Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1548

Early literacy tutoring and children's reading trajectories: Evidence from a four-year randomized controlled trial

Jilli Jung
Stanford University

Carly D. Robinson
Stanford University

Tye Ripma
Stanford University

Kalena E. Cortes
Texas A&M University

Karen Kortecamp
George Washington University

Susanna Loeb
Stanford University

Most evidence on early literacy tutoring comes from intensive one-year interventions. Much less is known about whether schools can sustain individualized instructional support as students' reading needs evolve across the early grades. We present findings from a four-year randomized controlled trial evaluating Chapter One, an adaptive instructional support system designed to translate the instructional principles underlying effective tutoring into a scalable model for early literacy. We randomly assigned 815 kindergarten students in a large urban district in the Southeast to Chapter One or business-as-usual instruction and followed them through 3rd grade. Students assigned to Chapter One scored 0.12 to 0.17 standard deviations higher on state-administered reading assessments across grades 1 through 3, although the third-grade estimate is imprecise. They were also less likely to have an individualized education program (IEP) or be retained by the end of the study. The reduction in IEP status was concentrated among students who already had an IEP at baseline. The estimated longitudinal effects on reading are smaller than those typically reported for intensive tutoring models but were achieved under ordinary school staffing, scheduling, budget, and student mobility constraints. The findings suggest that schools can translate the instructional principles underlying effective tutoring into adaptive instructional support systems that improve students' reading trajectories and academic progression while remaining feasible to sustain across the learning-to-read years.

VERSION: August 2026

Suggested citation: Jung, Jilli, Carly D. Robinson, Tye Ripma, Kalena E. Cortes, Karen Kortecamp, and Susanna Loeb. (2026). Early literacy tutoring and children's reading trajectories: Evidence from a four-year randomized controlled trial. (EdWorkingPaper: 26-1548). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/g7sg-m495

Early literacy tutoring and children's reading trajectories: Evidence from a four-year randomized controlled trial

Jilli Jung, Carly D. Robinson, Tye Ripma, Kalena Cortes, Karen Kortecamp & Susanna Loeb

Abstract

Most evidence on early literacy tutoring comes from intensive one-year interventions. Much less is known about whether schools can sustain individualized instructional support as students' reading needs evolve across the early grades. We present findings from a four-year randomized controlled trial evaluating Chapter One, an adaptive instructional support system designed to translate the instructional principles underlying effective tutoring into a scalable model for early literacy. We randomly assigned 815 kindergarten students in a large urban district in the Southeast to Chapter One or business-as-usual instruction and followed them through 3rd grade. Students assigned to Chapter One scored 0.12 to 0.17 standard deviations higher on state-administered reading assessments across grades 1 through 3, although the third-grade estimate is imprecise. They were also less likely to have an individualized education program (IEP) or be retained by the end of the study. The reduction in IEP status was concentrated among students who already had an IEP at baseline. The estimated longitudinal effects on reading are smaller than those typically reported for intensive tutoring models but were achieved under ordinary school staffing, scheduling, budget, and student mobility constraints. The findings suggest that schools can translate the instructional principles underlying effective tutoring into adaptive instructional support systems that improve students' reading trajectories and academic progression while remaining feasible to sustain across the learning-to-read years.

Introduction

Developing strong reading skills in the early grades is central to students' long-term academic success, yet many students struggle to do so: fewer than one-third of U.S. fourth graders read at a proficient level (National Center for Education Statistics, 2024). States and districts have increasingly adopted comprehensive early literacy policies centered on universal screening, evidence-based reading instruction, and targeted intervention for students who fall behind (ExcelinEd, 2026). These efforts aim to identify struggling readers early and provide support before initial gaps in foundational reading skills become persistent barriers to later learning.

Schools face the challenge of providing instructional support that evolves as children's reading develops. Identifying struggling readers is only the first step. Students who struggle with letter-sound correspondence in kindergarten require different instruction than students struggling with fluency, vocabulary, or comprehension in later grades. Effective early literacy support therefore requires more than a one-time intervention. It requires a system that continually adjusts instruction as students' needs change while remaining feasible for schools to sustain over multiple years. This view is consistent with broader work on progress monitoring and data-based individualization, which emphasizes using ongoing evidence about student learning to continually adapt instruction to students' evolving needs (Jung et al., 2018; Stecker et al., 2008).

Research on intensive tutoring provides some of the strongest evidence we have about how individualized instruction can accelerate learning. Decades of research show that tutoring is most effective when instruction is individualized, responsive to students' progress, relationship-based, and delivered with sufficient frequency to address current learning needs (Nickow et al., 2024; Robinson & Loeb, 2021). An important unanswered question is whether schools can embed the principles that make intensive tutoring effective into instructional systems that are feasible to sustain over multiple years under real operating conditions.

Chapter One was designed to address that question. Chapter One uses tutoring as one component of an adaptive literacy support system that adjusts instructional content and support intensity as students' reading develops. Students who reach grade-level expectations transition to lighter maintenance support, while instructional resources are concentrated on students who continue to require individualized assistance. In this way, the program seeks to preserve the core features that make intensive tutoring effective while organizing them into a model that schools can sustain across the learning-to-read years.

The existing evidence is limited on whether adaptive literacy support systems can improve students' reading trajectories over multiple years. Most evaluations examine interventions delivered over only a semester or a single school year (Nickow et al., 2024), leaving little evidence about whether schools can sustain adaptive instructional support as children's literacy needs evolve. Only a small number of studies have followed students across multiple years (e.g., Bonesrønning et al., 2022; Finseraas et al., 2024), and few have focused on early literacy. Moreover, because outcomes are typically measured shortly after interventions end, little is known about whether sustained instructional support can alter children's longer-term reading trajectories (Bailey et al., 2020).

In this paper, we report findings from a four-year randomized controlled trial evaluating Chapter One in a large urban district in the Southeast. During the 2021-22 school year, we randomly assigned 815 kindergarten students either to be offered access to Chapter One or to receive business-as-usual instruction and followed them through the end of 3rd grade. The program combines part-time tutors, brief classroom-embedded instructional sessions, adaptive allocation of individualized support, and technology-supported progress monitoring to provide sustained literacy support within ordinary school routines at an annual cost of approximately $416 per student.

Students offered Chapter One scored 0.12 to 0.17 standard deviations higher on state-administered reading assessments across grades one through three, although the third-grade estimate is imprecise. Using typical annual growth in reading as a benchmark (Hill et al., 2008), our estimates translate to about 1.55 months of learning in 1st grade, 2.83 months in 2nd, and 3.33 months in 3rd. Students in the treatment group were also 4.7 percentage points less likely to have an individualized education program (IEP) by the end of 3rd grade and 2.7 percentage points less likely to be retained, although the retention estimate is imprecise. These findings suggest that schools can translate the principles that make intensive tutoring effective into adaptive instructional support systems that improve children's reading trajectories while remaining feasible to sustain across the early elementary grades.

Background

Early Reading Development Requires Sustained Adaptive Support

Early reading proficiency is foundational to students' long-term academic success. Students who struggle with reading in the early grades often continue to face academic difficulties in later grades (Cunningham & Stanovich, 1997; Duncan et al., 2007; Juel, 1988). Yet low reading proficiency is both widespread and persistent. On the 2024 National Assessment of Educational Progress, only 31 percent of U.S. fourth graders performed at or above NAEP Proficient in reading, while 40 percent performed below NAEP Basic (National Center for Education Statistics, 2024). Fourth-grade reading achievement has now returned to pre-2003 levels, extending a decline that began after progress stalled in the mid-2010s (Dewey et al., 2026). Over the past decade, states have responded to this challenge by adopting comprehensive early literacy policies, including evidence-based reading instruction, universal screening to identify struggling readers early, and grade retention requirements (ExcelinEd, 2026). Growing evidence shows such policies can improve early reading achievement, though primarily where they are

comprehensive (Spencer, 2024; Westall & Cummings, 2023). Even so, the persistence of low reading performance nationally underscores the need for instructional systems that provide individualized support throughout the learning-to-read years, adapting as students' literacy needs evolve.

