***	1.547	1.198	-0.318***
	(0.665)	(0.426)	(0.05)	(0.650)	(0.427)	(0.05)
Observations	135	193	328	117	179	296

Note: The "Randomized Sample" comprises all students who are randomly assigned to the treatment of control groups, while the "Final Analytic Sample" comprises students with non-missing fall *and* spring scores. Columns (1), (2), (4), and (5) report the means and standard deviations (in parentheses) for each variable and columns (3) and (6) present the residualized difference between the control and treatment groups for each variable along with the standard error (in parentheses) for the difference. The residuals are obtained by regressing the variables on the randomization block fixed effect. *p<0.10, **p<0.05, ***p<0.01

As a result, observed differences in literacy achievement between the treatment and control groups may be attributable to instructionally aligned tutoring as well as tutor type. Later we examine whether the impacts differ by tutor type to quantify the extent to which this difference influences the overall treatment effect.

Impact Analysis

RQ2: For students who start the year performing below the 40th percentile on aimswebPlus, what is the effect of participating in instructionally aligned tutoring on literacy outcomes in grades K-3, compared to business as usual tutoring?

Table 5 presents results for the main impact question, as measured by end-of-year aimswebPlus standardized literacy achievement scores. The first column presents the impact of instructionally aligned tutoring on literacy achievement, with the inclusion of only randomization block fixed effects (FE), the second column adds the baseline achievement control, and column three runs the full model described in Equation (1). As seen in Table 5, the effect ranges from 0.12 to 0.19 standard deviations, and is statistically significant at the 5% and 10% levels. Our preferred specification in Column (3) shows that on average, students assigned to receive instructionally aligned tutoring score 0.12 standard deviations higher than students in the control group (p<0.10). Based on average annual growth estimates from Hill et al. (2008), this effect size is equivalent to an additional 1.3 months of learning for treatment students.

Table 5. Effect of Instructionally Aligned Tutoring on Literacy Achievement

	(1) Standardized Score	(2) Standardized Score	(3) Standardized Score
Instructionally Aligned Tutoring	0.189**	0.119*	0.120*
	(0.083)	(0.063)	(0.064)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	No	Yes	Yes
Demographic Characteristics	No	No	Yes
Observations	296	296	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. *p<0.10, **p<0.05, ***p<0.01

To assess whether these impacts are consistent, we also test the impact of instructionally aligned tutoring on state standardized English Language Arts (ELA) test for third grade students.[4] Table B2 shows that we find an average impact of 0.17 standard deviations.

[4] Only third grade students in the sample take this test.

Although the estimate is not statistically significant, the impacts on the standardized ELA test are consistent with the above findings.

Exploratory Analysis:

RQ3: How do the effects of instructionally aligned high-dosage tutoring on academic achievement differ by students' prior achievement levels and gender?

Next, we examine the heterogeneity of effects by students' prior achievement and gender. Tables 6 and 7 report the estimates based on our preferred specifications, while Table B3 demonstrates the robustness of these estimates when controls are excluded.

Policymakers are always on the lookout for interventions that improve outcomes for at-risk students. To test whether the intervention had a differential effect on high versus low performing students, we divide the students into two groups based on their baseline achievement. We classify students as high (low) performing if they have baseline achievement above (below) the grade level median. The pattern of results in Table 6 suggests that instructionally aligned tutoring benefits low performing students more than high performing students, as the impact is higher and marginally significant for the high performing students. However, the difference between the two groups is not statistically significant.

Table 6. Effect of Instructionally Aligned Tutoring, by Prior Achievement

	(1) Low Performing Students	(2) High Performing Students	(3) Difference
Instructionally Aligned Tutoring	0.176*	0.065	0.105
	(0.097)	(0.104)	(0.143)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	Yes	Yes	Yes
Demographic Characteristics	Yes	Yes	Yes
Observations	143	153	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. The difference in column (3) is obtained from a regression of the literacy achievement on an interaction of treatment and low performing student subgroup and other variables in Equation (1). *p<0.10, **p<0.05, ***p<0.01

In the current study, although the intervention was not designed to favor one group over the other and both groups received tutoring, it is possible that instructional alignment has a differential gender effect. Table 7 shows that the treatment seems to have a greater impact for boys, with an effect size of 0.22 standard deviations (p<0.10). While we cannot reject the null hypothesis of no differential effect between genders, the estimated impact is higher and marginally significant for boys relative to girls.

