Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1655

Listening at Scale: A Pre-Registered Quasi-Experimental Evaluation of an Automatic Speech Recognition Reading Tool in Kindergarten Through Grade 2

Yun-En Liu
SETA-ED

Melina Uncapher
SETA-ED

Oral reading fluency develops through guided oral reading with corrective feedback, yet a teacher in Kindergarten through Grade 2 (K-2) can listen to only one child at a time. We evaluate LitLab, a supplemental tool that uses automatic speech recognition (ASR) to score children's oral reading and report results to teachers, in a pre-registered quasi-experimental study in nine elementary schools in one high-poverty Arkansas district during the 2025-26 school year. Students in the classrooms of 18 volunteer teachers were matched by propensity score, within grade and school, to students in other classrooms. On the state summative literacy assessment given in the spring, matched LitLab students scored 6.38 scale points higher than comparison students (Hedges' g = 0.38, 95% confidence interval [CI] 0.15 to 0.60; 410 students in 41 classrooms) after just under 14 weeks of use at roughly half the planned dose. The estimate held under an alternative model specification and against an unmeasured confounder that raised the odds of treatment by half, though not one that doubled them. We detected no moderation by grade, gender, race and ethnicity, English learner status, special education status, or baseline readiness, with low power to do so. The tool's ASR accuracy scores agreed closely with human raters (intraclass correlation = 0.953). An accuracy-weighted reading rate we derived from the platform's practice data correlated moderately with the state test (r = .52), as expected for brief, untimed practice recordings on texts matched to each student's phonics skill; these practice metrics are not standardized oral reading fluency measures and should not substitute for one. Because teachers selected into treatment, the estimate may overstate the tool's effect, and a randomized trial is warranted.

VERSION: November 2026

Suggested citation: Liu, Yun-En, and Melina Uncapher. (2026). Listening at Scale: A Pre-Registered Quasi-Experimental Evaluation of an Automatic Speech Recognition Reading Tool in Kindergarten Through Grade 2. (EdWorkingPaper: 26-1655). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/n43k-cr86

Listening at Scale: A Pre-Registered Quasi-Experimental Evaluation of an Automatic Speech Recognition Reading Tool in Kindergarten Through Grade 2

Yun-En Liu
Melina Uncapher
SETA-ED (Scientific and Engineering Technical Assistance for Education)

November 2026

Abstract

Oral reading fluency develops through guided oral reading with corrective feedback, yet a teacher in Kindergarten through Grade 2 (K-2) can listen to only one child at a time. We evaluate LitLab, a supplemental tool that uses automatic speech recognition (ASR) to score children's oral reading and report results to teachers, in a pre-registered quasi-experimental study in nine elementary schools in one high-poverty Arkansas district during the 2025-26 school year. Students in the classrooms of 18 volunteer teachers were matched by propensity score, within grade and school, to students in other classrooms. On the state summative literacy assessment given in the Spring, matched LitLab students scored 6.38 scale points higher than comparison students (Hedges' $g = 0.38$, 95% confidence interval [CI] 0.15 to 0.60; 410 students in 41 classrooms) after just under 14 weeks of use at roughly half the planned dose. The estimate held under an alternative model specification and against an unmeasured confounder that raised the odds of treatment by half, though not one that doubled them. We detected no moderation by grade, gender, race and ethnicity, English learner status, special education status, or baseline readiness, with low power to do so. The tool's ASR accuracy scores agreed closely with human raters (intraclass correlation = 0.953). An accuracy-weighted reading rate we derived from the platform's practice data correlated moderately with the state test ($r = .52$), as expected for brief, untimed practice recordings on texts matched to each student's phonics skill; these practice metrics are not standardized oral reading fluency measures and should not substitute for one. Because teachers selected into treatment, the estimate may overstate the tool's effect, and a randomized trial is warranted.

Keywords: oral reading fluency; automatic speech recognition; educational technology; early literacy; quasi-experimental design; propensity score matching; pre-registration

Author note. This study was funded by Accelerate. The pre-analysis plan was registered on the Open Science Framework (OSF) on November 20, 2025 (https://osf.io/yzet5), before any outcome data were accessed. The study was reviewed by Solutions IRB and determined exempt under the Code of Federal Regulations, 45 CFR

46.104(d)(1) (approval November 13, 2025). Student data were provided by the developer under a data sharing agreement with SETA-ED; the developer holds a student data privacy agreement with the North Little Rock School District. The authors have no financial relationship with the developer, Koalluh, Inc. Correspondence: Dr. Melina Uncapher, melina@seta-ed.com.

Acknowledgments. We thank the North Little Rock School District and its K-2 teachers, students, and families, and the LitLab team for supplying platform data and supporting the study.

Data availability. Student data are held under data sharing agreements and are not publicly available.

1. Introduction

Forty percent of U.S. fourth graders scored below the Basic level on the 2024 National Assessment of Educational Progress (NAEP) in reading (National Center for Education Statistics [NCES], 2025), and the gap opens well before fourth grade. Oral reading fluency is the hinge. A student who decodes accurately and at a reasonable rate has attention left over to process meaning, but a student who does not spends that attention predominantly on the words. The instructional implication has been clear for two decades: children need repeated, guided oral reading practice with corrective feedback (National Reading Panel, 2000).

The constraint in implementing this insight is supply. Guided oral reading requires an adult listening to one child at a time, the scarcest resource in a Kindergarten through Grade 2 (K-2) classroom. A synthesis of reading programs for struggling elementary readers finds positive outcomes for one-to-one tutoring and positive but smaller outcomes for small-group tutoring (Neitzel et al., 2022). Educational technology has been proposed to fill the gap for two decades, with modest returns. Across 84 rigorous studies of more than 60,000 K-12 students, technology applications for reading produced an average effect of +0.16 standard deviations (Cheung & Slavin, 2012), and across 20 studies of struggling readers in Grades 1 through 6 the average was +0.14, with the strongest effects for small-group integrated applications closely tied to reading instruction (Cheung & Slavin, 2013). Most of the tools in those syntheses present text and record clicks. Very few listen.

Automatic speech recognition (ASR) changes what a tool can do here, because a system that transcribes a child reading aloud can score accuracy and rate on every attempt and route that signal to the teacher. The capability is new enough in classroom products that the evidence base is thin, and existing evaluations tend to share three weaknesses: they are small, they are run by the developer, and they measure outcomes with instruments the developer built. Those features are consequential. Across 645 studies drawn from 12 research syntheses, effect sizes were about twice as large in small studies and on researcher-made measures as in large studies and on independent measures, and quasi-experiments produced significantly larger effects than randomized trials (Cheung & Slavin, 2016).

This paper evaluates one ASR-based oral reading tool, LitLab, at district scale. The study has four features that address the weaknesses above. The outcome is the state's end-of-year

literacy assessment for Kindergarten through Grade 2, which neither the developer nor the evaluators designed. The study spans nine schools. The comparison group and the analytic sample were defined from pre-intervention data, and the pre-analysis plan was registered before any outcome data were accessed. And the study asks a second question that evaluations of such tools rarely ask: whether the tool's fluency metrics track an independent measure of literacy. That question matters because a teacher reading a dashboard makes instructional decisions on the tool's metric, and the state test arrives too late to inform them.

We find that LitLab students scored 0.38 standard deviations higher than matched comparison students on the spring state test. The tool's transcription agreed closely with human raters, and an accuracy-weighted reading rate we derived from its practice data correlated moderately with the state test, as expected given how the practice data are generated. Teachers chose whether to use LitLab, and we argue that the estimate describes the tool as used by teachers who chose it and may overstate the effect of the tool alone. Section 2 describes the intervention, Section 3 the pre-registered research questions, Section 4 the setting and data, and Section 5 the empirical strategy. Section 6 reports results and Section 7 discusses interpretation, limitations, and implications.

2. The Intervention

LitLab, developed by Koalluh, Inc., provides structured oral reading practice for K-2 students. A child reads a short text aloud; the platform transcribes the recording with an ASR pipeline built on third-party speech recognition models, scores each word against the target text, and reports accuracy and words per minute (WPM) as separate metrics. The platform's WPM counts only words read correctly, divided by the time from the first to the last correct word in the recording, so pauses before and after reading are not counted. Teachers can override the ASR's score for individual words; overrides change accuracy but not WPM. According to the developer, the platform does not label this rate words correct per minute (WCPM), so that its practice metrics are not read as a standardized oral reading fluency measure. Each page of a text is recorded separately, so a recording can be as short as 5 seconds. The practice texts are phonics-aligned decodable texts matched to each student's assigned phonics skill, which need not be the student's grade level. In this district they were aligned to the core phonics program used in every K-2 classroom, Wilson Fundations, so the tool is designed to connect the phonics

curriculum to connected-text reading. Teachers see results on a dashboard that reports student accuracy, rate, and skill scores.

Each student has an individual login. The expectation was that students would use LitLab during classroom time on school computers, working independently. The pre-analysis plan scheduled 15 to 20 minutes per session twice weekly, or 30 to 40 minutes per week over a 12-week window. According to teacher survey responses, most teachers assigned LitLab in small groups or at centers, usually while they worked with other groups; a few designated a period for the whole class to use it at once. The hypothesized mechanism is that the tool multiplies the number of oral reading attempts that receive immediate, word-level feedback, which should build fluency, and that the dashboard lets teachers direct their time toward the students who need it. The district paid nothing for LitLab licenses in the study year.

3. Research Questions

The pre-analysis plan registered five research questions, RQ1 to RQ5 (McCann et al., 2025). We retain its numbering.

RQ1 (Impact, confirmatory). Do LitLab classrooms exhibit greater gains than matched business-as-usual classrooms on district fluency and decoding measures over about 12 weeks? The plan named grade-specific fluency and decoding components of the state assessment as the outcome; for reasons given in Section 5.9, the confirmatory analysis uses the overall score, and the study window ran just under 14 weeks. The unit of intervention is the classroom, since each teacher implemented LitLab during literacy instruction, and the analysis is at the student level with standard errors clustered by classroom.

RQ2 (Dose-response). Within LitLab classrooms, what is the association between dosage and student outcomes?

RQ3 (Usability and feasibility). What do teachers report about the usability and feasibility of LitLab, and how is their dashboard use associated with student outcomes?

RQ4 (Heterogeneity). Do effects vary by English learner status, individualized education program (IEP) status, free or reduced-price lunch eligibility, baseline achievement quartile, gender, or grade? Because the baseline measure arrived as a two-level status, baseline achievement was analyzed as Ready versus Potential Risk.

RQ5 (Measurement validity). How strongly do LitLab's internal accuracy and rate metrics correlate with the independent outcome measure? The plan set a target correlation of $r \geq .70$ and specified a signal-to-noise screen on recordings.

4. Setting, Sample, and Data

4.1 Setting

The study took place in the North Little Rock School District in central Arkansas during the 2025-26 school year. According to 2026 Arkansas Department of Education (ADE) data, the district enrolls 7,039 students in prekindergarten through Grade 12 across 14 schools. Its student body is 59% Black or African American, 19% White, 15% Hispanic or Latino, 6% two or more races, 1% Asian, 0.3% American Indian, and 0.09% Native Hawaiian or Pacific Islander; 74.3% of students qualify for free or reduced-price lunch, 13.8% are designated "Special Education," 9.5% are English learners, and 7.6% are designated "Students with Disabilities."

Nine elementary schools took part. They differ substantially in composition and in incoming literacy readiness (Table 1): readiness on the prior year's state literacy screener ranges from 5% to 36% across schools, and free or reduced-price lunch eligibility from 33% to 99%, so the sample spans a wide band of school-level advantage within a single high-poverty district.