The learning-to-read period, spanning kindergarten through 3rd grade, requires support that is both continuous and responsive to students' changing needs. Over these years, students move through a sequence of increasingly complex reading demands, from early emphasis on phonological awareness, letter knowledge, and decoding to growing demands on fluency, vocabulary, and comprehension. As the task of reading changes, so do the instructional supports students need (Language and Reading Research Consortium, 2015). This developmental progression implies that effective early literacy support is unlikely to take the form of a one-time remediation of an initial skill deficit; students may encounter new instructional challenges as reading demands change across grades. Effective intervention therefore requires ongoing progress monitoring, in which educators use repeated data on student learning to adjust instruction (Jung et al., 2018; Stecker et al., 2008). Schools must sustain instructional support as children's literacy needs evolve. Students who initially struggle with decoding may later require support with fluency or comprehension, while students who catch up may require progressively less individualized instruction. Effective literacy systems therefore need to adjust both instructional content and instructional intensity over time.

The policy importance of early literacy extends beyond test scores, because low reading proficiency contributes to consequential and potentially avoidable educational decisions. Because retention decisions rely heavily on academic performance, students with low reading proficiency face elevated risk of grade retention. Early-grade retention can produce short-term achievement gains (Hwang & Koedel, 2025; Schwerdt et al., 2017), but these gains often fade, and a growing body of work links retention to adverse longer-run outcomes, including lower

high school graduation, reduced adult earnings, and increased behavioral and criminal involvement (Jacob & Lefgren, 2009; Zhong, forthcoming).

Low reading proficiency is also a primary consideration in special education identification (Fletcher et al., 2018). Under the Individuals with Disabilities Education Act (IDEA) regulations, students may not be determined eligible for special education if the determinant factor is lack of appropriate instruction. Tiered, evidence-based instruction is therefore intended to ensure that students receive adequate instruction before being evaluated and to help distinguish students with learning disabilities from those who have not yet had access to rigorous, research-validated instruction (Fuchs et al., 2014). Although instructional adequacy is most often discussed in relation to initial identification for special education, it may also matter after students have been identified. If some students receive IEPs in part because they have not had access to sufficiently effective early reading instruction, then sustained academic intervention may help them make enough progress to exit special education. This possibility is of interest because students who exit special education in the early grades demonstrate higher academic achievement and lower rates of disciplinary involvement after transitioning back to general education (Curran et al., 2021; Gilmour et al., 2023).

High-Impact Tutoring Identifies Principles of Effective Individualized Instruction

Research on intensive tutoring provides some of the strongest evidence available about the conditions under which individualized instruction improves student learning. A large body of experimental evidence shows that tutoring can substantially improve academic achievement, with particularly strong effects for programs delivered in the early grades, during the school day, and through frequent one-on-one or small-group instruction (Nickow et al., 2024). This literature identifies several features of effective individualized instruction that consistently improve student learning. Effective tutoring increases instructional time, tailors instruction to students' current skill levels, provides frequent feedback, fosters sustained instructional relationships that

support engagement and motivation, and adapts instruction in response to students' progress (Nickow et al., 2024; Robinson & Loeb, 2021). These findings suggest that tutoring is effective because it embodies a set of instructional principles that can inform the design of broader instructional systems. These features are especially valuable during the learning-to-read years when students require explicit instruction in phonological awareness, phonics, and fluency together with instruction that responds to their developing skills.

Although this literature identifies the ingredients of effective individualized instruction, it leaves open a broader question: how should schools organize instructional support over time? Most tutoring studies evaluate interventions lasting a semester or a single school year, with outcomes measured shortly after the intervention ends (Nickow et al., 2024). As a result, we know relatively little about how schools can sustain individualized instructional support as children's literacy needs evolve across grades. Multi-year tutoring studies have begun to emerge, primarily in mathematics (Bonesrønning et al., 2022), but remain rare, particularly in early literacy. Moreover, even when tutoring produces substantial short-run gains, the effects of educational interventions often diminish after support ends (Bailey et al., 2017). The only long-term follow-up of a multi-year tutoring intervention found that approximately 60 percent of the mathematics achievement gains persisted 3.5 years after the program concluded (Finseraas et al., 2024). Comparable evidence for early literacy is lacking.

A second challenge concerns implementation. The tutoring programs that produce the largest achievement gains typically require frequent instructional sessions, low student-to-tutor ratios, trained staff, and dedicated instructional time, all of which are difficult for school systems to sustain over multiple years. Programs that substantially reduce these features by using existing school staff still face implementation challenges, generally producing smaller or null effects on average (Huffaker et al., 2025). Consequently, recent research has increasingly focused on how schools can implement individualized instructional support equitably and cost-effectively under

real-world constraints (Cortes et al., 2025; Kortecamp & Peters, 2024; Kraft & Lovison, 2025; Robinson et al., 2025, 2026). The next challenge is designing instructional systems that preserve the core principles underlying tutoring's effectiveness while remaining feasible to sustain within ordinary school routines.

Chapter One was designed as one approach to this challenge. Chapter One uses tutoring as a flexible instructional resource within an adaptive instructional support system that adjusts instructional content and support intensity as students' reading develops. Individualized instruction changes as students' reading develops, allowing instructional content and support intensity to evolve while concentrating tutoring resources on students who continue to need them. We evaluate whether this approach can translate the principles that make intensive tutoring effective into a model of sustained instructional support that schools can maintain throughout the learning-to-read years.

Treatment Description: Adaptive Instructional Support System

Chapter One was designed to provide individualized reading support that evolves as students' literacy needs change within and across the learning-to-read years while remaining feasible for schools to sustain. The program uses tutoring as one component of an adaptive instructional support system that adjusts instructional content and support intensity according to students' reading development. A distinguishing feature of the model is short burst tutoring. Part-time Early Literacy Interventionists (ELIs, i.e., tutors) provide individualized instruction during the school day through brief classroom-embedded sessions, typically lasting five to ten minutes, in the classroom or an adjacent space. This structure allows tutors to work with multiple students across classrooms while maintaining regular interactions with students and teachers that support instructional coordination and relationship building (Cortes et al., 2025; Robinson & Loeb, 2021).

The instructional core is a decoding curriculum delivered through these sessions, averaging 2.5 sessions and 15 minutes per week, supplemented by about 10 minutes of daily practice on the student platform. Placement follows a student's reading level rather than their grade. The sequence runs from letter names and sounds through digraphs, vowel patterns, and connected text, ending at a middle-of-second-grade reading level, and students progress through it continuously across grades. A student exits full-dosage decoding instruction on reaching what we call the benchmark: end-of-grade expectations in kindergarten and first grade, or completion of the decoding curriculum in second and third grade.

A distinguishing design feature is that Chapter One is not a fixed tutoring intervention delivered uniformly across four years. Instead, it continuously reallocates individualized instructional support as students' reading develops. When students reach a benchmark, the time released is redistributed to classmates still learning to decode, and what those students receive next depends on where the benchmark falls. In kindergarten and first grade, meeting end-of-grade expectations still leaves students within the decoding curriculum, so they move to weekly maintenance sessions that confirm they are holding their gains. From second grade on, reaching benchmark means completing the curriculum, and support shifts from building decoding to sustaining it: supported reading with grade-appropriate chapter books, alongside a small-group book club in second grade and paired with daily independent reading and a growing emphasis on fluency in 3rd grade. Table 1 summarizes this allocation. Technology supports lesson sequencing, progress monitoring, and independent practice throughout, while individualized instruction remains human delivered.

The program's staffing and delivery model is central to its policy relevance. Chapter One relies on trained part-time tutors supported by a structured curriculum and technology designed to promote consistent implementation across tutors and classrooms. Its short burst tutoring model helps reduce a practical barrier to sustained implementation by minimizing the disruption

associated with removing students from regular classroom instruction for extended periods, making the model more feasible to sustain across multiple years. Cost provides a second dimension of sustainability. At the time of the study, the program costs school districts approximately $375 per student annually in kindergarten and 1st grade, $450 in 2nd grade, and $465 in 3rd grade, substantially below the cost of many intensive tutoring models. For example, early literacy tutoring programs implemented in Massachusetts during the same academic year cost between $925 and $1,909 per student. These lower costs are particularly important for instructional systems intended to support students across multiple years. Programs that schools cannot sustain are unlikely to provide continuous individualized support as students' instructional needs evolve. Chapter One therefore provides an opportunity to examine whether an adaptive instructional support system built on the principles of intensive tutoring can improve students' reading development and academic progression while remaining feasible under real-world staffing, scheduling, and budget constraints.