Several plausible mechanisms may explain this pattern as existing research has identified gender-based differences in the effects of tutoring. For instance, a study conducted in Peru found that the impact of a targeted remedial education program was entirely driven by boys, despite both boys and girls having similar levels of prior achievement and comparable attendance rates (Saavedra et al., 2019). Researchers hypothesize that these differential effects may stem from preferential treatment by tutors or greater engagement by boys in small-group instructional settings. In the current context, boys might have responded more positively to aligned instruction, particularly when it included immediate feedback. Another possibility is that boys may have begun with lower baseline achievement, providing greater room for

improvement. Indeed, analysis shows that within the sample, the baseline achievement of boys was approximately 0.07 standard deviations lower than that of girls, conditional on randomization block fixed effects and demographic characteristics—though this difference was not statistically significant.

Table 7. Effect of instructionally aligned tutoring by gender

	(1) Males	(2) Females	(3) Difference
Instructionally Aligned Tutoring	0.215*	0.074	0.136
	(0.119)	(0.098)	(0.153)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	Yes	Yes	Yes
Demographic Characteristics	Yes	Yes	Yes
Observations	144	152	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. The difference in column (3) is obtained from a regression of the literacy achievement on an interaction of treatment and male student subgroup and other variables in Equation (1). *p<0.10, **p<0.05, ***p<0.01

RQ4: Do results vary by tutoring group characteristics?

Prior research suggests that the effectiveness of high-dosage tutoring can vary considerably depending on implementation characteristics such as the type of tutor (e.g., teacher vs. paraprofessional), tutor qualifications, and dosage (Nickow, Oreopoulos, & Quan, 2023; Guryan et al., 2023). In Tables 8 to 11 and the robustness checks presented in Table B4, we investigate whether the tutoring group characteristics are associated with differential treatment effects. Importantly, because both groups in our study received tutoring, we can identify the causal effect of these characteristics on student literacy outcomes. This design strengthens our ability to isolate the influence of specific tutoring features.

Given that tutoring effects often vary by tutor type, it is plausible that the effectiveness of instructionally aligned tutoring may also differ depending on who delivers it. Table 8 reports the

estimates for the differential impact of instructionally aligned tutoring by type of tutor. Although none of the estimates are statistically significant, interestingly we see some suggestive evidence that the treatment effects are higher when the tutor is a paraprofessional relative to a teacher. Insights from district leaders suggest this may be due to paraprofessionals' focused role. Unlike teachers, who juggle multiple instructional responsibilities and preparation tasks, paraprofessional tutors are dedicated solely to tutoring. Accordingly, paraprofessionals can focus their planning and preparation in ways that allow them to adhere closely to curricula expectations.[5]

Table 8. Effect of Instructionally Aligned Tutoring, by Tutor Type

	(1) Teacher	(2) Paraprofessional	(3) Difference
Instructionally Aligned Tutoring	0.009	0.214	-0.204
	(0.122)	(0.143)	(0.184)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	Yes	Yes	Yes
Demographic Characteristics	Yes	Yes	Yes
Observations	166	130	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. The difference in column (3) is obtained from a regression of the literacy achievement on an interaction of treatment and an indicator for having a teacher tutor, along with other variables in Equation (1). *p<0.10, **p<0.05, ***p<0.01

We also examine the differences in literacy achievement based on the tutor's highest degree. Students are divided into two groups: those whose tutor holds a master's degree and those whose tutor holds a degree below the master's level (i.e., bachelor's, associate's, or high school diploma). We find that the impact of instructional alignment is higher and marginally more

[5] As paraprofessional tutors are more likely to have another adult present in the classroom, we also test whether the impact of instructionally aligned tutoring is greater for treatments students that have two adults present in the classroom. While there is some evidence that having two adults in the classroom enhances the effectiveness of instructionally aligned tutoring, the estimates are not statistically significant and are sensitive to the inclusion of controls (Table B4).

significant for students whose tutors do not hold a master's degree. Specifically, the estimated treatment effect is 0.17 standard deviations (p<0.05) for students tutored by individuals without a master's degree, 0.34 standard deviations (p<0.10) greater than the effect observed for students with tutors who hold a master's degree. These findings align with the theory that tutors without advanced degrees may be more inclined to adhere to curricula expectations of aligned tutoring, whereas those with higher degrees may struggle to meet those expectations on top of other responsibilities.[6] Furthermore, there is a strong association between having a master's degree and being a teacher. Hence, this finding is consistent with the suggestive evidence that impact of instructionally aligned tutoring seems to be higher with paraprofessionals than teachers.