Table 1

Characteristics of Participating Schools

School	Students	% Black	% White	% Hispanic/ Latino	% FRL	% EL	% Ready
Amboy Elementary	416	66	9	18	93	16	17.5
Boone Park Elementary	372	78	5	8	98	5	10
Crestwood Elementary	466	23	56	10	33	4	36
Glenview Elementary	269	81	4	8	99	6	24
Indian Hills Elementary	398	46	36	7	47	6	22
Lakewood Elementary	427	42	42	8	48	2	28
Meadow Park Elementary	411	81	3	7	96	3	13
North Little Rock 6 Academy	220	81	6	6	96	4	5
Ridge Road Elementary	516	40	11	44	92	39	9

Note. Enrollment and demographics are 2025 Arkansas Department of Education data for Kindergarten through Grade 5. FRL = free or reduced-price lunch eligible; EL = English learner. Ready = percentage of students rated Ready on the 2024-25 Arkansas Teaching, Learning, and Assessment System (ATLAS) Kindergarten through Grade 3 Literacy Screener, averaged across those grades.

4.2 Teacher Recruitment and Treatment Assignment

In fall 2025, the district's Coordinator of Reading, Literacy, and English Language Arts (PreK-12) emailed an open invitation to all of the district's K-5 teachers, with a LitLab flyer linking to a sign-up form; the first response arrived November 14, 2025. Thirty staff responded, 23 of them K-2 classroom teachers. Twenty of those K-2 teachers completed onboarding, and 18 obtained consent from at least one family; those 18 teachers and their classrooms form the treatment group. The other respondents taught Grades 3 to 5 or were instructional coaches and fall outside this study. The number of teachers who received the invitation was not tracked, so we cannot report how many K-2 teachers declined or did not respond. Each step was the teacher's choice, which is the central threat to inference in this design (Section 7.2).

Treatment was defined at the classroom level from district rosters. Parents or guardians of students in LitLab classrooms had to opt in to the study. Students first used LitLab on January 7, 2026; according to the developer, about 20 consent forms arrived between January 8 and January 15, 2026, after use had begun. Students without consent could still use LitLab, but they were marked as opted out in the platform's rosters, their platform data were never transferred to the evaluators, and they were removed before any analysis. The comparison pool consisted of all other students in the district with classroom and demographic records. Because students moved between classrooms, or district rosters were out of date, a few classrooms contained both LitLab and comparison students. To prevent contamination, we labeled any classroom containing a student who received a consent form as a treatment classroom and removed comparison students in those classrooms (2.5% of comparison students).

4.3 Analytic Sample

Table 2 traces the sample from the initial rosters to the analytic sample. Students missing a valid beginning-of-year score (missing or marked "Insufficient to score") were excluded before matching: 3.3% of treatment and 3.6% of comparison students. Matching was run on all 1,568 remaining students (294 treatment, 1,274 comparison) and produced 214 matched pairs. Students missing the spring post-test were then removed from the matched groups, leaving an analytic sample of 410 students (208 LitLab, 202 comparison) in 41 classrooms: 18 LitLab classrooms, each taught by a different teacher, and 23 comparison classrooms in 7 of the 9 schools. District

rosters identify comparison classrooms but not their teachers, so comparison teachers are not counted separately.

Table 2

Sample Construction

Stage	Treatment	Comparison	Notes
Initial roster	374	1,395	304 consented, 70 unconsented
Exclude students missing classroom identifiers	371	1,356	304 consented, 67 unconsented
Label classrooms with any treatment student as treated	405	1,322	A few classrooms contained both groups, likely from student movement or roster errors
Exclude unconsented students and comparison students in treatment classrooms	304	1,322	Prevents contamination of comparison students
Exclude invalid beginning-of-year scores	294	1,274	Missing or “Insufficient to score”
Propensity score matching	214	214	Matched pairs for RQ1
Remove students missing the spring post-test	208	202	Analytic sample, $N = 410$

4.4 Measures

Outcome. The outcome is the overall scale score on the spring summative literacy test of the Arkansas Teaching, Learning, and Assessment System (ATLAS), administered April 14 to May 14, 2026. The study’s variable documentation records a score range of 1000 to 1105. It is the state’s end-of-year literacy assessment for Kindergarten through Grade 2 and was designed by neither the developer nor the evaluators.

Baseline achievement. The pre-test is the ATLAS beginning-of-year (BOY) literacy screener, administered August 11 to September 26, 2025 (four outliers as late as December 10, 2025). The district supplied it only as a two-level status, Ready or Potential Risk. In the study dataset, 49% of Kindergarten, 40% of Grade 1, and 50% of Grade 2 students were rated Ready.

Student characteristics. District administrative data, received March 11, 2026, supplied grade, school, classroom, age in months, gender, race and ethnicity, English learner status, and IEP status.

Usage. LitLab platform logs supplied each student’s session time and the duration of each oral reading recording, and each teacher’s dashboard time, from January to April 2026; we

counted sessions that started on or before April 13, 2026, the day before the state test window opened.

Teacher surveys. LitLab administered three surveys to the 18 treatment teachers, on January 16, March 13, and May 19, 2026, which returned 17, 14, and 16 responses. Items included the Net Promoter Score (NPS) and Product Market Fit (PMF) items described in Section 5.7, years in education, and planned weekly minutes of LitLab use. No survey reached comparison teachers.

Transcription audit. Two human raters, one from the evaluation team and one from the developer's staff, independently scored 200 recordings randomly selected from those collected during the study, binned where possible to sample evenly by race, gender, and grade, between May 7 and May 21, 2026 (Section 5.8).

Data quality was assessed by the ability to join LitLab rosters, district rosters, district demographic data, and test scores; students with missing values on fields required for an analysis were removed. Data were stored in an encrypted Amazon Web Services storage bucket.

4.5 Treatment and Comparison Conditions

Students in LitLab classrooms first used the tool on January 7, 2026, and the summative test window opened on April 14, 2026, just under 14 weeks later. School closures for snow and ice during the week of January 24-29, 2026, disrupted schedules and made dosage targets harder to reach.

All K-2 classrooms in both groups used Wilson Fundations, so the contrast is Fundations plus LitLab practice against Fundations alone. Treatment teachers were asked to use LitLab as a supplement and not to stop using the district's existing literacy software, Magpie. About 25% of treatment teacher survey responses reported a change in Magpie use, generally a reduction, and the remaining 75% reported no change; a school literacy coach reported no observed change in Magpie use in either treatment or comparison classrooms. These reports suggest that LitLab mainly displaced other independent small-group or center activities. Magpie did not share usage data, so this could not be measured.

On Survey 1, 10 of the 17 responding treatment teachers reported more than 10 years in education. Asked how many minutes a week they planned for each student to use LitLab, 15

reported 21 to 45 minutes. Business-as-usual instruction in comparison classrooms is characterized only by the shared curriculum and the literacy coach's report, and it is the most weakly evidenced component of this study.

5. Empirical Strategy

5.1 Matching

Because teachers selected into treatment, we constructed the comparison group by student-level propensity score matching with the MatchIt package in R (Ho et al., 2011). The propensity score model was a logistic regression of treatment on student characteristics (Equation 1). Matching was nearest-neighbor, 1:1 without replacement, with a caliper of 0.2 standard deviations of the logit of the propensity score (Austin, 2011) and common-support trimming, and was exact on grade and school (Stuart, 2010). In one Grade 1 school and two Grade 2 schools, nearly every teacher volunteered, leaving too few within-school comparison students. For each, we paired the school with a similar-sized school that had only comparison students and treated the pair as one school for matching; where several candidates existed, we chose the one yielding the most matches. The balance targets were an absolute standardized mean difference (SMD) below 0.10 and a variance ratio between 0.5 and 2.

$$\text{logit}\ \Pr(T_i = 1 \mid \mathbf{X}_i) = \alpha + \mathbf{X}_i'\boldsymbol{\beta} \qquad (1)$$

Here $T_i = 1$ if student i is in a LitLab classroom, and $\mathbf{X}_i$ contains gender, race and ethnicity indicators, age in months, IEP status, English learner status, and BOY status (Ready versus Potential Risk). Grade and school enter as exact-matching strata. The R specifications for every model in this section are listed in Appendix D.

Matching was run on the full baseline sample, and students missing the post-test were removed afterward, so the analytic groups are defined from baseline data and attrition can be computed by condition (Center for Research and Reform in Education, 2026).

5.2 Estimand and Outcome Model

The confirmatory estimand is the average treatment effect on the treated (ATT) for matched LitLab students. On the matched sample we fit a linear outcome model of the spring ATLAS score on the treatment indicator, a natural cubic spline (three knots) of the logit

propensity score, age in months, English learner status, IEP status, race and ethnicity, gender, school, grade, and BOY status (Equation 2).

$$Y_i = \beta_0 + \tau T_i + \mathbf{Z}_i'\boldsymbol{\gamma} + \varepsilon_i,\ \mathbf{Z}_i = [s(\ell_i), \mathbf{X}_i, \mathbf{G}_i, \mathbf{S}_i] \qquad (2)$$

Y_i is the spring ATLAS overall scale score, ℓ_i is the logit of the estimated propensity score from Equation 1, $s(\cdot)$ is a natural cubic spline with three knots, $\mathbf{X}_i$ is the covariate vector from Equation 1, and $\mathbf{G}_i$ and $\mathbf{S}_i$ are grade and school indicators. We estimated the ATT by g-computation (Equation 3) with the avg_comparisons function of the marginaleffects package (Arel-Bundock et al., 2024): the fitted model predicts each LitLab student's outcome with treatment set to 1 and to 0, and the ATT is the mean difference over the N_1 matched LitLab students in the analytic sample. This is regression standardization and is not doubly robust. Because Equation 2 has no treatment interactions, the ATT in Equation 3 equals τ.

$$\text{ATT} = (1/N_1)\ \Sigma\ [\hat{Y}_i(1) - \hat{Y}_i(0)], \text{ summed over matched LitLab students } (T_i = 1) \qquad (3)$$

The estimate includes every matched consented student in a LitLab classroom regardless of how much the student used the platform, so it is an intention-to-treat contrast among consented students, as pre-registered. RQ2 is the only analysis conditioned on dosage, and we report no treatment-on-the-treated estimate.

Students in the same classroom share a teacher and tend to score alike. Classroom is also the level at which LitLab was taken up, since each teacher chose whether to use it, and when treatment is assigned by cluster, standard errors should be clustered at that level (Abadie et al., 2023). We therefore use cluster-robust standard errors at the classroom level with the bias-reduced linearization (CR2) correction (Bell & McCaffrey, 2002; Pustejovsky & Tipton, 2018) and Satterthwaite degrees of freedom for the treatment contrast.

We express effects as Hedges' *g*, dividing the adjusted difference by the standard deviation of the spring ATLAS score across all students in the study dataset before matching (about 17.0 scale points), so that matching does not change the yardstick.[1]

[1]The What Works Clearinghouse convention divides by the pooled standard deviation of the two analytic groups (15.9), which gives $g = 0.40$ (95% confidence interval about 0.16 to 0.64). We report the smaller value; the conclusions do not change.

5.3 Robustness and Sensitivity

The pre-registered robustness check is an analysis of covariance (ANCOVA) on the matched sample, regressing the outcome on treatment and the same covariates without the propensity spline (Equation 4), with CR2 classroom-clustered standard errors and Satterthwaite intervals; τ is the ANCOVA estimate. The plan required that it agree in sign with the primary estimate and fall within 25% of its magnitude. A second check repeats the full procedure on the grade-specific ATLAS components the plan named, rematching each grade on its own component pre-tests, with Benjamini-Hochberg control of the false discovery rate across the nine grade-by-component tests (Benjamini & Hochberg, 1995). Those samples are not merely subsets of the RQ1 matched sample, and their estimates are not directly comparable to the grade-specific estimates in Section 6.5.

$$Y_i = \beta_0 + \tau T_i + \mathbf{X}_i'\boldsymbol{\gamma} + \mathbf{G}_i'\boldsymbol{\varphi} + \mathbf{S}_i'\boldsymbol{\psi} + \varepsilon_i \qquad (4)$$

To assess sensitivity to unmeasured confounding, we computed Rosenbaum bounds with the sensitivitymw package (Rosenbaum, 2002, 2022). The parameter Γ is the factor by which an unmeasured confounder could multiply the odds of treatment among students matched on observed characteristics; the plan set a target of significance at $\Gamma = 1.3$.