Method

Study design and sample

This study follows the cohort from the preregistered randomized controlled trial reported in Cortes et al. (2025) from the beginning of kindergarten through the end of 3rd grade (see www.socialscienceregistry.org/trials/10810/). Cortes et al. (2025) evaluated the program's first-year effects on proximal, program-aligned kindergarten literacy measures. The present study addresses a distinct question: whether the effects of access to tutoring persist across four school years and extend to independent state-administered reading assessments and administrative measures of academic progression. During the 2021–22 school year, Chapter One partnered with a large urban district in the Southeast to evaluate its early-literacy tutoring program. The district identified 49 kindergarten classrooms across 13 schools to participate; these schools enrolled a

total of 7,891 students (average enrollment of 607), with school-level rates of free- and reduced-price lunch eligibility ranging from 79% to 92%. All kindergarten classrooms in the 13 schools were part of the study, and students were eligible if their parents consented to participate in Chapter One.

We conducted a student-level randomization stratified by classroom. Within each of the 49 classrooms, eligible students were randomly assigned either to the treatment group, which was offered access to the Chapter One adaptive instructional support system, or to the control group, which continued with business-as-usual instruction. Students assigned to treatment received individualized instructional support that evolved over time according to the program's decision rules and their reading progress. As described in Table 1, both instructional content and support intensity changed as students' reading developed, with students who reached grade-level expectations transitioning to lower-intensity support while students who continued to struggle received sustained individualized instruction. Because we randomized students within classrooms and instruction was delivered in the classroom or an adjacent space, treatment and control students shared the same teacher and tutor presence. This raises the possibility of interference between conditions, a violation of SUTVA. Interference could run in either direction. Control students might benefit indirectly through a tutor or teacher who reallocates some attention toward them or through improving peers, which would narrow the treatment-control gap toward zero. Alternatively, the program could draw attention away from control students, inflating the gap. We cannot rule out either channel, but the more plausible case is positive spillover to control students, which means our estimates likely understate the effect of the intervention relative to a classroom with no Chapter One instructional support.

In classrooms with an odd number of eligible students, the additional student was assigned to treatment. We randomized 815 students (420 assigned to treatment, 395 to control)[1] and followed them across four school years: kindergarten (2021–22), 1st grade (2022–23), 2nd grade (2023–24), and 3rd grade (2024–25). Students assigned to the treatment group who remained in participating schools were intended to continue receiving Chapter One tutoring in subsequent grades. Because the program is adaptive and embedded within classrooms, the amount of tutoring students received over the four years varied across students.

Table A1 summarizes the realized dosage among treated students with a nonmissing primary outcome. About 41 percent received tutoring in all four years, while the remaining 59 percent exited earlier. Most of this early exit reflects student mobility: roughly three-fifths of early exits occurred when students switched schools, and school switching accounts for the large majority of students who received only a single year of tutoring. Consistent with this pattern, early exit is largely unrelated to students' baseline characteristics (Table A2): students who exited early did not differ from those who completed all four years on baseline reading achievement or IEP status, and most demographic characteristics are similarly unpredictive, the one exception being that English learners were somewhat more likely to remain in the program. This pattern indicates that incomplete dosage is driven primarily by student mobility, with realized dosage modestly higher among students identified as English learners at baseline. Incomplete take-up therefore attenuates the intent-to-treat estimates relative to a fully implemented four-year program.

Consistent with the preregistered experimental design, all analyses follow an intent-to-treat framework: students are analyzed according to their original random assignment, regardless of the amount of tutoring they actually received in any given year. The cohort available for analysis changes across years as students leave the district, and thus we construct two analytic samples.

[1] Of the 818 students randomized in Cortes et al. (2025), we removed three students that the district confirmed to be duplicate records.

For our main cross-sectional analyses, we use the maximum available sample for each outcome and testing period. We also report robustness checks of the cross-sectional estimates on the balanced sample. For the event-study analysis, we restrict the sample to the balanced panel of students with scores in all windows, so that the trajectory of the treatment effect is not confounded by changes in sample composition over time.

Data and measures

We obtained administrative records from the district and Chapter One, including students' gender, race and ethnicity, English learner status, whether they had an IEP, and their scores on state- and program-administered assessments. As a baseline measure of reading readiness, we use each student's fall-2021 score on the state's Kindergarten Readiness Screener, which must be administered to all kindergarten students within the first 30 instructional days of the school year. In 2021–22, the screener was the Renaissance Star Early Literacy assessment. We include the baseline readiness score as a covariate in all models, along with an indicator for students whose baseline score was missing and imputed. We do not observe state-administered reading assessment outcomes during the kindergarten year beyond this baseline measure (see Cortes et al. (2025) for end-of-kindergarten impacts on a program-administered test).

Primary outcome: reading achievement (grades 1–3)

Our primary outcome is students' reading achievement in first, second, and third grade. Beginning in 2022–23, the state measured reading through their statewide assessment system administered three times each year during beginning, middle, and end of year progress-monitoring windows. The reading instrument differs by grade level. In grades 1 and 2, reading is measured using Renaissance's Star assessments. Students begin on Star Early Literacy and transition to Star Reading once they reach the established threshold. As a result, students within the same grade may be assessed on either instrument in a given window. Because both

assessments report scores on a common Unified scale, scores are comparable across the two instruments. We accordingly use each student's unified scale score regardless of which Star assessment they completed in a given window. Most students took Star Early Literacy at the beginning and middle of grade 1 and Star Reading from the end of grade 1 and throughout grade 2. In grade 3 (2024–25), reading is measured by the ELA Reading assessment, which is not directly comparable to the K–2 Star unified scale.

To place scores on a common metric for analysis, we standardize each score within its grade-by-window cell using the distribution of the control group in that cell. These standardized scores capture relative performance within each assessment period and thus should not be interpreted as directly comparable in magnitude across grades. Because the standard deviation used for scaling is specific to each period, and because the underlying assessment changes between second and third grade, an effect of a given size in one grade does not necessarily represent the same quantity of learning as an effect of the same size in another. We therefore focus on the direction and persistence of effects across grades, recognizing that precise comparisons of effect magnitudes across grades are less informative.

Reading achievement for retained students

Measuring 3rd grade reading achievement is complicated by how students progress through the early grades. Under the state's student progression framework, each district sets promotion and retention criteria at every grade, including kindergarten through grade 2. In the district, students identified as having a substantial reading deficiency—through screening, progress monitoring (including the Star assessments), or teacher observation—and those who fail to meet grade-level promotion criteria are considered for retention through a multi-tiered system of support (MTSS) review. As a result, some students in our cohort were retained and had not reached 3rd grade by 2024-25, and they were assessed on a lower-grade Star test rather than the grade 3 ELA Reading assessment, leaving their 3rd grade reading score unobserved.

Restricting the 4th year analysis to students who have a 3rd grade score would condition on grade progression, which itself is an outcome the intervention may affect. If tutoring helps lower-achieving students remain on track, then the treatment group's grade 3 test-takers would disproportionately include students who would otherwise have been retained, biasing the estimated treatment effect downward. To avoid this bias, we impute grade 3 reading scores for the retained students who did not take the 3rd grade assessment. We obtain predicted values from a regression of 3rd grade reading on 2nd grade reading, the baseline score, demographics, and treatment assignment, estimated among students with observed 3rd grade scores, and assign each retained student their predicted value. We include the prior-year (2nd grade) score because predictions from the baseline measure alone perform poorly. The kindergarten baseline is temporally distant and only weakly predictive of 3rd grade achievement. We retain the treatment indicator in the imputation model so that the imputation does not mechanically attenuate treatment-control differences. We do not impute scores for students who simply missed a test in a given year. Consequently, the 4th year reading sample includes observed 3rd grade test-takers plus imputed retained students, but excludes ordinary test-missers.

Additional exploratory outcomes

Oral reading fluency. Chapter One's Early Learning Interventionists administered the program's Oral Reading Fluency (ORF) benchmark assessment, a curriculum-based measure modeled on DIBELS Next (now Acadience Learning) and developed using a battery of readability indices to calibrate passage difficulty by grade level, with each benchmark passage approximately 250 words in length. We use the grade 3 ORF composite score, which ranges from 1 to 4 and combines words correct per minute (WCPM) and accuracy, following the program's scoring rubric and benchmarked against Hasbrouck and Tindal's (2017) compiled ORF norms. We standardize the score using the distribution of the control group. Because the ORF was administered by program staff to a limited subset of students and contains substantial missing

data, we present it as supplementary evidence. Our primary conclusions rely on the state-administered assessments.