Table 9. Effect of Instructionally Aligned Tutoring, by Tutor Highest Degree

	(1) Master's	(2) Bachelor's or Lower	(3) Difference
Instructionally Aligned Tutoring	-0.180	0.167**	-0.342*
	(0.181)	(0.077)	(0.185)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	Yes	Yes	Yes
Demographic Characteristics	Yes	Yes	Yes
Observations	75	219	294

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. The difference in column (3) is obtained from a regression of the literacy achievement on an interaction of treatment and indicator for tutor having a master's degree, along with in Equation (1). *p<0.10, **p<0.05, ***p<0.01

Next, we examine the differences in the impacts by dosage – specifically tutoring session length and group size. While the district's minimum requirement for session length was

[6] We further disaggregate the group of students whose tutors do not hold a master's degree into two subgroups: those tutored by individuals with a bachelor's degree and those tutored by individuals with an associate's degree or high school diploma. However, we do not find statistically significant differences in the impact between these two subgroups.

40 minutes (and most students received that length), a subset of students received 45-minute sessions, primarily concentrated in one school. Table 10 presents the estimated effects of instructionally aligned tutoring by session length. We find that 45-minute treatment group sessions yielded the highest impact on literacy achievement, with an effect size of 0.31 standard deviations ($p<0.05$). The difference in treatment impact with 40-minute versus 45-minute treatment sessions is 0.25 standard deviations ($p<0.10$). It is important to note, however, that the school implementing 45-minute sessions also received additional grade-level instructional coaching and was subject to heightened district oversight. As such, we cannot conclusively attribute the observed effects solely to session length; the additional support may have contributed to the improved outcomes.

Table 10. Effect of Instructionally Aligned Tutoring, by Tutoring Session Length

	(1) 45-minute sessions	(2) 40-minute sessions	(3) Difference
Instructionally Aligned Tutoring	0.306**	0.052	0.255*
	(0.124)	(0.074)	(0.147)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	Yes	Yes	Yes
Demographic Characteristics	Yes	Yes	Yes
Observations	88	208	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. The difference in column (3) is obtained from a regression of the literacy achievement on an interaction of treatment and an indicator for 45-minute tutoring sessions, along with other variables in Equation (1). *p<0.10, **p<0.05, ***p<0.01

Recent research consistently finds that smaller tutoring group sizes are associated with stronger academic outcomes (Robinson, et al., 2024). However, group sizes of three or more students appear to perform particularly well in grades 2–5 (Nickow, Oreopoulos, & Quan, 2023). Table 11 presents estimates of the effect of instructionally aligned tutoring by group size, distinguishing between groups of three or fewer students and those with more than three (up to six). The effect of instructionally aligned tutoring is higher and marginally significant for tutoring groups with more than three students. Notably, the impact is greatest at 0.49 standard deviations *($p < 0.01$)* for groups of four students (Table B4).

The implementation of the aligned curriculum included a peer interaction component, which likely contributed to the stronger outcomes observed in larger groups. Additionally, larger group sizes may have encouraged tutors to adhere more closely to the instructional materials, strategies, and pacing, rather than spending more or less time than recommended on specific content areas, or improvising in other ways that undermine effectiveness of tutoring. While one-on-one tutoring is often considered the gold standard, it is also more costly. The observed effectiveness of slightly larger groups suggests a promising avenue for reducing costs without

compromising instructional quality. This has important implications for scaling instructionally aligned tutoring.

Table 11. Effect of Instructionally Aligned Tutoring, by Tutoring Group Size

	(1) Three or less students	(2) More than three students	(3) Difference
Instructionally Aligned Tutoring	0.040	0.188*	-0.141
	(0.111)	(0.102)	(0.149)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	Yes	Yes	Yes
Demographic Characteristics	Yes	Yes	Yes
Observations	133	163	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. The difference in column (3) is obtained from a regression of the literacy achievement on an interaction of treatment and an indicator for tutoring group comprising three or less students, along with other variables in Equation (1). *p<0.10, **p<0.05, ***p<0.01

Discussion

This study investigates the effect of instructionally aligned tutoring on the literacy outcomes of students in kindergarten to grade three using a randomized controlled trial. District leaders were interested in instructionally aligned tutoring as a way of providing more coherent instructional experiences for both students and educators (State Collaborative on Reforming Education, 2023). Students struggling to read had opportunities to practice with familiar routines and materials and were provided additional support when engaging with foundational texts and writing tasks they needed to achieve grade-level reading success. Teachers provided additional support using materials and data to monitor and guide student learning. While educators have long considered instructional coherence and alignment best practice, the current study adds to the empirical evidence by providing an estimate of the causal impact of tutoring coherence with core instruction on early literacy.