5.4 Statistical Power

The pre-analysis plan did not register a minimum detectable effect size (MDES), so we report the achieved value. With CR2 classroom-clustered standard errors and Satterthwaite degrees of freedom, the achieved MDES for 80% power at a two-tailed α of .05 is 0.31 standard deviations. It is the standard error of the confirmatory contrast in standardized units (0.107) multiplied by 2.92, the sum of the .975 and .80 quantiles of a *t* distribution on that contrast's 23.4 degrees of freedom (Bloom, 1995). The binding constraint is the number of treatment classrooms, which is why the effective degrees of freedom are about 23 on a sample of 410. An MDES of 0.31 is nearly twice the average effect for reading technology in K-12 (Cheung & Slavin, 2012), so the study was not positioned to detect a field-typical effect. The dose-response, subgroup, and grade-specific analyses rest on smaller samples, with achieved MDES from about 0.37 in the largest grade cell to 0.83 in the thinnest race cell, so they could not detect effects in the small to moderate range. Null results in those analyses are absence of evidence at the available precision and should not be read as evidence of absence.

5.5 Dose-Response (RQ2)

The exposure is each LitLab student's total recording time, the summed duration of all oral reading recordings. We chose it as the best available measure of productive use; it is substantially shorter than session time, which also includes navigation, silent reading, other activities, and idle time. Let D_i be student i's total recording time in seconds. We estimated inverse probability weights for this continuous exposure (Equation 5) with a generalized linear model of recording time on gender, race and ethnicity, age in months, IEP status, English learner status, grade, and BOY status, so that recording time is independent of those covariates in the weighted sample. We then fit a weighted outcome model of the spring score on a natural cubic spline (three knots) of recording time and BOY status (Equation 6), with classroom-clustered standard errors, and estimated the average dose-response function (ADRF) by g-computation (Equation 7) with the adrftools package (Greifer, 2026). Recording time varies mostly between the 18 LitLab classrooms, so we report the curve descriptively without a significance test.

$$w_i = f(D_i) \,/\, f(D_i \mid \mathbf{X}_i, \mathbf{G}_i) \qquad (5)$$

$$Y_i = \beta_0 + g(D_i) + \gamma\ \text{Ready}_i + \varepsilon_i, \text{ weighted by } w_i \qquad (6)$$

$$\mu(d) = (1/N)\ \Sigma_i\ \hat{Y}_i(d) \qquad (7)$$

In Equation 5, $f(\cdot)$ is the marginal density of recording time and $f(\cdot \mid \mathbf{X}_i, \mathbf{G}_i)$ its conditional density from the generalized linear model; in Equation 6, $g(\cdot)$ is a natural cubic spline with three knots; and in Equation 7, $\hat{Y}_i(d)$ is student i's predicted score at recording time d, averaged over the $N = 285$ LitLab students in the dose analysis.

5.6 Heterogeneity (RQ4)

For each subgroup variable, we collapsed categories with fewer than 20 students in either condition into an "Other" category, added a treatment-by-subgroup interaction to the outcome model (Equation 8), and ran a joint Wald test of H_0: $\boldsymbol{\lambda} = 0$ with CR2 classroom-clustered errors and the approximate Hotelling's T-squared (HTZ) small-sample test (Pustejovsky & Tipton, 2018). Benjamini-Hochberg correction was applied across the six moderation tests.

Subgroup-specific ATT estimates and pairwise contrasts are exploratory and uncorrected, and the confirmatory RQ1 estimate is outside any correction family.

$$Y_i = \beta_0 + \tau T_i + \mathbf{M}_i'\boldsymbol{\theta} + T_i\mathbf{M}_i'\boldsymbol{\lambda} + \mathbf{W}_i'\boldsymbol{\gamma} + \varepsilon_i \qquad (8)$$

$\mathbf{M}_i$ contains the indicators for the levels of the subgroup variable being tested, and $\mathbf{W}_i$ the remaining covariates.

5.7 Teacher Acceptability and Dashboard Use (RQ3)

NPS was computed from the item "On a scale of 1-10, how likely are you to recommend LitLab to a friend or colleague?" as the proportion of responses of 9 or 10 minus the proportion of 1 through 6. PMF was the proportion of teachers answering "Very disappointed" to "How would you feel if you could no longer use LitLab?" (options: not, somewhat, or very disappointed). Confidence intervals are 95% intervals from a 1,000-sample bootstrap (Waskom, 2021). Teacher dashboard time takes one value per teacher, so the dashboard analysis rests on 18 classrooms, too few for a clustered dose-response test; we show students' post-test scores against their teacher's dashboard time without weighting, modeling, or significance testing.

5.8 Measurement Validity (RQ5)

Transcription audit. For each of the 200 sampled recordings, raters received the sentence the student was attempting to read and rated each word on a 0 to 3 scale that mirrors the platform's scoring (Table 3). Raters also transcribed extra words, from the student or from another speaker such as a teacher; these received no rating and did not count toward accuracy. Each recording's accuracy score for a rater was the proportion of target words rated 2 or 3, that is, recognizable as the intended word. We computed the mean absolute error (MAE) and the intraclass correlation, ICC(2,1), between the two raters, and between the platform's score and the mean of the two human scores. For example, if a student reading "The fan is pale" produced "The f, fa, fan is pole. What?" with noise obscuring "The," one rater might score 0.75 and the other 1.0, and the platform, scoring "The" as deleted and "pale" as replaced, 0.50; the human-human MAE is then 0.25 and the platform-human MAE 0.375. Appendix E gives the full scoring procedure, this worked example as a table, and the MAE and ICC definitions.

Table 3

Word-Level Rating Scale for the Transcription Audit

Rating	Interpretation	Equivalent platform label
0	Word completely missing	Delete
1	Word replaced with another word or grossly mispronounced (one phoneme replaced with another)	Replace
2	Word clearly understandable but enunciated incorrectly, or the recording is unclear (e.g., faint "-ed" in "reached")	Equals
3	Word pronounced correctly	Equals

Platform metrics against the state test. We restricted platform recordings to the end of the study (after March 1 and before the post-test cutoff of April 13, 2026) and to recordings passing the platform's signal-to-noise gate of at least 10 decibels, a marker of audio quality. For each recording we multiplied the WPM and accuracy values the developer supplied. Because WPM already counts only correct words, this product weights the rate by accuracy a second time, so we call it an accuracy-weighted reading rate and not WCPM. We averaged it across recordings within student, and correlated the student mean with the spring ATLAS score, against the pre-registered target of $r \geq .70$. This rate is not a metric LitLab reports, and it differs from WCPM as used in standardized oral reading fluency (ORF) assessment, including the ATLAS ORF component. Standardized ORF measures a one-minute cold read of a grade-level passage, so every student reads text of the same difficulty. LitLab recordings are untimed readings of single pages, some as short as 5 seconds, from decodable texts matched to each student's instructional phonics level, so the rate reflects both the reader and the difficulty of the assigned text. A moderate correlation with the state test is therefore expected.

5.9 Missing Data and Deviations From the Pre-Analysis Plan

Students with missing data were dropped without imputation. We define attrition as loss of the post-test among students otherwise eligible with a valid pre-test, the point at which a student could leave the study for reasons related to the intervention. Two smaller sources of missingness are distinct from attrition: invalid BOY scores (Section 4.3), and students who could not be joined across the LitLab roster, district roster, and demographic file. Within the matched sample, attrition was 2.8% for treatment (6 of 214) and 5.6% for comparison students (12 of 214), 4.2% overall (18 of 428), a differential of 2.8 percentage points. What Works

Clearinghouse (WWC) attrition standards govern randomized and regression discontinuity designs, so for a quasi-experiment the weight is carried by baseline equivalence on the analytic sample (What Works Clearinghouse, 2022); as a reference point, Evidence for ESSA (Every Student Succeeds Act) requires differential attrition of no more than 15 percentage points for a Strong or Moderate rating (Center for Research and Reform in Education, 2026). To check whether attrition was selective, we compared attrited and retained students within each condition on the propensity model covariates in the full eligible sample, since the matched sample contains too few attrited students for the comparison to be informative (Appendix Table A1).

The analyses depart from the registered plan in the following ways, each driven by a data condition that could not be anticipated at registration. We list every departure so readers can judge whether any could have favored a positive result.

Outcome measure. The plan named grade-specific ATLAS components as the confirmatory outcome. The components differ across Kindergarten, Grade 1, and Grade 2, so no single component exists in all three grades, and pooling requires a common metric. The confirmatory analysis therefore uses the overall ATLAS score, and the registered components are a robustness check (Section 5.3). This is the one departure that bears directly on the confirmatory finding.

Order of sample construction. The plan defines the analytic sample as students with both baseline and post-test scores and does not specify whether attriters are removed before or after matching. An earlier version of this analysis removed them before matching, which yielded 211 matched pairs and, under the same outcome model, $g = 0.45$ (95% CI 0.24 to 0.66). We reversed the order, as described in Section 5.1, so that groups are defined from the full baseline sample, as Evidence for ESSA requires (Center for Research and Reform in Education, 2026); the reordered estimate is $g = 0.38$.

Estimator. The plan specified doubly robust augmented inverse probability weighting after matching. We ran regression standardization by g-computation on an outcome model with a propensity score spline and the matching covariates (Section 5.2), which is not doubly robust and is not described as such.

Dose-response. The plan specified marginal structural models with inverse probability weights, using total platform minutes as a continuous exposure together with an

adequate-versus-inadequate dosage contrast. We used generalized linear model weights, a weighted outcome model, and g-computation, with all LitLab students included. The exposure is recording time, which excludes navigation and idle time, and the adequate-versus-inadequate contrast was not run. The adequate-dosage threshold was registered as 25 to 30 minutes per week in at least 8 of 12 weeks with at least one valid recording per week; we deviated from it, pre-committing 15 minutes per week for at least 6 weeks in the dated mid-year report to the funder, before summative outcome data were received.

Data that did not arrive. Magpie usage was registered as a covariate and as the basis for business-as-usual sensitivity analyses; Magpie declined to share usage data and district reports were not granular enough, so those analyses were not run. The registered fallback, a monthly teacher survey of both LitLab and comparison classrooms, reached treatment teachers only. Individual free or reduced-price lunch status was registered as a propensity model covariate and an equity subgroup and could not be used. The ATLAS interim assessment, registered as a secondary midline, was not analyzed. Core Phonics and Acadience, registered as secondary measures, were not available. The BOY screener arrived only as a two-level status, so the registered baseline composite scores could not be used.

Matching within school. The plan specified exact matching within school and grade; for three school-grade cells we used a paired comparison school (Section 5.1).

Targets not met. The plan set an effective sample size loss target below 20%; the achieved loss is 29%. The plan set a balance target of absolute SMD below 0.10; English learner status sits at 0.120 after matching and is adjusted for in the outcome model. The plan committed to a teacher survey nonresponse analysis, which was not performed (16 of 18 teachers responded to the final survey).