Grade retention and IEP status. We additionally examine two consequential, administratively recorded outcomes observed in the 3rd grade year: grade retention and IEP status. Retention is coded 1 if a student was retained and did not reach 3rd grade, and 0 if the student was on track in 3rd grade. IEP status is coded 1 if a student had an IEP in the grade 3 year and 0 otherwise. We treat both outcomes as exploratory.

Analysis plan

Cross-sectional (endline) specification. Our primary specification estimates the intent-to-treat (ITT) effect at the end of each grade. For each grade *g*, we estimate:

$$Y_{ic} = \beta_0 + \beta_1 Treated_i + X_i'\phi + \mu_c + \varepsilon_{ic},$$

where Y_{ic} is the within-period standardized reading score of student i in baseline classroom c; $Treated_i$ indicates random assignment to Chapter One; X_i is the vector of baseline covariates, including baseline score bins, the missing-baseline indicator, race and ethnicity, gender, baseline English learner status, and baseline IEP status; and μ_c are classroom (randomization strata) fixed effects. Given the imbalance in baseline test scores, we adjust for baseline achievement non-parametrically by grouping students into score bins instead of modeling baseline achievement with a single linear term. The coefficient β_1 is the ITT effect of being assigned to tutoring on end-of-grade reading. We estimate this model on the maximum available sample in each grade and, as a robustness check, on the common (balanced) sample of students with valid scores in all three grades.

Event-study (longitudinal) specification. To trace the treatment effect across all nine post-randomization windows, we estimate a single pooled event-study model on the balanced panel:

$$Y_{ict} = \gamma\, Treated_i + \sum_{p \neq base} \delta_p \left(Treated_i \times 1[t = p] \right) + X_i' \phi + \mu_c + \varepsilon_{ict},$$

where t indexes the ten periods (the fall-of-kindergarten baseline plus the beginning, middle, and end of year windows of grades 1-3), and the baseline period is omitted. Because the outcome is standardized within each period to the control distribution, the control mean is zero in every period; period-level differences are therefore already removed from the outcome, and the model requires no period fixed effects. Each δ_p is the treatment effect in period p relative to the baseline difference as we omitted the baseline period. In this specification, X_i excludes the baseline score bins and the missing baseline indicator, because the baseline score enters through the reference period. The event study shows when effects emerge and whether they persist. Because each estimate is expressed in its own period-specific units and the third-grade estimate comes from a different assessment, we interpret the trajectory qualitatively, focusing on the direction and persistence of effects across periods.

Treatment-on-the-treated (TOT) specification. Assignment to Chapter One did not guarantee four years of tutoring: students received tutoring for varying numbers of years. To estimate the effect of the tutoring actually received, we use two-stage least squares, instrumenting the realized dose with random assignment:

$$First\ stage: \quad \widetilde{D}_i = \pi\, Treated_i + X_i' \phi + \mu_c + \nu_{ic},$$

$$Second\ stage: \quad Y_i = \rho\, \hat{\widetilde{D}}_i + X_i' \phi + \mu_c + \varepsilon_{ic},$$

where $\widetilde{D}_i = Dose_i/4$ and $Dose_i$ is the number of school years (0–4) in which student i received Chapter One tutoring. Scaling the dose by four lets us interpret ρ as the effect of a full four-year course of tutoring relative to none. This estimate rescales the ITT to the treated dose, and because the endogenous regressor enters linearly, the four-year estimate assumes a constant marginal effect of each additional year of tutoring (i.e., that the treatment effect is linear and continuous in dosage). We therefore interpret it as an approximation and consider it alongside the ITT estimates.

Results

Balance and attrition

Table 2 reports baseline descriptive statistics and the balance test. Treatment and control students are similar on IEP status and on the broad distribution of race and ethnicity. The treatment group, however, is slightly less likely to be White, more likely to be female, more likely to be classified as an English learner, and has lower baseline reading scores (−0.08 vs. 0.08 SD) than the control group. We account for these differences by including them as baseline covariates in all models.

Table 3 examines attrition over the four years using two indicators: leaving the district and missing the endline reading test. A non-trivial share of students left the district (19 percent of the control group) and were missing an endline test score (26 percent of the control group). However, we find little evidence that attrition differs by treatment status: the estimated treatment-control difference is close to zero for both indicators (Model 1) and the treatment-by-covariate interactions show no concerning patterns. These results reduce the concern that our main findings are driven by differential sample loss across conditions.

Effects on students' reading trajectories

Our primary question is whether an adaptive instructional support system built on the principles of effective tutoring can improve students' reading trajectories across the learning-to-read years. We first examine effects on reading achievement before turning to measures of academic progression. Table 4 presents the intent-to-treat (ITT) effects of assignment to the Chapter One adaptive instructional support system on state-administered reading assessments. Scores were measured at the end of each grade and standardized within each testing period using the control group distribution. Students assigned to Chapter One scored higher than students assigned to business-as-usual instruction. Specifically, treated students scored 0.15 SD higher in 1st grade (SE = 0.077, p = 0.063), 0.17 SD higher in 2nd grade (SE = 0.074, p = 0.022), and 0.12 SD higher in 3rd grade (SE = 0.074, p = 0.117).

The common-sample estimates (Table A3) follow the same pattern, though less precise because the sample is smaller, indicating that the results are not driven by changes in sample composition across grades. Because the 3rd grade estimate is positive but imprecise, we complement it with a Bayesian interpretation using the BASIE framework (Deke et al., 2022). The posterior analysis provides additional evidence that the endline estimate is more consistent with a positive effect than with a null or negative one: we estimate a 93 percent probability that the tutoring intervention had a positive effect on third-grade reading scores, and a 79 percent probability that the effect was at least 0.05 SD.[2]

Figure 1 and Table A4 provide complementary evidence on how students' reading trajectories evolved over time under the Chapter One adaptive instructional support system. The top panel plots event-study estimates relative to the baseline period, and the bottom panel shows unadjusted mean trajectories for the treatment and control groups. The mean trajectories show

[2] For the prior, we use BASIE's evidence-based prior for elementary ELA achievement, estimated from impact estimates in the What Works Clearinghouse (WWC) database that meet evidence standards and adjusted for small-study effects.

that treated students began below control students at baseline but closed much of the gap over time as they gained more relative to control students. The event-study estimates show the same pattern: relative to the baseline treatment–control gap, the treatment effect increases during first grade and remains positive and broadly stable thereafter, ranging from roughly 0.15 to 0.24 SD across the windows of 2nd and 3rd grade. It is important to note that magnitudes are not strictly comparable across grades because scores are standardized within each period and the third-grade assessment differs from the instruments used in grades 1 and 2. The event-study estimates are somewhat larger than the cross-sectional estimates in Table 4, which is expected. The event study measures each period's effect relative to the baseline gap and thus assumes the gap would have persisted absent tutoring, whereas the cross-sectional models compare students with similar baseline scores and so net out the growth that low-scoring students would have experienced anyway. We therefore read the event study as evidence on when the effect emerged and whether it persisted, and rely on the cross-sectional estimates for its magnitude.

Effects on longer-term academic progression

As exploratory analyses, we examine the effects of tutoring on two administratively recorded outcomes that capture students' academic progression and are likely shaped by early reading performance: IEP status and grade retention. Table 5 reports ITT effects on both at the end of the study. Students assigned to tutoring were 4.7 percentage points less likely to have an IEP by 3rd grade (SE = 0.022, p = 0.040), relative to a control rate of 22 percent. They were also 2.7 percentage points less likely to be retained (SE = 0.019, p = 0.180), relative to a control rate of 9 percent, though the retention estimate is imprecise. Both estimates directionally suggest that the program reduced IEP designation and retention. On the provider-administered Oral Reading Fluency assessment, treated students scored 0.40 SD higher (SE = 0.115, p = 0.001). Because this assessment was administered to a limited subset of students who remained in the sample in

the 4th year of the study and contains substantial missing data, we treat it as supporting evidence for the test score results and do not rely on it as independent confirmatory evidence.

Heterogeneity

Table 6 asks whether these effects differ by students' baseline IEP status. While these analyses are exploratory, they reveal a consistent pattern. For students who had an IEP at baseline, the effects of tutoring are substantially larger than for their peers. Assignment to tutoring lowered endline IEP status by an additional 13.9 percentage points (SE = 0.064, p = 0.035) and lowered retention by an additional 13.4 percentage points (SE = 0.069, p = 0.058). The reading-score effect is likewise larger for this group (0.29 SD, SE = 0.242, p = 0.237), though it is imprecisely estimated. The progression effects documented in Table 5 are thus concentrated among the students who entered with the greatest identified need. Across the other subgroups we examined (baseline reading ability, English language learner status, race and ethnicity, and gender), we find no meaningful heterogeneity.