We find that students assigned to receive instructionally aligned tutoring have slightly more improvement in literacy over the course of the school year, relative to those being tutored using strategies and materials that are not aligned to core instruction. We also find that boys appear to benefit slightly more than girls from the instructional alignment, suggesting instructional alignment may mitigate some of the gender gaps in early literacy identified in previous research. Additionally, students with lower fall performance appear to benefit more from instructional alignment and coherence. This pattern suggests instructional alignment may be most important for students who might be most likely to receive tutoring. In this study, tutoring occurred in-school, supplanting rather than supplementing classroom instruction. Future research could explore whether alignment also bolsters effectiveness of tutoring that takes place outside the school day, supplementing classroom instruction.

We do not find any statistically significant differences in the impacts by tutor type (teacher or paraprofessional). From past research, we might have expected tutors with stronger credentials to generate larger gains in literacy achievement. If anything, we observe larger effects for tutors that do not hold a master's degree, relative to those who have one. The influence of tutor qualifications and degree may have been mitigated by training and ongoing support from school leaders and instructional coaches. This speaks to the district's role in promoting not just instructional alignment but also instructional coherence, by allocating resources to support effective implementation. This also has important implications for scaling tutoring interventions cost effectively without compromising quality.

We caution that the study design does *not* allow us to say whether one type of curriculum is more effective than another. Rather, our interpretation is that using an aligned curriculum appears more effective than using a different curriculum during tutoring, in a context where all curricula in the study were considered high-quality by the state.

Limitations

Constraints in the data available presented a few challenges for this project. First, since state and district policies focus on the RTI tiers, we would prefer to present findings by these categorizations. While the study context allowed for a multi-site randomized controlled trial, only a subset of students in kindergarten through third grade were eligible for tutoring. The number of students in the different RTI tiers was small, and analyzing the data by tier sacrificed statistical power to detect effects. As a result, we created broader (and less policy relevant) categories of high and low achievement.

Additionally, while district and school leaders were interested in the effects on specific early literacy skills, teacher discretion in administering and scoring skills assessments led to inconsistent and potentially unreliable data. Consequently, we used the standardized interim assessment for the study. While the standardized data are less actionable for school leaders and classroom teachers, they provide more valid results.

We would like to have examined tutoring dosage using a measure of student attendance at each session. The state in which this study took place had established a platform for documenting such information. However, we found that the data were incomplete in the year in which the student took place. Consequently, as discussed before, we had to rely on a proxy for student attendance.

While this study shed light on the effectiveness of instructionally aligned tutoring, we do not have data to examine the cost of this intervention. Understanding cost is essential for evaluating feasibility, scalability, and cost-effectiveness relative to alternative strategies. Future research could explore intervention costs, including whether a less resource-intensive version of instructionally aligned tutoring yields similar results.

Implications for Policy

This study provides evidence that instructionally coherent materials and strategies promote student learning during early grade tutoring sessions. Districts' procurement of instructional material; instructional coaches and school leaders' support for alignment in strategies used; and educators' practices in the classroom can all advance student learning through instructional alignment. The district where this study took place is working to expand the instructionally aligned tutoring. As part of this scaling effort, district leaders have tapped school leaders who participated in the pilot to share what they have learned. This allows students across the district to benefit from the experiences of leaders who created the working conditions necessary to implement instructionally aligned tutoring effectively.

We recognize that disparities in literacy are perpetuated throughout students' lives by a variety of societal inequities. As such, in-school solutions are just a starting point. A variety of actors at different levels of the system have roles to play in addressing inequitable literacy outcomes. For example, greater support for in-home literacy activities and stronger state policies to screen for and address challenges such as dyslexia could prove valuable.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics*, *138*(1), 1–35. https://doi.org/10.1093/qje/qjac045

Accelerate. (2025, August 28). *Defining tutoring dosage for program implementation and applied research*. Retrieved from https://accelerate.us/research/defining-tutoring-dosage-for-program-implementation-and-applied-research/

Allensworth, E. & Schwartz, N. (2020). *School Practices to Address Student Learning Loss*. EdResearch for Recovery Project, Brief No. 1. Annenberg Institute at Brown University.