6. Results

6.1 Baseline Equivalence

Before matching, LitLab students were more often White, older, more often in Grade 2, and more often rated Ready at baseline than comparison students (Table 5). After matching, the groups were nearly identical on BOY status (SMD = −0.002) and differed by no more than 0.12 in absolute SMD on any characteristic (Figure 1); the largest gap, English learner status (0.120),

is adjusted for in the outcome model. Every characteristic with an absolute SMD above 0.05 is included in the outcome model. Because BOY status has only two levels, balance on it does not ensure that students started at the same reading level within each rating. Table 4 describes the analytic sample.

Table 4

Characteristics of the Analytic Sample

Characteristic	Overall	Comparison	LitLab
n	410	202	208
Kindergarten, *n* (%)	111 (27.1)	54 (26.7)	57 (27.4)
Grade 1, *n* (%)	139 (33.9)	68 (33.7)	71 (34.1)
Grade 2, *n* (%)	160 (39.0)	80 (39.6)	80 (38.5)
Age in months, *M* (*SD*)	86.63 (11.44)	86.17 (11.46)	87.08 (11.43)
Male, *n* (%)	194 (47.3)	98 (48.5)	96 (46.2)
Black, *n* (%)	229 (55.9)	116 (57.4)	113 (54.3)
White, *n* (%)	98 (23.9)	49 (24.3)	49 (23.6)
Hispanic/Latino, *n* (%)	47 (11.5)	21 (10.4)	26 (12.5)
Asian, American Indian/Alaska Native, Pacific Islander, or two or more races, *n* (%)	36 (8.8)	16 (7.9)	20 (9.6)
English learner, *n* (%)	32 (7.8)	12 (5.9)	20 (9.6)
IEP, *n* (%)	54 (13.2)	26 (12.9)	28 (13.5)
Ready on BOY screener, *n* (%)	217 (52.9)	107 (53.0)	110 (52.9)
Spring ATLAS score, *M* (*SD*)	1056.30 (16.22)	1053.07 (17.93)	1059.44 (13.71)

Note. Analytic sample after matching and attrition. *M* = mean; *SD* = standard deviation; IEP = individualized education program; BOY = beginning of year; ATLAS = Arkansas Teaching, Learning, and Assessment System.

Table 5

Covariate Balance Before and After Matching

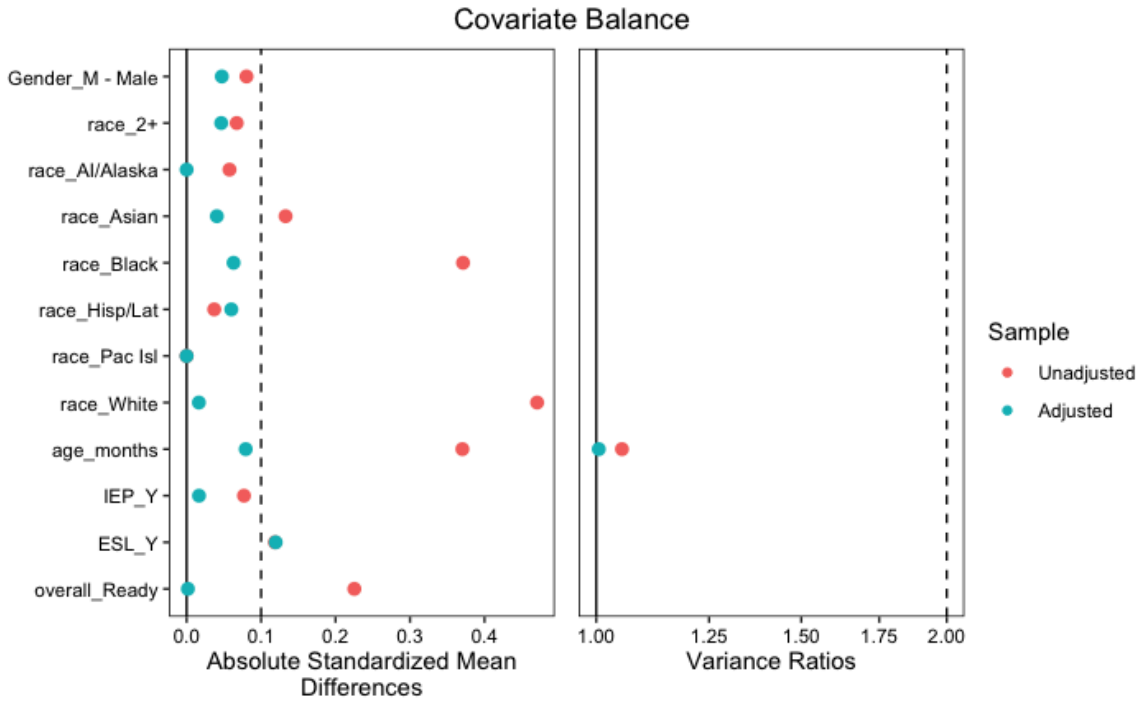
Variable	Before: Comparison	Before: LitLab	Before: SMD	After: Comparison	After: LitLab	After: SMD
Propensity score (logit)				−1.437	−1.424	0.016
Male	0.505	0.466	−0.078	0.485	0.462	−0.047
Two or more races	0.085	0.068	−0.063	0.074	0.087	0.046
American Indian/Alaska Native	s	0.000	−0.056	0.000	0.000	0.000
Asian	s	s	0.132	s	s	0.041
Black	0.610	0.429	−0.369	0.574	0.543	−0.063
Hispanic/Latino	0.148	0.133	−0.045	0.104	0.125	0.060
Pacific Islander	s	0.000	−0.040	0.000	0.000	0.000
White	0.149	0.350	0.478	0.243	0.236	−0.017
Age in months	83.671	87.724	0.350	86.168	87.077	0.079
IEP	0.164	0.139	−0.069	0.129	0.135	0.017

Variable	Before: Comparison	Before: LitLab	Before: SMD	After: Comparison	After: LitLab	After: SMD
English learner	0.126	0.085	−0.132	0.059	0.096	0.120
Ready on BOY screener	0.441	0.561	0.242	0.530	0.529	−0.002
Kindergarten	0.377	0.218	−0.353	0.267	0.274	0.015
Grade 1	0.326	0.296	−0.064	0.337	0.341	0.010
Grade 2	0.297	0.486	0.394	0.396	0.385	−0.024

Note. Values are means (proportions for binary variables). s = suppressed because the group has fewer than 10 students in that condition; standardized mean differences are shown. Before matching: all eligible students with a valid beginning-of-year (BOY) score (294 LitLab, 1,274 comparison). After matching: the analytic sample (208 LitLab, 202 comparison). SMD = standardized mean difference using the pooled standard deviation; IEP = individualized education program.

Figure 1

Absolute Standardized Mean Differences and Variance Ratios Before and After Matching



Note. Unadjusted = all eligible students before matching; Adjusted = matched sample. Dashed lines mark the pre-registered balance targets (absolute SMD of 0.10; variance ratio of 2). overall_Ready = Ready on the beginning-of-year screener.

Matching removed more LitLab students than attrition did: 78 had no comparison student within the caliper and 2 fell outside common support. The 80 unmatched LitLab students (27% of the 294) were mostly in Grade 2 (60 of 80), mostly White (53 of 80), and mostly in two schools (63 of 80; Table 6). The estimate therefore describes a group that is younger and more often Black than LitLab students overall.

Table 6

Matched and Unmatched LitLab Students

Characteristic	Matched (n = 214)	Unmatched (n = 80)
Kindergarten or Grade 1	131 (61%)	20 (25%)
Grade 2	83 (39%)	60 (75%)
Black, Hispanic/Latino, Asian, or two or more races	164 (77%)	27 (34%)
White	50 (23%)	53 (66%)
Ready on BOY screener	114 (53%)	51 (64%)
Schools A and G	103 (48%)	63 (79%)
Schools B to F	111 (52%)	17 (21%)

Note. Counts with column percentages for the 294 LitLab students who entered matching. Schools A and G are the two schools with the most unmatched students; Schools B to F are combined, and the two comparison-only schools are omitted. Categories are combined so that no cell has fewer than 10 students. BOY = beginning of year.

6.2 Impact on the Spring State Literacy Assessment (RQ1)

LitLab students scored 6.38 scale points higher on the spring ATLAS than matched comparison students (Table 7), an effect size of g = 0.38 (95% confidence interval [CI] 0.15 to 0.60; t = 3.5 on 23.4 degrees of freedom, p = .002; 410 students in 41 classrooms). The pre-registered ANCOVA gives +6.30 points (p = .002), a 1.3% difference from the primary estimate, so the registered robustness criterion is met. The estimate exceeds the achieved MDES of 0.31, so it is an effect this study could detect.

Table 7

Effect of LitLab on the Spring ATLAS Overall Score

Estimator	Difference (scale points)	n	df	t	p	g	SE (g)	95% CI (g)
G-computation ATT (primary)	6.38	410	23.4	3.5	.002	0.375	0.107	0.153, 0.597
ANCOVA (robustness)	6.30	410			.002			

Note. ATT = average treatment effect on the treated, estimated by g-computation on the matched sample with CR2 standard errors clustered by classroom (41 classrooms) and Satterthwaite degrees of freedom. Effect sizes use the pre-matching standard deviation of the outcome (about 17.0). ANCOVA = analysis of covariance; SE = standard error; df = degrees of freedom; CI = confidence interval.

The result is moderately robust to unmeasured confounding. The treatment effect remains significant at Γ = 1.5 (p = .01) and not at Γ = 2.0 (p = .267; Appendix Table A2). An unmeasured confounder that raised a student's odds of being in a LitLab classroom by half, beyond every characteristic already matched on, would not overturn the finding, while one that doubled those

odds could. The result clears the pre-registered target of $\Gamma = 1.3$ ($p = .004$). These bounds treat matched pairs as independent and do not account for classroom clustering, so they somewhat overstate robustness.

On the grade-specific components named in the pre-analysis plan, all nine point estimates are positive (g from 0.07 to 0.56), and none is statistically distinguishable from zero after Benjamini-Hochberg adjustment (Appendix Table A3). These analyses rematch each grade on component pre-tests and have matched samples of 101, 140, and 156 students, so they are uninformative about small and moderate effects. The confirmatory finding is an effect on overall literacy performance in a K-2 sample.

6.3 Implementation and Dose-Response (RQ2)

Students used LitLab about half as much as planned. Among the 285 LitLab students with a valid pre-test and post-test, median total session time was 256 minutes, roughly 18 to 20 minutes per week, students met the mid-year report's threshold of 15 minutes in a week for an average of 6.8 weeks, and 61% met it for at least 6 weeks (Table 8). Across all consented LitLab students, Kindergarten use was higher and far more variable than use in Grades 1 and 2 (Table 9). The week of weather closures accounts for part of the shortfall. The confirmatory effect was therefore observed at about half the intended dose.

Table 8

LitLab Session Time by Student Subgroup

Group	Subgroup	*n*	Median minutes	Mean weeks at minimum dose
Overall		285	256	6.84
Gender	Female	153	237	6.69
	Male	132	279	7.02
IEP	No	247	249	6.78
	Yes	38	331	7.26
English learner	No	260	262	6.98
	Yes	25	183	5.36
Race and ethnicity	Asian or two or more races	25	219	5.88
	Black	120	318	7.66
	Hispanic/Latino	39	237	5.97
	White	101	225	6.45

Group	Subgroup	*n*	Median minutes	Mean weeks at minimum dose
Grade	Kindergarten	61	410	8.52
	Grade 1	84	180	5.93
	Grade 2	140	237	6.66
BOY status	Potential Risk	125	235	6.54
	Ready	160	262	7.08

Note. Minimum dose = at least 15 minutes of total LitLab session time in a week, the threshold set in the mid-year report; the threshold is for total session time, not recording time. IEP = individualized education program; BOY = beginning of year.