Treatment-on-the-treated estimates for sustained participation

Because not every student assigned to treatment remained in the Chapter One adaptive instructional support system throughout the full four-year study period (Table A1), we estimated treatment-on-the-treated (TOT) effects by instrumenting years of participation with random assignment. Table A5 reports the resulting two-stage least squares estimates. Scaled to continuous participation throughout the four-year study period, the estimated effect on third-grade reading is 0.18 SD (SE = 0.11, p = 0.123), larger than the corresponding ITT estimate but not statistically significant. The TOT estimates for the academic progression outcomes are likewise larger than their ITT counterparts and point in the same direction: continuous participation is associated with a 7.2 percentage-point lower probability of having an IEP at endline (SE = 0.034, p = 0.039) and a 4.1 percentage-point lower probability of grade retention

(SE = 0.030, p = 0.180). Because years of participation enter the model linearly, these estimates assume that each additional year of participation has the same marginal effect. We therefore interpret them as rescalings of the ITT under continuous participation in the adaptive instructional support system and not as precise estimates of the cumulative effects of additional years of support.

Discussion

We provide longitudinal experimental evidence on an adaptive instructional support system designed to translate the instructional principles underlying effective intensive tutoring into a model schools can sustain throughout the learning-to-read years. We followed a single cohort from kindergarten through 3rd grade, extending the first-year findings of Cortes et al. (2025), which documented large kindergarten gains on program-aligned measures. This study addresses an important gap in the tutoring literature, which is dominated by single-year interventions evaluated immediately after programs end (Nickow et al., 2024). Specifically, we ask whether schools can organize individualized instructional support over multiple years by adapting instructional content and support intensity as children's reading develops. Using independent state-administered reading assessments and administrative measures of academic progression, we find modest but consistently positive improvements in reading achievement that persist from 1st through 3rd grade, together with reductions in IEP status and grade retention, although the latter estimate is imprecise.

These findings suggest that schools can organize individualized instructional support as an adaptive system that evolves with students' reading development. As illustrated in Table 1, instructional content and support intensity change as students' needs change. Students who continue to struggle receive sustained individualized instruction, while students who reach grade-level expectations transition to lower-intensity support that maintains progress and allows

instructional resources to be concentrated where they are most needed. In this model, tutoring functions as one component of a broader instructional system whose goal is to match individualized support to students' evolving learning needs across the learning-to-read years.

This design also changes the appropriate benchmark for interpreting the magnitude of the effects. At 0.12 to 0.17 SD, the estimated impacts are smaller than the average intensive tutoring effect reported in recent meta-analyses (0.29 SD; Nickow et al., 2024), but the program was not designed to maximize instructional intensity for every student throughout the duration of the intervention. It was designed to sustain individualized instructional support under the staffing, scheduling, budget, and student mobility constraints that districts routinely face. The relevant comparison is therefore not the largest effect achieved by the most intensive tutoring models under ideal conditions, but the level of individualized instructional support that school systems can realistically sustain over multiple years. Nor is the appropriate benchmark simply four times the effect of a one-year intervention because achievement gains from early academic interventions typically fade after support ends (Bailey et al., 2017; Hart et al., 2024).

In addition, two features of the assessments can make the estimates look smaller in later grades even if the underlying effect does not weaken. First, Chapter One's instruction centers on foundational literacy skills, whereas reading assessments capture an increasingly broad construct as students advance, incorporating vocabulary, comprehension, and grade-level standards. A constant amount of foundational support may therefore register as a smaller standardized effect. Second, typical annual growth in reading declines sharply over the early grades, roughly 0.97 standard deviations in 1st grade, 0.60 in 2nd grade, and 0.36 in 3rd grade (Hill et al., 2008). Against this benchmark, our estimates translate to about 1.55 months of learning in 1st grade, 2.83 months in 2nd, and 3.33 months in 3rd. In other words, the effects represent a larger share of a year's expected growth in each successive grade, even though the standardized coefficients

themselves shrink. This translation is approximate as the growth norms come from nationally normed samples.

The interpretation of these estimates also depends on students' participation in the adaptive instructional support system over time. Assignment to treatment did not guarantee continuous participation throughout the learning-to-read years. Fewer than half of treatment students in the analytic sample remained in the Chapter One program through all four years (Table A1), largely because school mobility interrupted participation. The estimates therefore capture the effect of offering an adaptive instructional support system under the conditions districts actually face. They do not estimate the effect of uninterrupted participation from kindergarten through 3rd grade. This incomplete participation was largely unrelated to students' baseline characteristics (Table A2), suggesting that its primary effect was to attenuate the intent-to-treat estimates, with little evidence of selective non-compliance. We view the intent-to-treat estimate as the primary policy-relevant quantity because it reflects what districts can expect when implementing a multi-year adaptive instructional support system in a mobile student population. The pattern of participation also has an important practical implication. Realizing the full benefits of sustained individualized instructional support may require systems that maintain continuity as students move across schools, such as transferring records, coordinating transitions, and re-enrolling students when appropriate. For districts seeking to sustain adaptive instructional support over multiple years, continuity across student moves may be as important as the amount of support students receive within any single school year. At the same time, the adaptive nature of the support may partially offset these transitional costs. Because the system diagnoses each student's current skill level rather than assuming a fixed sequence of prior instruction, it can recalibrate quickly when a student re-enters after a move or gap in participation. If these approaches were adopted systemically by schools, adaptive instructional support may be particularly well suited to mobile student populations, even as continuity of enrollment remains important for maximizing its cumulative benefits.

The dosage pattern also shapes what the 3rd grade estimate reflects. About 41 percent of treated students received individualized instructional support through 3rd grade, and the rest had stopped one to three years earlier mostly because they changed schools. Therefore, the 3rd grade estimate combines the effects of continued participation with the persistence of gains among students whose participation ended early. We cannot separate the two directly because the number of years students stayed in the program was not randomly assigned. Appendix Table A6 shows what a range of fade-out rates implies for this mix. At one extreme, full fade-out for early exiters would require an effect of roughly 0.29 SD among the students tutored in all four years to produce the pooled estimate of 0.12 SD. At the other extreme, full persistence would mean every treated student, including those tutored only in kindergarten, carried a roughly 0.12 SD advantage years after their tutoring ended. Applying the average fade-out rate documented in the literature (Hart et al., 2024) lands between the two: persisting gains from earlier tutoring and gains from continued tutoring each account for roughly half of the 3rd grade effect.

Beyond test scores, the findings suggest that sustained adaptive instructional support may influence students' academic progression. Treatment students were less likely to have an IEP by 3rd grade and less likely to be retained, though the retention estimate is imprecise. These outcomes are important because early reading difficulty can shape students' later access to grade-level content and contribute to referral for additional services or delayed progression.

The reduction in the likelihood of having an IEP appears to reflect students exiting special education services, with little evidence that it was driven by reduced entry into special education. Of the roughly 12 percent of the sample who had an IEP in kindergarten—a group balanced across conditions—the students who no longer had an IEP by 3rd grade were almost entirely in the treatment group. This pattern is notable because it suggests that the overall reduction in IEP status was not simply the result of fewer students being newly identified, but also reflected changes among students who entered the treatment with identified needs. IEP status reflects both

genuine student need and administrative decision-making, and exiting special education services is not inherently beneficial when those services are appropriate and necessary. However, the reduction in IEP status appears alongside positive reading effects and lower retention among students who had an IEP at baseline. The convergence of these outcomes strengthens the interpretation that tutoring may have helped some students with IEPs make enough academic progress to remain on track without continued special education services. This pattern suggests that sustained individualized instructional support strengthened foundational literacy for students with the greatest identified need at the outset. As a result, some students remained on a grade-level trajectory, lowering their risk of retention, and, in some cases, reducing the need for continued special education services.

The broader implications of this study are both conceptual and practical. The tutoring literature has largely viewed tutoring as an intervention whose effectiveness depends on its intensity, frequency, and duration. Our findings suggest broadening this perspective. Rather than viewing these features as fixed characteristics of a tutoring intervention, tutoring can be understood as one instructional resource within a broader adaptive system that reallocates individualized support as students' needs change. Under this perspective, the central design question shifts from determining how much tutoring students should receive to determining how schools should organize individualized instructional support across the learning-to-read years.