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory Into Practice, 41*(4), 255-260.

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). *Evaluation of Response to Intervention Practices for Elementary School Reading* (NCEE 2016-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bhatt, M. P., Chau, T., Condliffe, B., Davis, R., Grossman, J., Guryan, J., Ludwig, J., Magnaricotte, M., Mattera, S., Momeni, F., Oreopolous, P. &S toddard, G. (2025). *Initiative Interim Report: Findings from 2023-24*. University of Chicago Education Lab. Retrieved from https://educationlab.uchicago.edu/resources/personalized-learning-initiative-interim-report-findings-from-2023-24/

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347-364.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. University of Chicago Press.

Carbonari, M. V., Davison, M., DeArmond, M., Dewey, D., Dizon-Ross, E., Goldhaber, D., Hashim, A. K., Kane, T. J., McEachin, A., Muroga, A., Morton, E., Patterson, T., & Staiger, D. O. (2024). *The Impact and Implementation of Academic Interventions During COVID: Evidence from the Road to Recovery Project.* CALDER Working Paper No. 275-0624-2.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology, 98*(3), 489-507.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471-533.

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*(4), 605-619.

Fitzgerald, J. (2001). Can minimally trained college student volunteers help young at-risk children to read better? *Reading Research Quarterly, 36*(1), 28-46.

Gaab, N. & Petscher, Y. (2022). Screening for Early Literacy Milestones and Reading Disabilities: The Why, When, Whom, How, and Where. *Perspectives on Language and Literacy*, Winter.

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M. V., Dodge, K., Farkas, G., Fryer, R. C., et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, 113(3), 738–765. https://www.nber.org/papers/w28531

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.

Kohlmoos, L., & Steinberg, M. P. (2024). *Contextualizing the Impact of Tutoring on Student Learning: Efficiency, Cost Effectiveness, and the Known Unknowns*. Accelerate. Retrieved from https://accelerate.us/wp-content/uploads/2024/05/Accelerate-Research-Report-Efficiency-and-Cost-Effectiveness-1.pdf

Kraft, M. A., Schueler, B. E., & Falken, G. (2024). *What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability*. (EdWorkingPaper: 24-1031). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/zygj-m525

Lynch, J. (2002). Parents' self-efficacy beliefs, parents' gender, children's reader self-perceptions, reading achievement and gender. *Journal of Research in Reading, 25*(1), 54-67.

Markovitz, C. E., Hernandez, M. W., Hedberg, E. C., & Whitmore, H. W. (2022). Evaluating the effectiveness of a volunteer one-on-one tutoring model for early elementary reading intervention: A randomized controlled trial replication study. *American Educational Research Journal*, *59*(4), 788-819. https://doi.org/10.3102/00028312211066848

McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., Quiroga, T., & Gray, A. L. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities, 35*(1), 69-86.

McKenna, M. C., Kear, D. J., & Ellsworth, R. A. (1995). Children's attitudes toward reading: A national survey. *Reading Research Quarterly, 30*(4), 934-956.

National Assessment of Educational Progress (2022). *1992–2022 Reading Assessments.* Retrieved from: https://www.nationsreportcard.gov/highlights/reading/2022/

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development.

National Student Support Accelerator (n.d.). *District Playbook: High Impact Tutoring.* https://nssa.stanford.edu/district-playbook-high-impact-tutoring/section-6/scheduling-sessions

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis*, *23*(4), 297–321. http://www.jstor.org/stable/3594132

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The promise of tutoring for preK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, *61*(1), 74-107. https://doi.org/10.3102/00028312231208687

Pearson (2017). *aimswebPlus Technical Manual*. https://resources.finalsite.net/images/v1744836724/marshfieldschoolsorg/lv1cgfxv3inkz62zl8ax/PlusTechnicalManual.pdf

Pearson (n.d.). aimswebPlus Assessment Matrix. Retrieved from https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/aimsweb/awp-assessment-matrix-update-us-ca.pdf

Pressley, M., Wharton-McDonald, R., Allington, R., Block, C. C., Morrow, L., Tracey, D., Baker, K., Brooks, G., Cronin, J., Nelson, E., & Woo, D. (2001). A study of effective first-grade literacy instruction. *Scientific Studies of Reading, 5*(1), 35-58.