Table 9

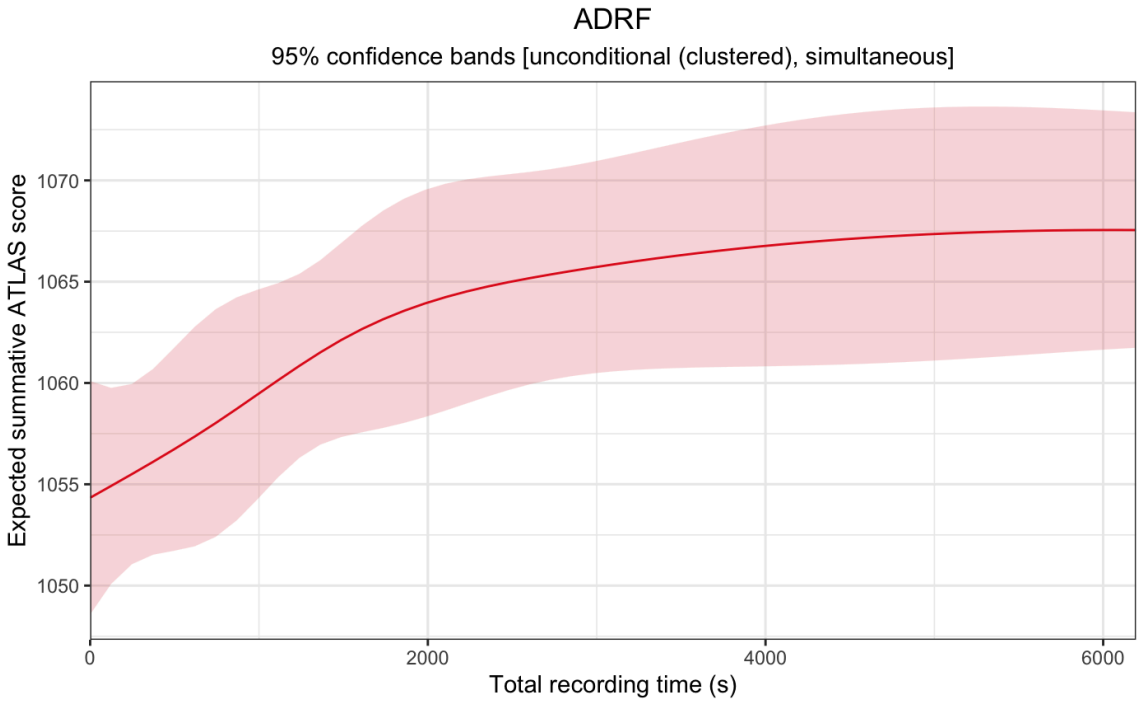
Total LitLab Minutes per Student, All Consented LitLab Students

Group	*M* (*SD*)	Median	Minimum	Maximum
Overall	325 (337)	255	0	2,583
Kindergarten	560 (578)	398	97.8	2,583
Grade 1	193 (111)	172	0	489
Grade 2	288 (166)	239	0	914

Scores rose with total recording time up to about 2,000 seconds (roughly 33 minutes) and leveled off after that (Figure 2; 285 students in 18 classrooms). Because recording time varies mostly between classrooms, students who recorded more were also in classrooms whose teachers implemented the tool more, and weighting on student characteristics cannot separate the two. We present the pattern descriptively.

Figure 2

Average Dose-Response Function of Spring ATLAS Score on Total Recording Time



Note. Inverse probability weighted g-computation estimate with 95% simultaneous confidence bands from classroom-clustered standard errors. Recording time is the summed duration of a student's oral reading recordings. ADRF = average dose-response function.

6.4 Teacher Acceptability and Dashboard Use (RQ3)

Teachers reported favorable views of LitLab in all three surveys (Table 10). NPS was 0.59, 0.57, and 0.75 and PMF was 0.53, 0.43, and 0.69 in January, March, and May. Practitioner sources commonly describe an NPS above 0.5 as excellent and above 0.8 as world-class, although Bain & Company, whose Fred Reichheld introduced the measure (Reichheld, 2003), publishes no fixed thresholds (Bain & Company, n.d.); a PMF at or above 0.40 is the conventional marker of product-market fit (Ellis & Brown, 2017). Every wave clears both thresholds.

Three qualifications apply. First, with 14 to 17 respondents per wave the intervals are wide: the second-wave NPS interval runs from roughly 0.22 to 0.86 and the second-wave PMF interval from roughly 0.21 to 0.71, each spanning its threshold, and wave-to-wave differences are not interpretable. Second, the NPS item used a 1 to 10 scale with detractors at 1 through 6, while the standard instrument uses 0 to 10 with detractors at 0 through 6 (Bain & Company, n.d.), so the values are not precisely comparable to published benchmarks. Third, both thresholds come

from commercial software benchmarking and have no validated K-12 referent, and the respondents are volunteer teachers surveyed by the developer. These results describe the tool's acceptability to an already favorable group and are not evidence of impact.

Table 10

Teacher Net Promoter Score and Product Market Fit by Survey Wave

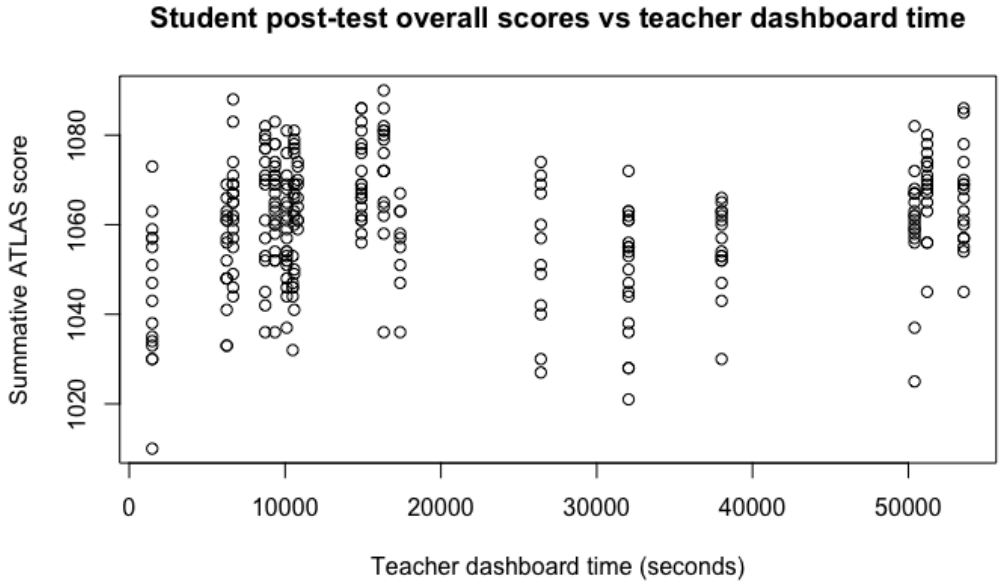
Survey wave	Date	Respondents	NPS	PMF
1	January 16, 2026	17	0.59	0.53
2	March 13, 2026	14	0.57	0.43
3	May 19, 2026	16	0.75	0.69

Note. NPS = Net Promoter Score (proportion of 9-10 responses minus proportion of 1-6 responses on a 1-10 scale); PMF = Product Market Fit (proportion answering "Very disappointed" if they could no longer use LitLab). Respondents are the 18 LitLab teachers.

Post-test scores vary widely within every classroom, and the plot of scores against teacher dashboard time shows no consistent trend across the 18 classrooms (Figure 3). With no weighting or modeling, the figure is not evidence about whether dashboard use affects outcomes.

Figure 3

Student Spring ATLAS Scores by Their Teacher's Total Dashboard Time



Note. Each point is a student; each vertical column of points is one LitLab classroom (18 classrooms). Descriptive only.

6.5 Heterogeneity (RQ4)

None of the six moderation tests was statistically significant after Benjamini-Hochberg correction (Table 11). The tests use all 41 classrooms, but with denominator degrees of freedom from 9.5 to 29 they have little power, so a null interaction is not evidence that the effect is uniform. Point estimates of the subgroup-specific effects were positive in every grade, both baseline groups, both genders, and each race and ethnicity group large enough to estimate (Black, White, and Hispanic/Latino students), and are reported with pairwise contrasts in Appendix Tables B1 and B2. Individual free or reduced-price lunch status was unavailable (Section 5.9), and the English learner, two-or-more-races, and Asian cells were too small for subgroup estimates.

Table 11

Tests of Effect Moderation

Moderator	*F*	Numerator *df*	Denominator *df*	Adjusted *p*
Grade	0.048	2	17.36	.953
English learner	4.578	1	9.52	.357
IEP	0.640	1	21.00	.649
Gender	0.656	1	29.06	.649
Race and ethnicity	1.070	3	19.55	.649
BOY status	0.167	1	28.63	.823

Note. Joint Wald tests of treatment-by-subgroup interactions with CR2 classroom-clustered standard errors and the HTZ small-sample correction; *p* values adjusted with the Benjamini-Hochberg procedure across the six tests. IEP = individualized education program; BOY = beginning of year.

6.6 Measurement Validity (RQ5)

The two human raters' accuracy scores were nearly identical (MAE = 0.031; ICC[2,1] = 0.978). The platform's ASR scores, compared with the mean of the two human scores, had an MAE of 0.067 and an ICC(2,1) of 0.953, approaching but not matching human-human agreement.

The accuracy-weighted reading rate correlated with the spring ATLAS score at $r = .52$ overall and .49 to .59 within grade (Table 12), below the pre-registered target of .70. The relationship is asymmetric (Figure 4): students with high rates almost all scored well on the state test, while students with low rates spanned nearly the full range of state test scores.

Table 12

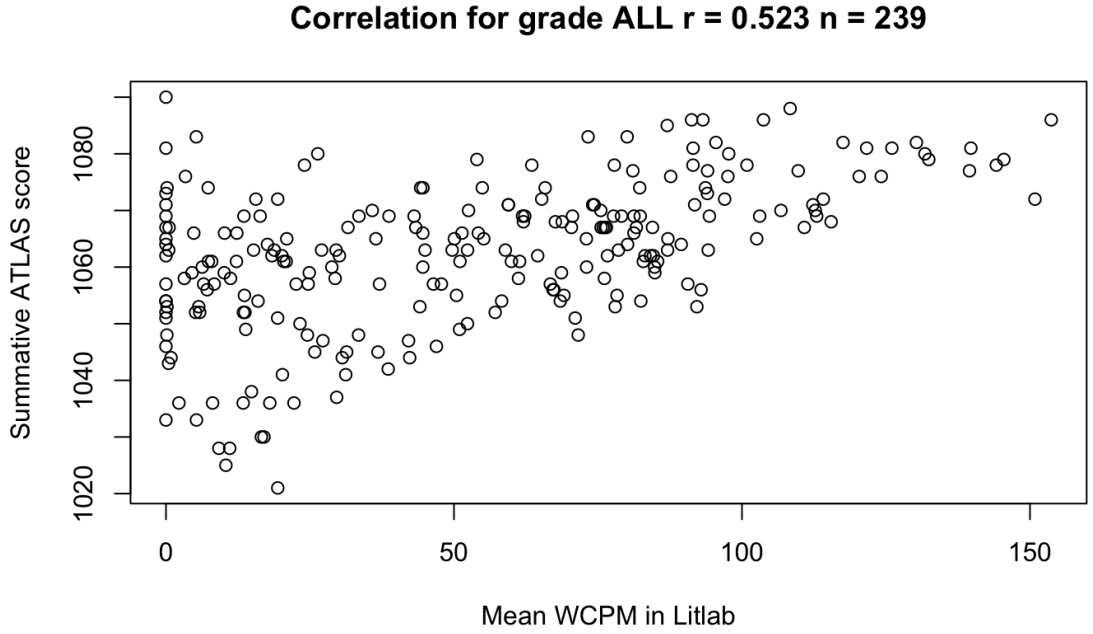
Correlation of the Accuracy-Weighted Reading Rate With the Spring ATLAS Score

Grade	*n*	*r*
All	239	.523
Kindergarten	50	.492
Grade 1	55	.515
Grade 2	134	.590

Note. Student mean of the accuracy-weighted reading rate (WPM × accuracy; Section 5.8) across recordings made after March 1 and before April 13, 2026 that passed the platform's 10-decibel signal-to-noise gate.