Our study has several limitations. First, although treatment assignment was randomized, treatment and control students were not perfectly balanced at baseline, and treatment students entered with lower reading scores. The preferred models adjust for baseline scores non-parametrically, so the estimates compare students with similar starting points. Second, the special education and retention analyses rely on rare events, and we observe whether students had an IEP but not the decision-making processes behind changes in IEP status. These outcomes were also not preregistered, and we therefore view these findings as exploratory. Third, since we

randomized students within classrooms and tutoring is delivered in the room, we cannot rule out spillovers between the two groups. The most likely form is positive spillover to control students, through reallocated teacher attention, which would lead our within-classroom estimates to understate the program's full effect on a classroom. But it complicates the interpretation of the control group as a clean no-tutoring counterfactual.

This study suggests that schools can translate the instructional principles underlying effective intensive tutoring into adaptive instructional support systems that remain feasible to sustain over multiple years. The resulting effects are more modest than those reported for the most intensive tutoring models, but they were achieved under ordinary staffing, scheduling, budget, and student mobility constraints while providing individualized support throughout the learning-to-read years. For policymakers, the findings suggest that adaptive systems built around brief individualized instruction, ongoing progress monitoring, and flexible allocation of instructional support can strengthen students' reading trajectories and academic progression while remaining feasible to sustain at scale. More broadly, the findings suggest that tutoring can function as one component of an adaptive instructional support system that adjusts instructional content and support intensity as students' needs change. In this approach, individualized instructional support evolves with children's reading development, allowing schools to concentrate intensive support on students who continue to need it while maintaining appropriate support for students who have reached grade-level expectations. This perspective highlights the importance of designing instructional systems that continually match support to students' changing needs across the learning-to-read years.

References

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*(6), 934–945. https://doi.org/10.1037/0012-1649.33.6.934

Curran, F. C., Bal, A., Goff, P., & Mitchell, N. (2021). Estimating the Relationship Between Special Education De-Identification for Emotional Disturbance and Academic and School Discipline Outcomes: Evidence From Wisconsin's Longitudinal Data. *Education and Urban Society*, *53*(1), 83–112. https://doi.org/10.1177/0013124520925053

Deke, J., Finucane, M., & Thal, D. (2022). *The BASIE (BAyeSian Interpretation of Estimates) framework for interpreting findings from impact evaluations: A practical guide for education researchers*. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee

Dewey, D. C., Fahle, E., Kane, T. J., Reardon, S., & Staiger, D. O. (2026). *From Learning Recession to Learning Recovery: Understanding the Sources of U.S. K-12 Improvement*. Center for Education Policy Research and The Educational Opportunity Project.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

ExcelinEd. (2026). *Comprehensive Early Literacy Policy: State-by-State Policy Implementation Report*. ExcelinEd.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2018). *Learning Disabilities: From Identification to Intervention*. Guilford Publications.

Fuchs, D., Fuchs, L. S., & Vaughn, S. (2014). What Is Intensive Instruction and Why Is It Important? *TEACHING Exceptional Children*, *46*(4), 13–18. https://doi.org/10.1177/0040059914522966

Gilmour, A. F., Shanks, C., & Winters, M. A. (2023). Choice, Mobility, and Classification: Disaggregating the Charter School Special Education Gap. *Remedial and Special Education*, *44*(4), 272–282. https://doi.org/10.1177/07419325221115421

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., & Watts, T. W. (2024). Fadeout and persistence of intervention impacts on social–emotional and cognitive skills in children and adolescents: A meta-analytic review of randomized controlled trials. *Psychological Bulletin*, *150*(10), 1207–1236. https://doi.org/10.1037/bul0000450

Hasbrouck, J., & Tindal, G. (2017). *An Update to Compiled ORF Norms*. Behavioral Research and Teaching.

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.x

Hwang, N., & Koedel, C. (2025). Helping or Hurting: The Effects of Retention in the Third Grade on Student Outcomes. *Educational Evaluation and Policy Analysis*, *47*(1), 65–88. https://doi.org/10.3102/01623737231197639

Jacob, B. A., & Lefgren, L. (2009). The Effect of Grade Retention on High School Completion. *American Economic Journal: Applied Economics*, *1*(3), 33–58. https://doi.org/10.1257/app.1.3.33

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*(4), 437–447. https://doi.org/10.1037/0022-0663.80.4.437

Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of Data–Based Individualization for Students with Intensive Learning Needs: A Meta–Analysis. *Learning Disabilities Research & Practice*, *33*(3), 144–155. https://doi.org/10.1111/ldrp.12172

Language and Reading Research Consortium. (2015). Learning to Read: Should We Keep Things Simple? *Reading Research Quarterly*, *50*(2), 151–169. https://doi.org/10.1002/rrq.99

National Center for Education Statistics. (2024). *The Nation's Report Card: 2024 NAEP reading assessment*. U.S. Department of Education, Institute of Education Sciences. https://www.nationsreportcard.gov/

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*, 154–169. https://doi.org/10.1016/j.jpubeco.2017.06.004

Spencer, N. (2024). Comprehensive early literacy policy and the "Mississippi Miracle." *Economics of Education Review*, *103*, 102598. https://doi.org/10.1016/j.econedurev.2024.102598

Stecker, P. M., Fuchs, D., & Fuchs, L. S. (2008). Progress Monitoring as Essential Practice within Response to Intervention. *Rural Special Education Quarterly*, *27*(4), 10–17. https://doi.org/10.1177/875687050802700403

Westall, J., & Cummings, A. (2023). *The Effects of Early Literacy Policies on Student Achievement*. Education Policy Innovation Collaborative. https://eric.ed.gov/?id=ED672706

Zhong, J. (forthcoming). Early Grade Retention Harms Adult Earnings. *American Economic Journal: Applied Economics*. https://doi.org/10.1257/app.20230121

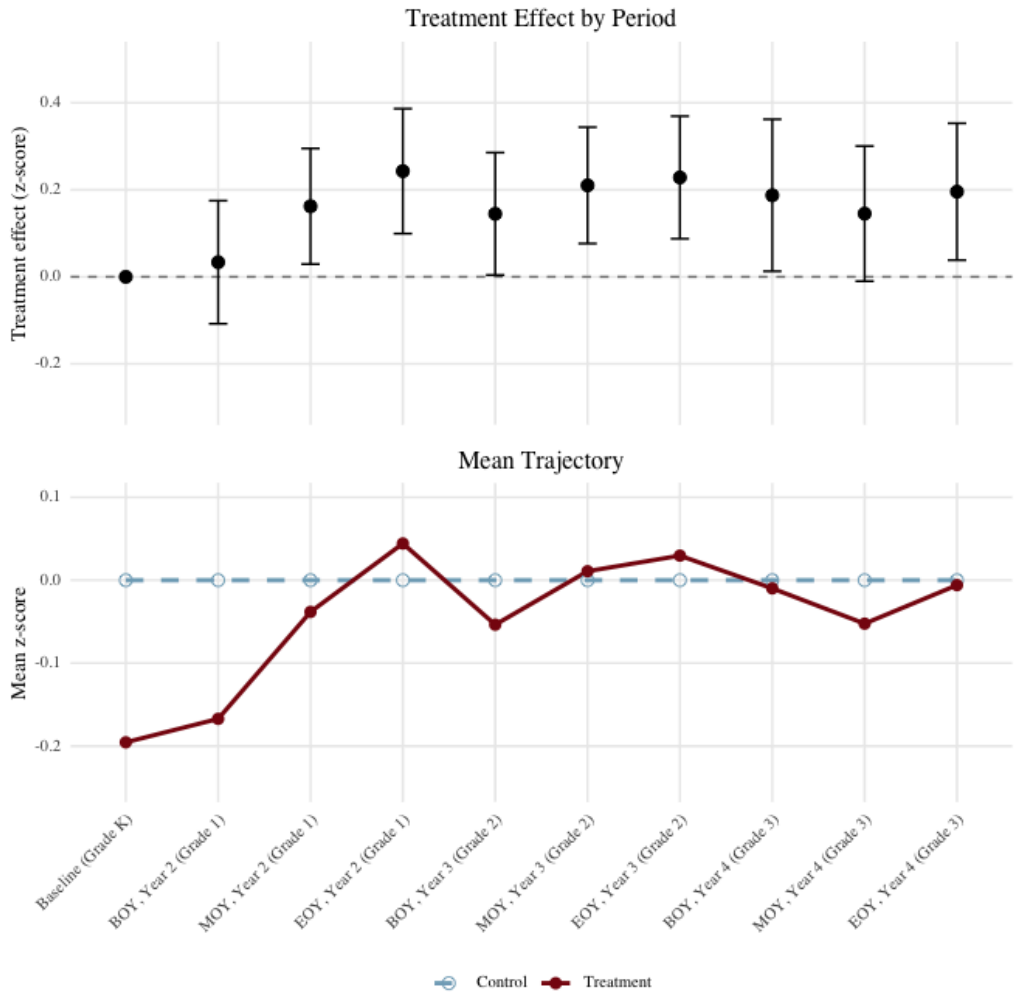

Figure 1: Event-study estimates of the effect of early literacy tutoring on state-administered reading test scores, with unadjusted mean trajectories for treatment and control groups

