Air Reading: A Randomized Evaluation of a Virtual Tutoring Model in Louisiana and Texas Schools

Amanda J. Neitzel, PhD, Nathan Storey, PhD,
Xue Wang, MA

December 2025

JOHNS HOPKINS
UNIVERSITY

Air Reading: A Randomized Evaluation of a Virtual Tutoring Model in Louisiana and Texas Schools

Amanda J. Neitzel, PhD, Nathan Storey, PhD, and Xue Wang, MA

Center for Research and Reform in Education
Johns Hopkins University School of Education
2800 N. Charles St
Baltimore, MD 21218
https://education.jhu.edu/crre

Preferred citation: Neitzel, A.J., Storey, N., & Wang, X. (2025). *Air Reading: A randomized evaluation of a virtual tutoring model in Louisiana and Texas schools.* Center for Research and Reform in Education, Johns Hopkins University.

INTRODUCTION

Overview of Air Reading

Air Reading is an assessment-driven virtual tutoring program designed to improve students' foundational reading skills in Kindergarten through 8th grade. Students receive live virtual instruction from consistent Air Reading tutors throughout the school year, addressing individual reading gaps. Small groups of up to four students are paired with the same highly-qualified tutor in 30-minute sessions four times each week.

Air Reading is grounded in the Science of Reading. Comprehensive, one-on-one diagnostics identify students' learning needs and inform group placement and bi-weekly assessments to track student progress. Highly qualified, paid Air Reading tutors deliver explicit, skill-based instruction using Air Reading's systematic reading curriculum. Ongoing training and support are provided to ensure high-quality teaching. All sessions are tracked and monitored on Air Reading's proprietary platform to maintain consistent standards. This model is scalable and can be replicated for schools regardless of their geography or staffing capabilities.

Overview of the Evaluation

Air Reading partnered with the Center for Research and Reform in Education (CRRE) to conduct an evaluation of Air Reading in a project funded by Accelerate and Arnold Ventures. The trial was conducted across the 2024-25 school year in a large suburban district in Louisiana as well as a district in Texas, and delivered tutoring for a full school year.

The present study used a randomized controlled trial to examine these research questions:

1. What is the effect of Air Reading on reading achievement for students performing below grade level, in comparison to similar students performing below grade level receiving business-as-usual teaching?
2. How do the effects of Air Reading differ by race, ethnicity, English learner status, special education status, economic status, and grade level?
3. To what extent is dosage received associated with better student outcomes?

METHOD

Research Design

This study employed a randomized controlled trial (RCT) to estimate the causal impact of Air Reading, a structured virtual tutoring program, on elementary students' literacy outcomes. Randomization occurred at the student level within schools and grade cohorts, ensuring that treatment and control students were directly comparable at baseline. Students assigned to treatment were offered Air Reading tutoring during the school day, while control students received business-as-usual literacy instruction.

The trial was conducted in a district in Texas and a large suburban district in Louisiana, and delivered tutoring for a full school year. This design allowed us to examine the program across distinct educational contexts.

Participants

The study included students from rural Texas schools and from a large suburban Louisiana district. The Texas schools were small and rural, serving predominantly Hispanic and economically disadvantaged students. The Louisiana district was much larger, with a racially and linguistically diverse population, including higher proportions of African American students and multilingual learners.

The sample included students in grades 1–4 randomized to either the Air Reading program (n = 174) or the business-as-usual control condition (n = 203).

Of these, the analytic sample consisted of 359 students, reflecting an overall attrition rate of 5 percent between baseline and endline. Attrition was balanced across conditions, with differential attrition of 3 percent. The flow of participants through randomization, assignment, and analytic inclusion is documented in Figure 1, which presents the study CONSORT diagram.

Figure 1.
CONSORT Diagram of Participant Flow

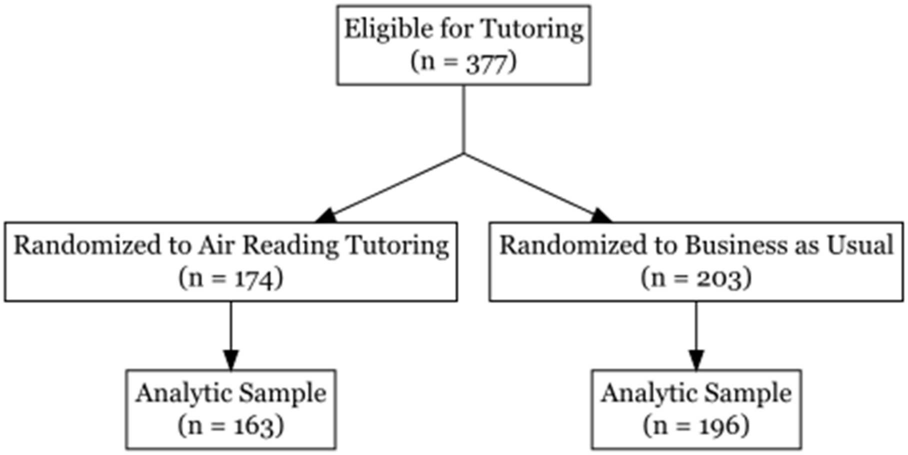

Table 1 provides descriptive statistics on student demographics and baseline achievement, pooled across the full sample and disaggregated by cohort and treatment status. Balance checks indicated no meaningful baseline differences between treatment and control groups overall or within cohorts.