Figure 4

Spring ATLAS Score by Mean Accuracy-Weighted Reading Rate



Note. All grades; $r = .523$, $n = 239$. Rate = accuracy-weighted reading rate (WPM × accuracy; Section 5.8); the axis label in the original figure reads "Mean WCPM in Litlab."

7. Discussion

7.1 Magnitude

An effect of 0.38 standard deviations is large by conventional benchmarks. Kraft (2020) classifies effects of 0.20 or more as large for causal studies of education interventions measured on standardized achievement tests, the median effect in randomized education trials is about 0.10 (Kraft, 2020), and the average across rigorous evaluations of reading technology in K-12 is about 0.16 (Cheung & Slavin, 2012). Normative growth gives a second benchmark. Bloom et al.

(2008) estimate average annual reading growth of about 1.52 standard deviations from Kindergarten to Grade 1 and 0.97 from Grade 1 to Grade 2, so 0.38 corresponds to roughly two to three and a half months of typical early-grade reading growth. On the Education Endowment Foundation's Teaching and Learning Toolkit scale, which benchmarks against average progress over a school year without adjusting for age, 0.38 falls in the band of 0.36 to 0.44 that corresponds to five months of additional progress (Higgins et al., 2016). The grade-specific benchmark gives the smaller figure. Effect sizes run large in the early grades because growth itself is fast, a reason to read 0.38 as substantial without reading it as extraordinary. Still, an estimate more than twice the field average, from a supplemental tool used at about half its scheduled dose for just under 14 weeks, warrants scrutiny before it warrants confidence.

7.2 Teacher Selection and Unmeasured Confounding

The main threat to the estimate is that teachers chose to use LitLab. Matching balanced students on every characteristic available, but no teacher data were available for comparison classrooms. Teachers who volunteer may already spend more time listening to children read aloud and giving feedback, be more comfortable with classroom technology, be more motivated to try new practices, or run small-group and center routines that make independent practice productive. Each of these would raise their students' scores with or without LitLab, so each would push the estimate up. Parental opt-in adds a second, student-level selection layer: treatment students are those whose families returned a consent form, and opt-in is plausibly associated with home literacy support. Neither mechanism is observable in our data. Effect sizes from quasi-experiments also run larger than those from randomized trials (Cheung & Slavin, 2016).

The Rosenbaum bounds give a moderate margin. A confounder that raised the odds of treatment by half would not overturn the result; one that doubled them could, and teacher selection of that size is possible. Our reading is that the direction of the effect is supported, its magnitude may be overstated, and a recommended future direction should include a randomized trial to confirm.

7.3 What the Estimate Describes

Three features bound the estimand. It describes the 214 LitLab students who had a close match, 208 of whom took the spring test; the 80 unmatched students, mostly White Grade 2

students in two schools, are outside it. It is an intention-to-treat effect among consented students at about half the planned dose. And it is an effect on the overall ATLAS score; the registered grade-specific components are all positive but individually indistinguishable from zero at the precision available. The plain reading is that 0.38 describes LitLab as used by teachers who chose it, for roughly three quarters of LitLab students, in a district where every K-2 classroom used the same phonics program, and that the effect of the tool alone may be smaller.

7.4 The Platform's Metrics

The measurement results separate two questions that are easy to conflate. The ASR itself is accurate: its disagreement with the mean of two human raters (6.7%) approaches the disagreement between the raters themselves (3.1%). The accuracy-weighted reading rate correlated with the state test at .52, which is expected given the construct differences described in Section 5.8: the platform's recordings are short, untimed readings of texts matched to each student's phonics skill, and the state test's overall score includes, in Grades 1 and 2, a timed, cold read of grade-level text. The relationship is asymmetric. Students who read well on the platform generally did well on the state test, while a low practice score can reflect the difficulty of the assigned text or factors unrelated to reading, including a quiet voice, a noisy room, misunderstood instructions, a technical problem, or not taking the task seriously. LitLab's practice WPM and accuracy are not standardized ORF measures and should not substitute for one; their role is to support practice and progress monitoring within the platform's skill progression.

7.5 Limitations

Beyond teacher and parental selection, several limitations qualify the findings. The baseline measure was supplied only as a two-level status, so matching could not ensure similar starting reading levels within each rating; this is the main weakness after selection. Magpie usage could not be measured, matched, or controlled because Magpie declined to share usage data. In three school-grade cells, exact matching on school was replaced by a paired comparison school, so school was not fully controlled there. Some classrooms contained both groups, likely from student movement or roster errors; the 2.5% of comparison students sharing a classroom with a treatment student were dropped, and because the developer offered LitLab only to treatment teachers, residual contamination should be minimal. Attrition was low in both groups (97% of

matched treatment and 94% of matched comparison students had a post-test), but in the comparison group attrited students differed from retained students on race and BOY status (SMDs of 0.45 and 0.34); the analytic sample is balanced on both, and attrition related to unobserved characteristics cannot be ruled out. Teacher surveys were administered by the developer to volunteer teachers. Business-as-usual instruction in comparison classrooms was not surveyed. The study took place in one district with one product, and districts weighing this evidence should compare their own context with a mid-sized, high-poverty, majority-Black district in central Arkansas.

7.6 Implications for Policy and Practice

Three implications follow. First, LitLab's practice WPM and accuracy are not standardized ORF measures and should not substitute for one; districts should use them for practice and progress monitoring alongside their standardized fluency assessments (Section 7.4). Second, dosage expectations should be realistic. Students reached about half the scheduled time in a year with a week of weather closures, and the effect appeared anyway, at a dose a real literacy block can absorb; low dosage should not by itself be read as implementation failure. Third, on cost, the district paid nothing for licenses in the study year, and teacher time, which was not measured, was the real cost: setup and onboarding, facilitation during the literacy block, and dashboard time that platform logs show reaching several hours across the study window for the most active users. The pre-analysis plan's conditions for a cost-effectiveness estimate were not met, so we report none; districts estimating cost should treat teacher time as the dominant line.

For evidence standards, the study was designed to meet the ESSA Tier 2 (moderate evidence) criteria, and in our assessment it meets them: a comparison group defined from baseline data, 410 students in nine schools, a statistically significant positive effect on an independent state assessment, baseline balance on the analytic sample, a differential attrition of 2.8 percentage points, more than 10 weeks of use, and no statistically significant negative effect on any outcome, subgroup, or grade. The assessment is ours; the study has not been reviewed by the WWC or Evidence for ESSA.

7.7 Future Research

The results justify a randomized controlled trial that separates the tool's effect from the teachers who chose to use it, ideally with more English learners, students with disabilities, and students in the race and ethnicity groups too small to analyze here, with individual free or reduced-price lunch data, and with finer-grained baseline and outcome reading measures that could show which skills the tool trains best. Randomizing specific platform features could identify which components drive effects. Future studies could also assign delivery format, small-group or center rotation compared with whole-class use, by design; the present data cannot answer that question, because teacher survey responses were anonymized and could not be linked to classrooms, and few of the 18 treatment classrooms used whole-class delivery.

8. Conclusion

In a high-poverty Arkansas district, K-2 students whose teachers used an ASR-based oral reading tool for roughly 18 to 20 minutes a week over just under 14 weeks scored 0.38 standard deviations higher on the state summative literacy test than matched peers, a difference equivalent to roughly two to three and a half months of typical early-grade reading growth, or five months on the Education Endowment Foundation's scale, which does not adjust for age. The tool's transcription was accurate, and an accuracy-weighted reading rate we derived from its practice data tracked the state test moderately, as expected for short, untimed readings of skill-matched texts; the platform's practice metrics support practice and progress monitoring and are not a substitute for a standardized fluency measure. Because teachers chose to use the tool, this design cannot separate the tool from those teachers, and the estimate may overstate the tool's effect. A randomized trial is the next step to settle the size of the effect.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics, 138*(1), 1–35. https://doi.org/10.1093/qje/qjac038

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software, 111*(9), 1–32. https://doi.org/10.18637/jss.v111.i09

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10*(2), 150–161. https://doi.org/10.1002/pst.433

Bain & Company. (n.d.). *Measuring your Net Promoter Score*. Retrieved September 11, 2026, from https://www.netpromotersystem.com/about/measuring-your-net-promoter-score/

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology, 28*(2), 169–181.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*(5), 547–556. https://doi.org/10.1177/0193841X9501900504

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328. https://doi.org/10.1080/19345740802400072

Center for Research and Reform in Education. (2026). *Evidence for ESSA: Standards and procedures* (Version 3). Johns Hopkins University. https://www.evidenceforessa.org/wp-content/uploads/2026/08/Standards-and-Procedures-Update-FINAL-08-26-26.pdf

Cheung, A. C. K., & Slavin, R. E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*(3), 198–215. https://doi.org/10.1016/j.edurev.2012.05.002

Cheung, A. C. K., & Slavin, R. E. (2013). Effects of educational technology applications on reading outcomes for struggling readers: A best-evidence synthesis. *Reading Research Quarterly, 48*(3), 277–299. https://doi.org/10.1002/rrq.50

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

Ellis, S., & Brown, M. (2017). *Hacking growth: How today's fastest-growing companies drive breakout success*. Currency.

Greifer, N. (2026). *adrftools: Estimating, visualizing, and testing average dose-response functions* (R package version 0.1.0.9001) [Computer software]. https://github.com/ngreifer/adrftools

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A. B. V., Coleman, R., Henderson, P., Major, L. E., Coe, R., & Mason, D. (2016). *The Sutton Trust-Education Endowment Foundation teaching and learning toolkit* (December 2016 ed.). Education Endowment Foundation. https://durham-repository.worktribe.com/OutputFile/1606383

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

McCann, D., Sandberg, I., Uncapher, M., Fitzgerald, B., & Gulati, V. (2025, November 20). *Evaluating the impact of an AI-powered, curriculum-aligned reading platform (LitLab.ai) on student literacy in K–2 classrooms* [Preregistration]. OSF Registries. https://doi.org/10.17605/OSF.IO/YZET5

National Center for Education Statistics. (2025). *NAEP report card: Reading*. U.S. Department of Education, Institute of Education Sciences. https://www.nationsreportcard.gov/reports/reading/2024/g4_8/national-trends

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.

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2022). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly, 57*(1), 149–179. https://doi.org/10.1002/rrq.379

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–54.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer.

Rosenbaum, P. R. (2022). *sensitivitymw: Sensitivity analysis for observational studies using weighted M-statistics* (R package version 2.1) [Computer software]. https://CRAN.R-project.org/package=sensitivitymw

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

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Handbooks

Appendix A. Attrition, Sensitivity, and Grade-Specific Outcomes

Table A1

Standardized Mean Differences Between Attrited and Retained Students, by Condition

Variable	LitLab	Comparison
Grade	0.347	0.087
Gender	0.186	0.054
Race and ethnicity	0.764	0.452
Age in years (district record)	0.203	−0.123
IEP	0.487	0.169
English learner	0.439	0.079
Age in months	0.409	−0.039
BOY status	0.012	0.340

Note. Computed on the full eligible sample before matching (294 LitLab, 1,274 comparison). Only 9 LitLab students were lost to attrition in this sample, so the LitLab column is unreliable. IEP = individualized education program; BOY = beginning of year.