Robinson, C. D., Kraft, M. A., Loeb, S., & Schueler, B. E. (2021). *Accelerating student learning with high-dosage tutoring*. EdResearch for Recovery Design Principles Series. Annenberg Institute for School Reform at Brown University. ERIC Number: ED613847. Retrieved from https://files.eric.ed.gov/fulltext/ED613847.pdf

Robinson, C. D., & Loeb, S. (2021). *High-Impact Tutoring: State of the Research and Priorities for Future Learning*. (EdWorkingPaper: 21-384). Annenberg Institute at Brown University.

Saavedra, J. E., Näslund-Hadley, E., & Alfonso, M. (2019). Remedial inquiry-based science education: Experimental evidence from Peru. *Educational Evaluation and Policy Analysis*, *41*(4), 483-509. https://doi.org/10.3102/0162373719867081 (Original work published 2019)

State Collaborative on Reforming Education (2023). *Early Literacy Success for All Students: A Coherent Path Forward* https://tnscore.org/resources/early-literacy-success-for-all-students-a-coherent-path-forward

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly, 28*(2), 178-200.

Appendix A: Randomization

Table A.1: Number of students randomized by school and grade

School	Grade	Treatment	Control	Total
School A	Kindergarten	15	15	30
	First	19	7	26
	Second	15	6	21
	Third	14	13	27
School B	Kindergarten	13	4	17
	First	18	13	31
	Second	21	16	37
	Third	8	6	14
School C	Kindergarten	8	9	17
	First	8	6	14
	Second	6	6	12
	Third	4	4	8
School D	Kindergarten	9	10	19
	First	15	12	27
	Second	13	7	20
	Third	5	3	8
Total		135	193	328

Appendix B: Results

Figure B1. Density Plot of the Percentage of Scheduled Sessions Attended, by Treatment Group Status

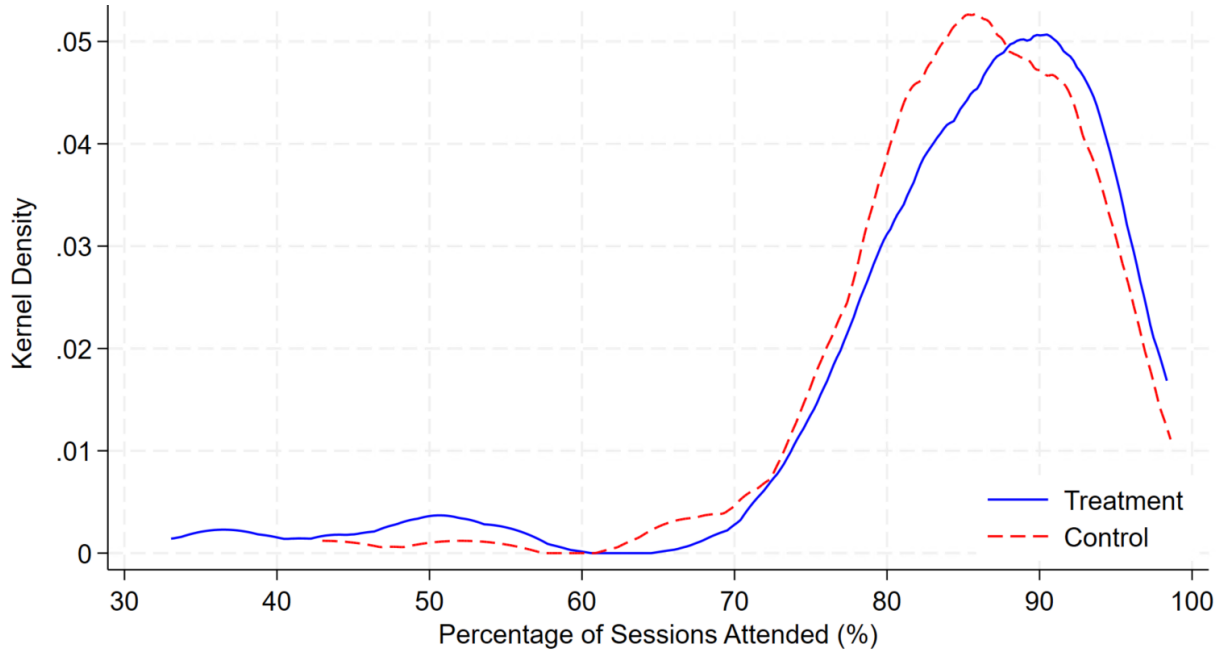


Table B1. Percentage of Students in Attendance Bins, by Treatment Group Status

Attendance	100%	99%-80%	79%-50%	Less than 50%
Dosage (Percentage of Students)				
Treatment	0%	77.08%	16.67%	6.25%
Control	0%	75.56%	18.52%	18.52%