Note. The top panel reports event-study estimates of the intent-to-treat effect of assignment to early literacy tutoring on reading test scores. Estimates are relative to the baseline period, Grade K, which is normalized to zero. Bars show 95% confidence intervals based on robust standard errors clustered at the classroom level. The bottom panel shows unadjusted treatment and control group means by period. Test scores are standardized within each period relative to the control group distribution of that period. Therefore, the control group mean is zero by construction in each period. The figure uses the balanced panel. BOY, MOY, and EOY denote the beginning, middle, and end of the school year, respectively.

Table 1. Allocation of individualized support across the learning-to-read years

Student's reading level	Kindergarten	Grade 1	Grade 2	Grade 3
Below Benchmark (learning to decode)	1:1 decoding, ~2.5 sessions/week (~15 min) + 10 min/day platform	Same	Same	Same
At or above benchmark	1:1 maintenance, ~1x/week	1:1 maintenance, ~1x/week	1:1 supported reading, 1x/week (10 min) + 1:4 book club, 1x/week (30 min)	1:1 supported reading, 2x/week (10 min) + 10 min/day independent reading
Purpose	Build decoding	Complete decoding	Move from decoding to fluent reading	Sustain fluent reading
Cost per student	$375	$375	$450	$465

Table 2. Sample Descriptive Statistics and Balance Test

Variable[1]	Overall, n = 815[1]	Control, n = 395[1]	Treatment, n = 420[1]	**p-value**[1]
White	0.044	0.068	0.021	0.001
Black	0.717	0.706	0.726	0.500
Hispanic	0.206	0.195	0.217	0.400
Other Race	0.029	0.028	0.031	0.800
Female	0.473	0.439	0.505	0.061
English Language Learner	0.278	0.246	0.309	0.047
Has an IEP	0.107	0.107	0.108	>0.900
Missing	0.004	0.003	0.005	0.600
Demographics				
Baseline Test Score	0.001	0.083	-0.076	0.024

[1]p-values are from two-sample t-tests comparing treatment and control groups. The baseline test score is from a screening assessment which must be administered to all public school kindergarten students within the first 30 instructional days of each school year.

Table 3. Sample Attrition

	Leaving the District			Missing Endline Test Score		
	(1)	(2)	(3)	(1)	(2)	(3)
Treatment	-0.011	0.001	0.007	-0.013	-0.000	-0.021
	(0.027)	(0.028)	(0.037)	(0.026)	(0.027)	(0.039)
White		0.205*	0.203+		0.223*	0.167
		(0.085)	(0.113)		(0.091)	(0.101)
Hispanic		-0.031	-0.023		-0.048	-0.041
		(0.045)	(0.059)		(0.061)	(0.072)
Other Race		-0.039	0.021		-0.069	-0.040
		(0.073)	(0.130)		(0.078)	(0.117)
Female		-0.034	-0.029		-0.048+	-0.050
		(0.027)	(0.040)		(0.028)	(0.051)
Missing demographics		-0.281***	-0.268***		-0.304***	-0.356***
		(0.024)	(0.026)		(0.025)	(0.029)
English Language Learner		-0.037	-0.047		-0.004	-0.011
		(0.039)	(0.064)		(0.051)	(0.076)
Has an IEP		-0.042	-0.034		-0.074+	-0.120
		(0.039)	(0.066)		(0.042)	(0.074)
Baseline Score		-0.022	-0.027		-0.020	-0.036+
		(0.017)	(0.021)		(0.018)	(0.019)

Treatment × White	0.011	0.232
	(0.202)	(0.172)
Treatment × Hispanic	-0.012	-0.006
	(0.080)	(0.088)
Treatment × Other Race	-0.110	-0.045
	(0.125)	(0.127)
Treatment × Female	-0.008	0.002
	(0.050)	(0.068)
Treatment × Missing demographics	-0.018	0.091*
	(0.044)	(0.040)
Treatment × English Language Learner	0.018	0.015
	(0.075)	(0.087)
Treatment × Has an IEP	-0.013	0.093
	(0.084)	(0.100)
Treatment × Baseline Score	0.009	0.033
	(0.028)	(0.035)

N	815	815	815	815	815	815
Adjusted R2	0.083	0.103	0.104	0.081	0.102	0.106
Control mean	0.190	0.190	0.190	0.258	0.258	0.258
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Robust standard errors clustered at the classroom level in parentheses.

Table 4. The Effect of Early Literacy Tutoring on State-administered Reading Test Scores

Year	Estimate	N
Year 2 EOY (Grade 1)	0.147+	694
	(0.077)	
Year 3 EOY (Grade 2)	0.174*	650
	(0.074)	
Year 4 EOY (Grade 3)	0.117	606
	(0.074)	
Baseline covariates	Yes	
Classroom FE	Yes	

+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001. Each row reports the coefficient from a separate regression. Test scores are standardized within each period relative to the control group distribution of that period. Baseline covariates include baseline test score bins (with scores imputed for students missing the assessment), an indicator for students missing baseline scores, race/ethnicity, gender, English language learner status, and an indicator for having an IEP at baseline. Robust standard errors clustered at the classroom level in parentheses.

Table 5. The Effect of Early Literacy Tutoring on Endline IEP Status, Grade Retention, and Oral Reading Fluency

	Has an IEP	Grade Retention	Oral Reading Fluency
Treatment	-0.047*	-0.027	0.399**
	(0.022)	(0.019)	(0.115)
N	661	661	235
Adjusted R2	0.528	0.205	0.439
Baseline Control Mean	0.12	-	-
Endline Control Mean	0.22	0.09	0.00
Baseline covariates	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Oral Reading Fluency (ORF), standardized based on the control group distribution, is a provider assessment of reading fluency administered at the end of year 4. The ORF sample is smaller because of missing provider assessment data; thus, results for this outcome should be interpreted with caution. Baseline covariates include baseline test score bins (with scores imputed for students missing the assessment), an indicator for students missing baseline scores, race/ethnicity, gender, English language learner status, and an indicator for having an IEP at baseline. Robust standard errors clustered at the classroom level in parentheses.

Table 6. Heterogeneity by Baseline IEP Status

	State-administered Reading Test Score	Has an IEP	Grade Retention
Treatment	0.082	-0.030	-0.010
	(0.076)	(0.024)	(0.020)
Treatment × IEP at Baseline	0.290	-0.139*	-0.134+
	(0.243)	(0.064)	(0.069)
IEP at Baseline	-0.041	0.860***	0.153*
	(0.167)	(0.042)	(0.065)
N	606	661	661
Adjusted R2	0.375	0.530	0.210
Baseline Control Mean	0.12	-	-
Endline Control Mean	0.22	0.09	0.00
Baseline covariates	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The reference group is students who did not have an IEP at baseline. Test scores are standardized within each period relative to the control group distribution of that period. Baseline covariates include baseline test score bins (with scores imputed for students missing the assessment), an indicator for students missing baseline scores, race/ethnicity, gender, English language learner status, and an indicator for having an IEP at baseline. Robust standard errors clustered at the classroom level in parentheses.

Appendix

Table A1. Distribution of Realized Tutoring Dosage Among Treated Students, by Reason for Early Exit

Dosage (years receiving tutoring)	Overall	Switching schools	Other reasons
1	73 (23.1%)	59 (51.8%)	14 (19.4%)
2	94 (29.7%)	45 (39.5%)	49 (68.1%)
3	19 (6.0%)	10 (8.8%)	9 (12.5%)
4	130 (41.1%)	—	—
—	316 (100.0%)	114 (100.0%)	72 (100.0%)

Note. The sample is restricted to treatment-group students with a nonmissing endline test score. Each cell reports the number of students with the percentage in parentheses; percentages are computed within each column. "Switching schools" indicates students who switched their school during the study; "Other reasons" indicates students who remained in the same school but stopped tutoring early.