Table A2

Rosenbaum Sensitivity Analysis

Γ	Trimmed *t* test *p* value
1.0	< .001
1.3	.004
1.5	.01
2.0	.267

Note. The pre-registered target was $p < .05$ at $\Gamma = 1.3$. Computed on the 196 matched pairs in which both students had a spring score.

Table A3

Effects on Grade-Specific ATLAS Components

Grade and outcome	Adjusted *p*	*g*	*SE*	95% CI
Kindergarten ($N = 101$; 50 of 64 treated retained; pre-test: Alphabet)				
Sound-symbol	.533	0.210	0.269	−0.440, 0.860
Decoding	.407	0.556	0.277	−0.116, 1.220
Encoding	.533	0.265	0.265	−0.377, 0.906
Grade 1 ($N = 140$; 69 of 87 treated retained; pre-tests: Decoding, Encoding)				
Oral reading fluency rate	.407	0.362	0.152	−0.015, 0.740
Decoding	.430	0.186	0.141	−0.164, 0.536
Encoding	.430	0.200	0.144	−0.157, 0.558
Grade 2 ($N = 156$; 77 of 143 treated retained; pre-tests: Oral Reading Fluency Rate, Encoding)				

Grade and outcome	Adjusted *p*	*g*	*SE*	95% CI
Oral reading fluency rate	.533	0.159	0.213	−0.320, 0.637
Decoding	.430	0.147	0.111	−0.102, 0.397
Encoding	.680	0.074	0.173	−0.316, 0.463

Note. Each grade was rematched on its own component pre-tests using the RQ1 procedure, so these samples are not merely subsets of the RQ1 matched sample. *N* is the effective sample size after matching and attrition; in Grade 2 the effective sample size loss is 46%, which shifts the estimand toward treated students with close matches. *p* values are adjusted with the Benjamini-Hochberg procedure across the nine tests. CI = confidence interval.

Appendix B. Subgroup Estimates

Table B1

Effects of LitLab by Subgroup (Exploratory)

Factor	Level	*n*	Estimate	*SE*	*t*	*p*	*g*	*SE* (*g*)	95% CI (*g*)
Grade	Kindergarten	111	7.409	2.538	2.919	.004	0.436	0.149	0.142, 0.729
	Grade 1	139	5.822	2.301	2.531	.012	0.342	0.135	0.076, 0.609
	Grade 2	160	6.218	2.104	2.955	.003	0.366	0.124	0.122, 0.609
English learner	No	378	7.003	1.364	5.136	< .001	0.412	0.080	0.254, 0.570
	Yes	32							
IEP	No	356	6.806	1.413	4.816	< .001	0.400	0.083	0.237, 0.564
	Yes	54	3.785	3.656	1.035	.301	0.223	0.215	−0.200, 0.646
Gender	Female	216	5.333	1.820	2.930	.004	0.314	0.107	0.103, 0.524
	Male	194	7.622	1.929	3.951	< .001	0.448	0.113	0.225, 0.671
Race and ethnicity	Black	229	8.021	1.759	4.559	< .001	0.472	0.103	0.268, 0.675
	Hispanic/Latino	47	2.331	3.940	0.592	.554	0.137	0.232	−0.318, 0.593
	White	98	3.875	2.723	1.423	.155	0.228	0.160	−0.087, 0.543
	Asian or two or more races	36							
BOY status	Potential Risk	193	7.215	1.928	3.742	< .001	0.424	0.113	0.201, 0.647
	Ready	217	5.703	1.811	3.150	.002	0.335	0.107	0.126, 0.545

Note. Estimates in ATLAS scale points; uncorrected for multiple testing. Blank rows had fewer than 20 students in a treatment or comparison cell and are not estimated. IEP = individualized education program; BOY = beginning of year; CI = confidence interval.

Table B2

Pairwise Contrasts Between Subgroup Effects (Exploratory)

Factor	Contrast	Estimate	*SE*	*t*	*p*	*g*	*SE* (*g*)	95% CI (*g*)
Grade	Grade 1 − Kindergarten	−1.587	3.444	−0.461	.645	−0.093	0.203	−0.492, 0.305
	Grade 2 − Kindergarten	−1.191	3.296	−0.361	.718	−0.070	0.194	−0.451, 0.311
	Grade 2 − Grade 1	0.397	3.121	0.127	.899	0.023	0.184	−0.338, 0.384
IEP	Yes − No	−3.021	3.923	−0.770	.442	−0.178	0.231	−0.631, 0.276
Gender	Male − Female	2.289	2.665	0.859	.391	0.135	0.157	−0.174, 0.443

Factor	Contrast	Estimate	*SE*	*t*	*p*	*g*	*SE* (*g*)	95% CI (*g*)
Race and ethnicity	Hispanic/Latino − Black	−5.690	4.309	−1.320	.187	−0.335	0.253	−0.833, 0.164
	White − Black	−4.145	3.258	−1.272	.204	−0.244	0.192	−0.621, 0.133
	White − Hispanic/Latino	1.545	4.797	0.322	.748	0.091	0.282	−0.464, 0.646
BOY status	Ready − Potential Risk	−1.511	2.649	−0.571	.569	−0.089	0.156	−0.395, 0.217

Note. Uncorrected for multiple testing; several cells are small and the study was not powered to detect differences of this size, so none of these contrasts is evidence for or against differential effects. IEP = individualized education program; BOY = beginning of year; CI = confidence interval.

Appendix C. Covariate Balance Within Subgroups

Figures C1 to C6 show mean differences before (unadjusted) and after (adjusted) matching within each level of each subgroup variable used in the heterogeneity analysis. Dashed lines mark ±0.10.