Notes: Dosage is calculated using student and tutor school attendance, tutoring exit date (if applicable), and days of tutoring

Table B2. Effect of Instructionally Aligned Tutoring on English Language Arts for Students in Grade 3

	(1) Standardized Score	(2) Standardized Score	(3) Standardized Score
Treatment	0.162	0.154	0.172
	(0.178)	(0.166)	(0.191)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	No	Yes	Yes
Demographic Characteristics	No	No	Yes
Observations	296	296	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1,2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. *p<0.10, **p<0.05, ***p<0.01

Table B3. Effect of Instructionally Aligned Tutoring on Literacy Achievement, by Student Prior Achievement and Gender

	(1) Standardized Score	(2) Standardized Score	(3) Standardized Score
Impact by Prior Achievement			
Treatment	0.218**	0.169*	0.176*
	(0.109)	(0.096)	(0.097)
Treatment *Achievement Above Median	-0.148	-0.133	-0.105
	(0.155)	(0.139)	(0.143)
Impact by Gender			
Treatment	0.256*	0.208*	0.201*
	(0.143)	(0.114)	(0.118)
Treatment *Female	-0.098	-0.122	-0.136
	(0.189)	(0.148)	(0.153)
Randomization Block FE	Yes	Yes	Yes
Baseline Achievement	No	Yes	Yes
Demographic Characteristics	No	No	Yes
Observations	296	296	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. Estimates are obtained from a regression of the literacy achievement on an interaction of treatment and an indicator for subgroup of interest. *p<0.10, **p<0.05, ***p<0.01

Table B4. Effect of Instructionally Aligned Tutoring on Literacy Achievement, by Tutoring Group Characteristics

	(1) Standardized Score	(2) Standardized Score	(3) Standardized Score
Impact by Tutor Type			
Treatment	0.502***	0.228*	0.214
	(0.159)	(0.132)	(0.138)
Treatment *Teacher	-0.395*	-0.210	-0.204
	(0.218)	(0.174)	(0.184)
Impact by Tutor Highest Degree			
Treatment	0.264***	0.176**	0.167**
	(0.101)	(0.078)	(0.078)
Treatment *Master's	-0.272	-0.247	-0.342*
	(0.219)	(0.158)	(0.185)
Impact by Number of Adults in the Classroom			
Treatment	0.700***	0.229	0.220
	(0.247)	(0.171)	(0.168)
Treatment *One Adult	-0.614**	-0.126	-0.123
	(0.276)	(0.196)	(0.195)
Impact by Tutoring Session Length			
Treatment	0.080	0.055	0.049
	(0.092)	(0.073)	(0.075)
Treatment *45 Minute Session	0.383**	0.224	0.255*
	(0.194)	(0.148)	(0.147)
Impact by Tutoring Group Size			
Treatment	0.233*	0.175*	0.188*
	(0.126)	(0.100)	(0.104)
Treatment *Three or Less Students	-0.120	-0.128	-0.141
	(0.184)	(0.148)	(0.149)
Randomization Block FE	Yes	Yes	Yes

Baseline Achievement	No	Yes	Yes
Demographic Characteristics	No	No	Yes
Observations	296	296	296

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. Estimates are obtained from a regression of the literacy achievement on an interaction of treatment and an indicator for subgroup of interest. *p<0.10, **p<0.05, ***p<0.01

Table B5. Effect of Instructionally Aligned Tutoring, by Tutoring Group Size

	(1)	(2)	(3)	(4)
Group Size	2 students	3 students	4 students	5 students
Instructionally Aligned Tutoring	0.230	-0.174	0.486***	0.092
	(0.379)	(0.207)	(0.154)	(0.147)
Randomization Block FE	Yes	Yes	Yes	Yes
Baseline Achievement	Yes	Yes	Yes	Yes
Demographic Characteristics	Yes	Yes	Yes	Yes
Observations	39	85	94	64

Notes: Baseline Achievement is standardized fall ORF scores for students in grades 1, 2, and 3 and fall EL for students in kindergarten. Demographic characteristics include gender and race. *p<0.10, **p<0.05, ***p<0.01