Table A2. Predicting Early Tutoring Exit from Baseline Characteristics

		By exit channel	
	Any early exit	Switching schools	Other reasons
Baseline reading score	0.021	0.030	-0.023
	(0.029)	(0.036)	(0.028)
Missing baseline score	0.075	-0.073	0.170
	(0.086)	(0.082)	(0.108)
White	0.200	0.232	
	(0.160)	(0.168)	
Hispanic	0.063	0.050	0.018
	(0.087)	(0.100)	(0.085)
Other race/ethnicity	-0.177+	-0.204*	-0.057
	(0.090)	(0.099)	(0.068)
Female	0.021	0.089	-0.068
	(0.059)	(0.066)	(0.063)
Has an IEP (baseline)	0.096	0.097	0.073
	(0.066)	(0.090)	(0.084)
English language learner (baseline)	-0.145*	-0.099	-0.178**
	(0.054)	(0.074)	(0.057)
N	314	239	198
Adjusted R2	0.387	0.419	0.569
Mean of outcome	0.586	0.456	0.354
Classroom FE	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Each column reports a linear probability model in which the outcome equals one if the student stopped tutoring through the indicated channel. The estimation sample is treatment-group students with a nonmissing endline test score. In columns (2) and (3) the comparison group is students who completed all four years; exits through the other channel are excluded. Race/ethnicity indicators are relative to Black students. Robust standard errors clustered at the classroom level in parentheses.

Table A3. The Effect of Early Literacy Tutoring on State-administered Reading Test Scores (Common sample across years)

Year	Estimate	N
Year 2 EOY (Grade 1)	0.143	496
	(0.091)	
Year 3 EOY (Grade 2)	0.141+	496
	(0.080)	
Year 4 EOY (Grade 3)	0.093	496
	(0.082)	
Baseline covariates	Yes	
Classroom FE	Yes	

+ p<0.1, * p<0.05, ** p<0.01, *** p<0.001. Each row reports the coefficient from a separate regression. Test scores are standardized within each period relative to the control group distribution of that period. Baseline covariates include baseline test score bins (with scores imputed for students missing the assessment), an indicator for students missing baseline scores, race/ethnicity, gender, English language learner status, and an indicator for having an IEP at baseline. Robust standard errors clustered at the classroom level in parentheses.

Table A4. Event Study Estimates of Tutoring Effects on Reading Test Scores

	(1)
BOY, Year 2 (Grade 1)	0.034
	(0.070)
MOY, Year 2 (Grade 1)	0.162*
	(0.066)
EOY, Year 2 (Grade 1)	0.243**
	(0.071)
BOY, Year 3 (Grade 2)	0.145*
	(0.070)
MOY, Year 3 (Grade 2)	0.210**
	(0.066)
EOY, Year 3 (Grade 2)	0.228**
	(0.070)
BOY, Year 4 (Grade 3)	0.187*
	(0.087)
MOY, Year 4 (Grade 3)	0.145+
	(0.077)
EOY, Year 4 (Grade 3)	0.196*
	(0.078)
N	4960
Adjusted R2	0.165
Demographics	Yes
IEP and English language learner controls	Yes
Classroom FE	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Test scores are standardized within each period relative to the control group distribution of that period. Covariates include race/ethnicity, gender, English language learner status, and an indicator for having an IEP at baseline. Baseline test scores are not included because the baseline period serves as the reference. Robust standard errors clustered at the classroom level in parentheses. The reference period is the baseline (Grade K).

Table A5. Effect of Four Full Years of Tutoring (Treatment-on-the-Treated Effects from 2SLS)

	Reading Test Score	Has an IEP	Grade Retention
Full four-year tutoring equivalent (TOT)	0.178	-0.072*	-0.041
	(0.113)	(0.034)	(0.030)
N	606	661	661
Endline Control Mean	0.00	0.22	0.09
First Stage Coefficient	0.661	0.652	0.652
F-statistic of IV	1422.8	1590.5	1590.5
Baseline covariates	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tutoring dosage is defined as the number of school years in which a student received tutoring and ranges from 0 to 4. For ease of interpretation, we divide dosage by four before using it as the endogenous treatment variable. The reported coefficient can therefore be interpreted as a treatment-on-the-treated estimate scaled to a full four-year tutoring equivalent. Test scores are standardized within each period relative to the control group distribution of that period. Baseline covariates include baseline test score bins (with scores imputed for students missing the assessment), an indicator for students missing baseline scores, race/ethnicity, gender, English language learner status, and an indicator for having an IEP at baseline. Robust standard errors clustered at the classroom level in parentheses.

Persistence versus continued tutoring in the 3rd grade effect

The 3rd grade estimate pools two components: gains from tutoring received through grade 1 that persisted, and gains from tutoring received in grades 2 and 3.[3] Our data cannot separate the two, because the number of years students remained in the program was not randomly assigned. Early exit was driven mainly by school moves and largely unrelated to students' baseline characteristics (Table A2); the calculation below assumes it is likewise unrelated to their potential outcomes. We can then gauge the split using fade-out rates documented in the literature. The 3rd grade estimate can be written as the sum of two parts:

$$\hat{\beta}_3 = p \cdot \hat{\beta}_1 + C,$$

where $\hat{\beta}_1$ and $\hat{\beta}_3$ are the ITT effects on reading scores at the end of grade 1 and the end of grade 3, p is the share of the end-of-grade-1 effect that would still remain at the end of grade 3 absent further tutoring, and C is the contribution of tutoring received in grades 2 and 3. We use the point estimates from Table 4: $\hat{\beta}_1 = 0.15$ SD and $\hat{\beta}_3 = 0.12$ SD. These estimates are imprecise, but the calculation requires point values, and the results should be read with that additional uncertainty in mind.

We consider three scenarios for p. Our preferred scenario uses the literature average, $p = 0.35$: in a meta-analysis of 86 randomized trials (Hart et al., 2024), the conditional persistence rate for cognitive skills, that is, the share of the end-of-treatment impact expected to remain one to two years later, is 0.35. Under full fade-out ($p = 0$), none of the end-of-grade-1 effect remains by the end of grade 3, so the entire 3rd grade effect comes from tutoring received in grades 2 and 3.

[3] A finer breakdown is not feasible. We observe no district-administered outcome before the end of grade 1, so the kindergarten and grade-1 contributions cannot be separated. Separating grade 2 from grade 3 tutoring is also uninformative: the share of treated students receiving tutoring falls only from 47 to 41 percent between those grades, and a year-by-year version of this calculation is too sensitive to the assumed fade-out rate to be meaningful.

Under full persistence ($p = 1$), the end-of-grade-1 effect carries forward in full whether or not a student continued receiving tutoring.

Under the literature-average rate, persisting gains from tutoring through grade 1 (0.05 SD) and gains from continued tutoring (0.07 SD) each account for roughly half of the 3rd grade effect. Under full fade-out, gains vanish once tutoring stops, so the entire 3rd grade effect comes from tutoring in grades 2 and 3; because only 41 percent of treated students were still tutored in grade 3, producing the pooled estimate of 0.12 SD requires an effect of roughly 0.29 SD among those students. Under full persistence, persisted gains alone (0.15 SD) slightly exceed the observed 3rd grade estimate, leaving nothing for continued tutoring to explain; the 0.03 SD gap is well within sampling error (SE = 0.074).

Table A6. Persisted Gains Versus Gains from Continued Tutoring in the 3rd Grade Effect, by Fade-Out Scenario

Scenario	Grade-1 effect remaining (p)	Persisted gains (0.15 × p)	Continued-tutoring gains (C)
Literature	0.35	0.05	0.07
Full fade-out	0.00	0.00	0.12
Full persistence	1.00	0.15	-0.03

Note. Effects are in within-period control-group SD units. 0.15 is the end-of-grade-1 ITT estimate and 0.12 the end-of-grade-3 ITT estimate (Table 4); C = 0.12 - 0.15 × p is the contribution of tutoring received in grades 2 and 3. The literature rate is the conditional persistence rate for cognitive skills at one-to-two-year follow-ups in Hart et al. (2024). Under full persistence, persisted gains slightly exceed the observed 3rd grade estimate; the gap (-0.03) is well within sampling error (SE = 0.074).