Figure C1

Balance Within Grade

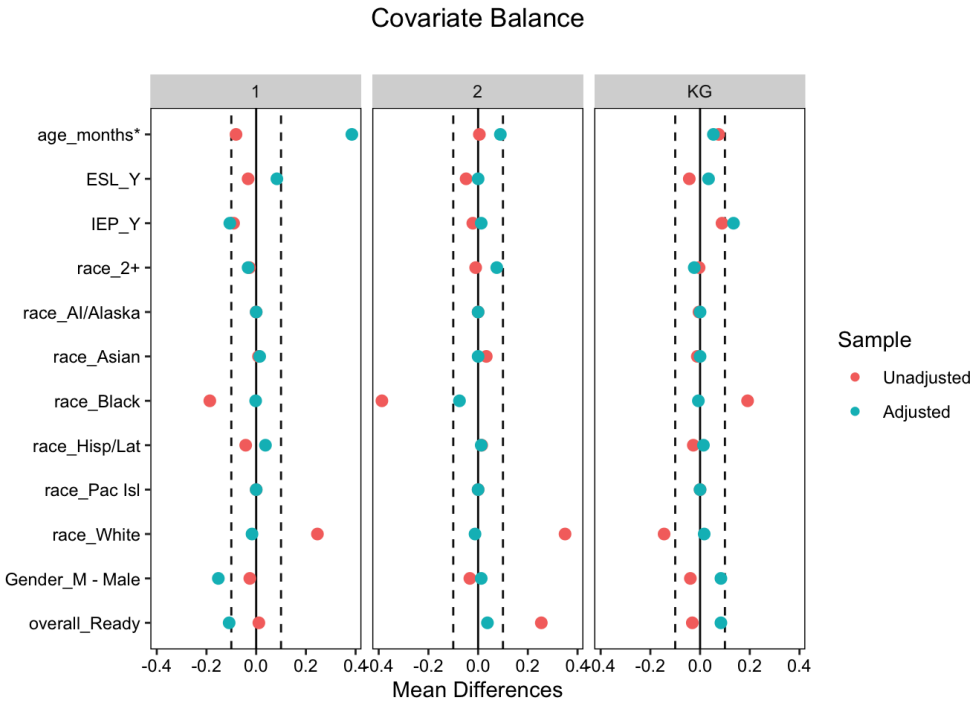


Figure C2

Balance Within English Learner Status

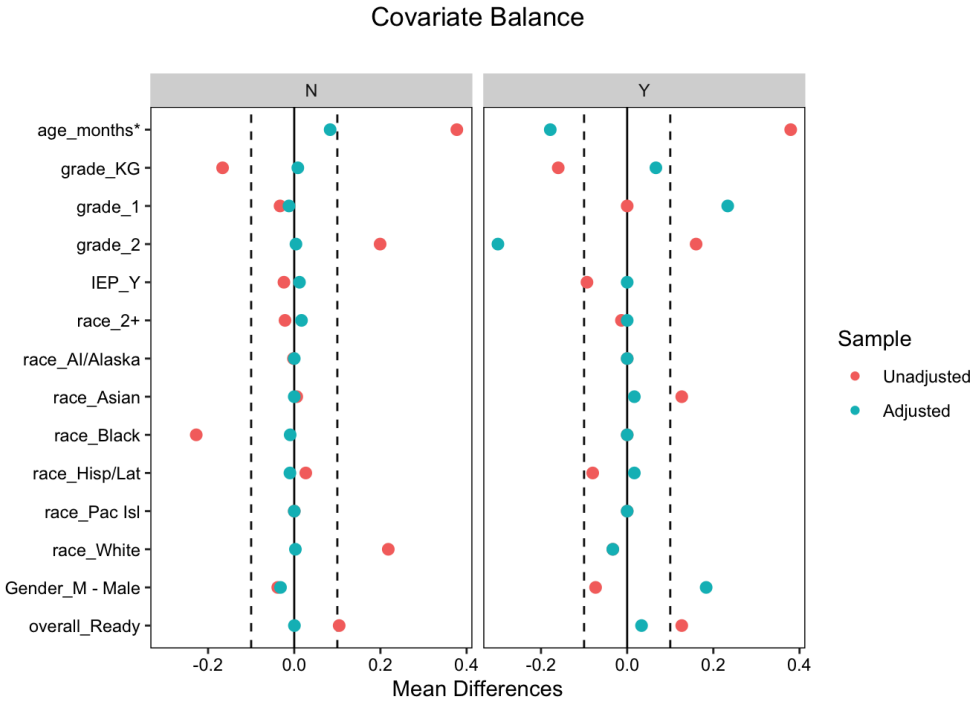


Figure C3

Balance Within IEP Status

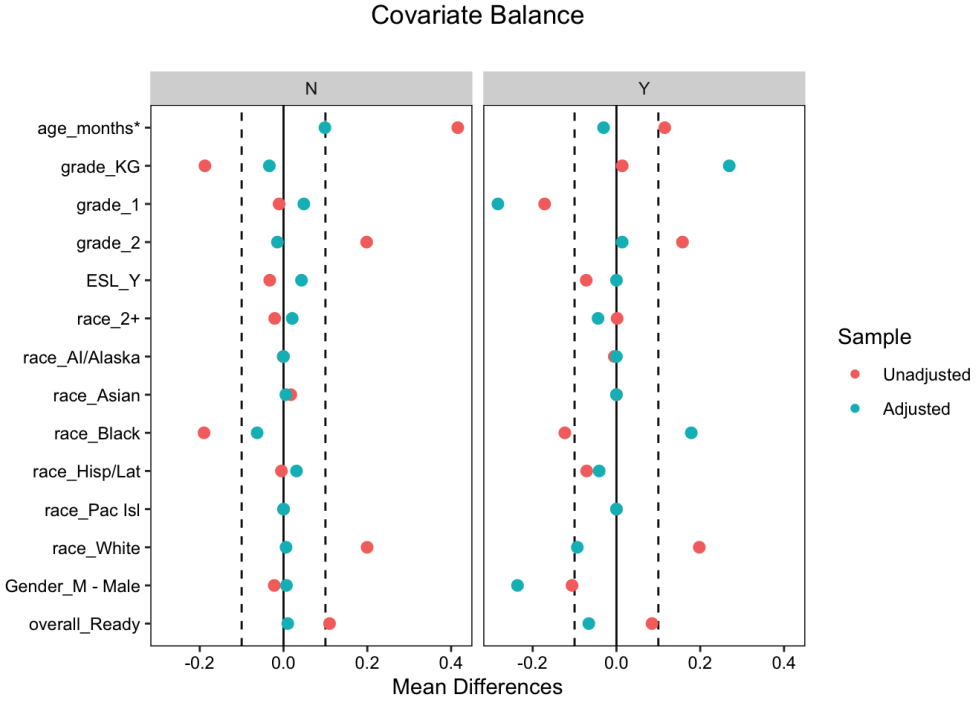


Figure C4

Balance Within Gender

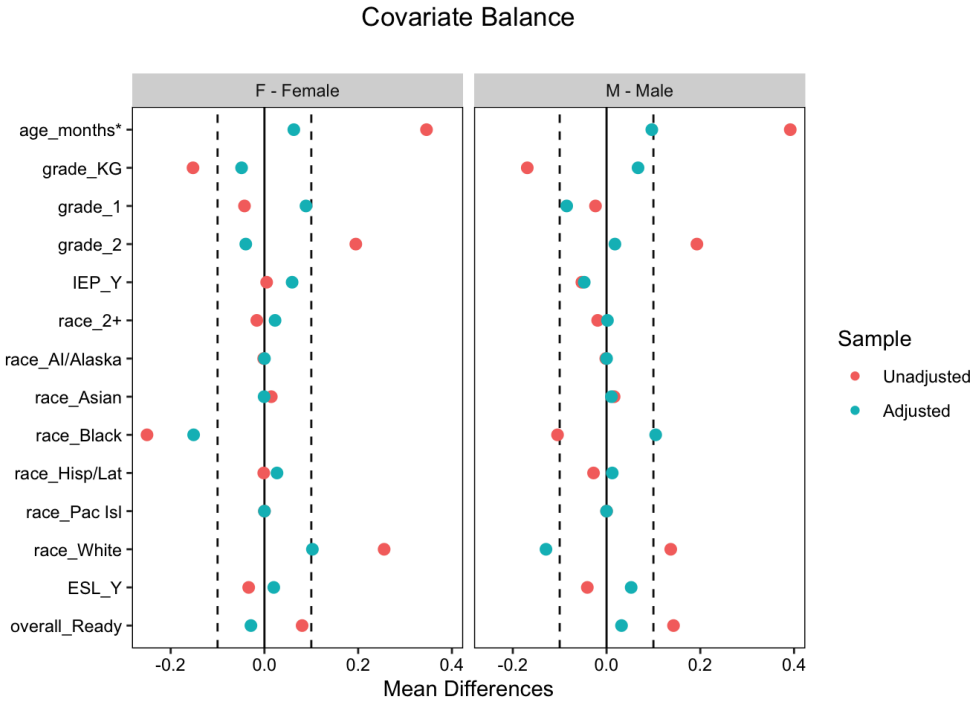


Figure C5

Balance Within Race and Ethnicity

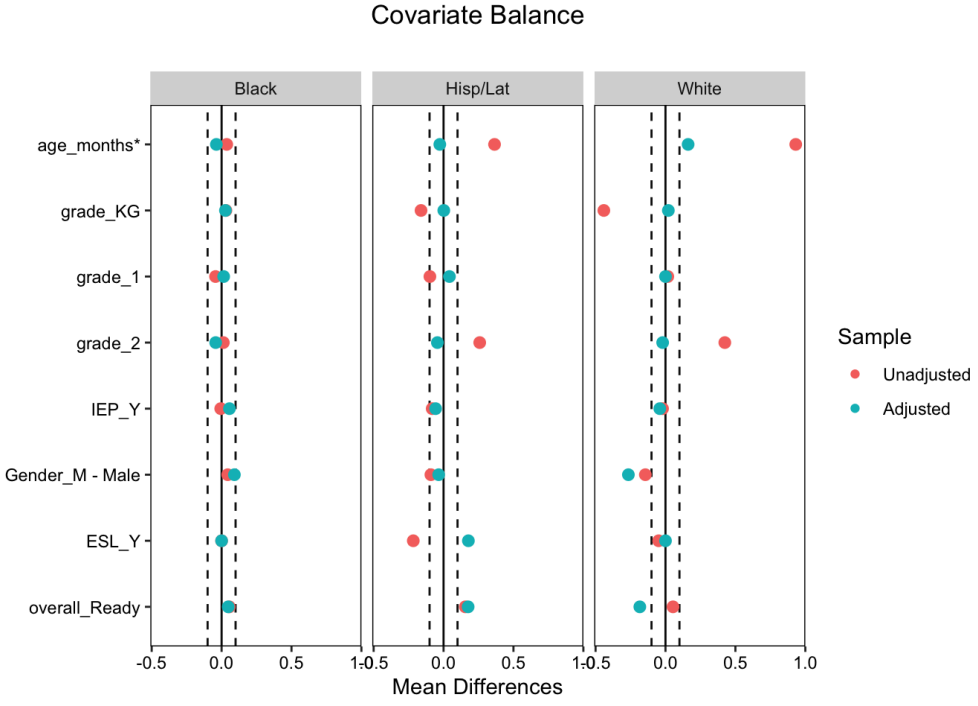
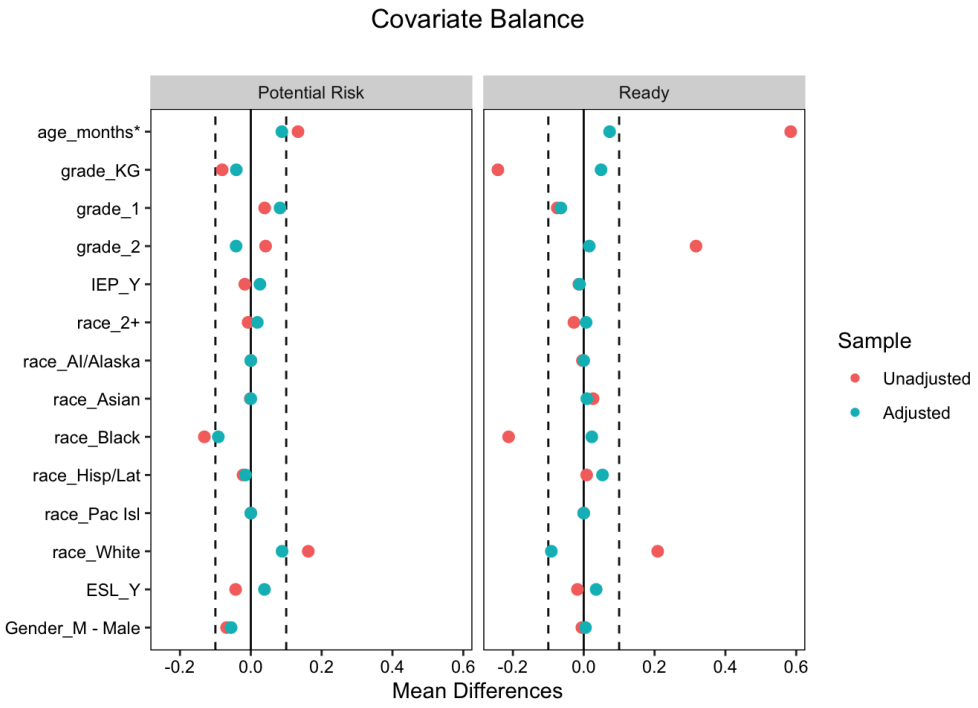


Figure C6

Balance Within Baseline (BOY) Status



Appendix D. Model Specifications

Table D1 lists each model as specified in R. In the formulas, pretest is the beginning-of-year (BOY) ATLAS status (Ready or Potential Risk), school is the school identifier after pairing the three fully treated school-grade cells with comparison schools, and ns() is a natural cubic spline.

Table D1

Model Specifications in R

Analysis	Specification	Estimation and inference
Propensity score and matching (RQ1)	treatment ~ Gender + race + age_months + IEP + ESL + pretest; exact = ~ grade + school	MatchIt: nearest neighbor, 1:1, without replacement, caliper = 0.2 SD of the logit propensity score, common-support trimming
Outcome model (RQ1, primary)	posttest ~ treatment + ns(logit propensity score, df = 4) + age_months + ESL + IEP + race + Gender + school + grade + pretest	marginaleffects::avg_comparisons on treated students (ATT by g-computation); CR2 standard errors clustered by classroom; Satterthwaite degrees of freedom
ANCOVA (RQ1, robustness)	posttest ~ treatment + age_months + ESL + IEP + race + Gender + grade + school + pretest	clubSandwich::vcovCR (CR2) clustered by classroom; Satterthwaite test

Sensitivity (RQ1)	Matched-pair outcome differences	sensitivitymw: trimmed *t* test at Γ = 1.0 to 2.0
Dose-response weights (RQ2)	recording_time ~ Gender + race + age_months + IEP + ESL + grade + pretest	Inverse probability weights for a continuous exposure, generalized linear model method
Dose-response outcome (RQ2)	posttest ~ ns(recording_time, df = 4) + pretest	Weighted; classroom-clustered standard errors; ADRF by g-computation with adrftools
Moderation (RQ4)	posttest ~ treatment * S + remaining covariates, for each subgroup variable S	Joint Wald test of interaction terms; CR2 by classroom; HTZ small-sample test; Benjamini-Hochberg across six tests

Note. ATT = average treatment effect on the treated; CR2 = bias-reduced linearization cluster-robust variance estimator; ANCOVA = analysis of covariance; ADRF = average dose-response function; HTZ = approximate Hotelling's *T*-squared test; SD = standard deviation.

Appendix E. Transcription Audit Procedure

Two human raters independently scored each of 200 recordings randomly selected from those collected during the study, using a procedure analogous to the one the platform's ASR pipeline uses to score recordings. For each recording, raters received the sentence the student was attempting to read and rated each word in that sentence from 0 to 3 (Table 3 in Section 5.8). These ratings were then compared with the platform's word labels. Raters also transcribed any additional words they heard, from the student or from another source such as a teacher speaking audibly in the recording; these extra words received no rating and did not count toward accuracy. Each recording received an accuracy score between 0 and 1 for each rater: the proportion of words in the target sentence rated 2 or 3, that is, clearly understandable as the word the student was meant to read. The platform's score is the proportion of words it labeled Equals.

As an example, consider the sentence "The fan is pale." Suppose the student said "The f, fa, fan is pole. What?" and the recording begins with noise, so "The" is hard to hear. Table E1 shows ratings the two humans and the platform might assign. Both raters and the platform would record the extra words, but those words do not enter the score.

Table E1

Worked Example of Word-Level Ratings and Recording Accuracy Scores

Rater	The	fan	is	pale	Accuracy score
Human 1	2 (Equals)	3 (Equals)	3 (Equals)	1 (Replace)	0.75
Human 2	3 (Equals)	3 (Equals)	3 (Equals)	2 (Equals)	1.00
Platform ASR	Delete	Equals	Equals	Replace	0.50

Note. Human ratings use the 0 to 3 scale in Table 3, with the equivalent platform label in parentheses.

We then computed two agreement statistics across the 200 recordings. The mean absolute error (MAE) between scores a and b is $\mathrm{MAE} = (1/n)\,\Sigma\,|a_i - b_i|$. For human-human agreement, a and b are the two raters' scores; in the example, $\mathrm{MAE} = |1.00 - 0.75| = 0.25$. For platform-human agreement, a is the platform's score and b is the mean of the two human scores; in the example, $\mathrm{MAE} = |(1.00 + 0.75)/2 - 0.50| = 0.375$. We also computed ICC(2,1), the intraclass correlation from a two-way random-effects model for absolute agreement of single ratings, for the same two comparisons.