Appendix C: Open Science Framework Pre-Registration

Hypotheses

RQ1 (impact): students in grades K-3 who are assigned to the instructionally aligned high-dosage tutoring groups will outperform control group students on literacy assessments. RQ2 (impact/mediator): the effects of instructionally aligned high-dosage tutoring will be greater for students with lower prior academic achievement. RQ3 (impact/mediator): the effects of instructionally aligned high-dosage tutoring will differ depending on whether tutoring focused on foundational reading skills vs. knowledge building competencies (non-directional hypothesis). RQ4 (descriptive): tutoring groups will vary in dosage and the qualifications of tutors. Students in Response to Intervention Tier (RTI) III are expected to have higher dosage (smaller group size, more time in tutoring) than students in RTI Tier II. Students in RTI II and III are expected to have more qualified tutors than students in RTI I who are eligible for tutoring.

Design Plan

Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials. No blinding is involved in this study.

Study design

This study uses a randomized controlled trial (RCT) to address the impact research questions. Students are randomly assigned to treatment or control. The students are stratified by school*grade*skill group and randomized from within those strata.

Sampling Plan

Explanation of existing data

We plan to use prior test scores and student demographics as covariates in our models, but the analysts do not yet have access to these data.

Data collection procedures

Abt will obtain data on the Quick Phonics Assessment directly from the schools; these data will be used in determining student needs for tutoring group formation (expected to be available by the end of August 2024). Schools will also provide: - Tutor qualifications (by October 2024) - Tutoring dosage (quarterly throughout SY24-25) The district will provide the following data: - aimswebPlus (beginning of year expected to be available by mid-September 2024; mid-year by December 2024, and end of year by June 2025) - Student demographics (by October 2024) - English Language Arts scores from Tennessee Comprehensive Assessment Program (expected to be available by July 2025).

Sample size

Below are the estimated numbers of students enrolled in each grade eligible for the study for the participating schools. The partners believe that approximately half the students will meet the eligibility criteria to participate in the study (approximately 518 students).

Sample size rationale

The minimum detectable effect size (with student-level randomization and 500 students) is 0.149.

Variables

Manipulated variables

The variable manipulated is the instructional materials used during tutoring. Students in the treatment group receive instructionally aligned (Benchmark) materials. Students in the control group who are in RTI Tiers II and III receive tutoring with other materials. Students in the control group who are in RTI I but are eligible for tutoring receive standard classroom instruction in the control condition.

Measured variables

Outcome measures: Average growth in composite score from the universal literacy screener (aimswebPlus) from the beginning to the middle to the end of the 2024-25 school year. Average growth in domain specific scores from the universal literacy screener (aimswebPlus) from the beginning to the middle to the end of the 2024-25 school year. 3rd grade ELA TCAP scores and proficiency rates for students served in the varied small group instructional settings. Covariates: Student grade (series of dummy indicators) Gender (binary dummy indicator) Race/ethnicity (dummy indicators for Black and Latino) Free/reduced price lunch status (binary dummy indicator) IEP status (binary dummy indicator)

Analysis Plan

Statistical models

We will use a linear regression model, in which the effect of treatment on an outcome measure is estimated controlling for a pretest measure of the outcome and other baseline student covariates. Dummy variables representing stratification blocks are included in the model. We will test for interactions between the treatment and students' prior achievement. We will also test for interactions between the treatment and the focus of the tutoring (foundational skills vs. knowledge building).

Transformations

We do not plan on transforming or centering the data.

Inference criteria

We will be using a *p*-value of 0.05 and two-tailed tests for the analyses.

Data exclusion

Students eligible for tutoring in RTI I, II, or III are eligible for the study. This excludes English language learners and some students with Individualized Education Plans that receive other support during the intervention block. Outliers will be included in the analysis.

Missing data

Students who do not complete both pre- and post-assessment will not be included in the analysis.

Exploratory analysis

RQ1 is the only impact analysis. RQs 2 and 3 are exploratory. For RQ2, we will examine whether the effects of instructionally aligned high-dosage tutoring on academic achievement differ by students' beginning-of-year performance on the universal literacy screener (aimswebPlus): the 26-40th percentile, the 11-25th percentile (RTI Tier 2), and the 1-10th percentile (RTI Tier 3). For RQ3, we will examine whether the effects of instructionally aligned high-dosage tutoring on academic achievement differ by tutoring focus: foundational reading skills, knowledge building, or both.