Table 1
Baseline Characteristics of the Analytic Sample

		All	Control	Treatment
n		359	196	163
State N (%)	Louisiana	170 (47.4)	89 (45.4)	81 (49.7)
	Texas	189 (52.6)	107 (54.6)	82 (50.3)
Race/Ethnicity N (%)	Black	125 (34.8)	64 (32.7)	61 (37.4)
	Hispanic	179 (49.9)	104 (53.1)	75 (46.0)
	Other	8 (2.2)	3 (1.5)	5 (3.1)
	White	47 (13.1)	25 (12.8)	22 (13.5)
English learner N (%)		90 (25.1)	49 (25.0)	41 (25.2)
Special education N (%)		46 (12.8)	23 (11.7)	23 (14.1)
Economic disadvantage N (%)		272 (75.8)	152 (77.6)	120 (73.6)
Grade N (%)	1	67 (18.7)	38 (19.4)	29 (17.8)
	2	176 (49.0)	93 (47.4)	83 (50.9)
	3	69 (19.2)	41 (20.9)	28 (17.2)
	4	47 (13.1)	24 (12.2)	23 (14.1)
Gender N (%)	Female	198 (55.2)	111 (56.6)	87 (53.4)
	Male	161 (44.8)	85 (43.4)	76 (46.6)
Standardized baseline achievement		0.01 (0.99)	0.03 (1.04)	0.00 (0.93)

Intervention

Air Reading is a structured, virtual tutoring program designed to support the development of foundational literacy skills. The program is aligned to the Science of Reading. Tutoring is delivered synchronously, with consistent tutors working directly with students during the school day.

Each tutoring session lasted approximately 30 minutes and was scheduled four times per week. Students were grouped by instructional need, with group sizes ranging from 1:1 to 1:4. Within these groups, tutors followed structured lesson plans.

Tutors were required to meet specific professional qualifications. Minimum requirements included a bachelor's degree, prior teaching experience at the K–3 level, and state certification.

Measures

Two primary types of data were collected to assess program impacts: student achievement outcomes and program usage indicators.

Student achievement

In Texas, the Texas Primary Reading Inventory (TPRI) was administered in grades 1–2, and the Great Reading assessment in grade 3. In Louisiana, student performance was measured using DIBELS (grade 2) and the state's LEAP assessments (grade 4). To account for variation across sites and cohorts, outcome scores were standardized (z-scores), with dummy variables included for assessment type in pooled analyses.

Air Reading usage

Tutoring dosage and participation were tracked through Air Reading's internal system. The primary measure of usage was number of sessions attended, analyzed both as a continuous variable and as a categorical indicator of dosage. Students were classified as high dosage if they completed 56 lessons or more, with students below this level categorized as low dosage.

Covariates

Additional data were collected from school records, including demographics (race/ethnicity, gender, English learner status, special education status, and economic disadvantage) and baseline achievement levels. These variables were incorporated in subgroup analyses and controlled for in models estimating program effects.

Analytical Approach

To estimate the impact of Air Reading on student literacy outcomes, we used ordinary least squares (OLS) regression models with covariate adjustment. Randomization was conducted at the student level within schools and grade cohorts, so models included school fixed effects and blocking variables from the random assignment process. This approach provided unbiased estimates of treatment impacts while accounting for stratification used in the randomization. Treatment and control students demonstrated baseline equivalence, with all standardized mean differences well below the What Works Clearinghouse (2022) threshold of 0.25 standard deviations. Table 2 reports descriptive statistics and effect sizes for the analytic sample.

Table 2
Baseline Equivalence of Standardized Pretest Scores for Treatment and Control Groups, Analytic Sample

Analytic sample	All students	Treatment			Comparison			Effect size
	n	*n*	*M*	*SD*	*n*	*M*	*SD*	
Full sample	359	163	0.00	0.93	196	0.03	1.04	-0.03

Notes. 1. Effect sizes are experimental minus control means divided by the control group standard deviation.

Analyses controlled for students' baseline achievement as measured by pretest scores on the relevant district assessment. Additional covariates included gender, race/ethnicity, English learner status, special education status, socioeconomic disadvantage, and grade level. These controls improved precision and ensured that any residual imbalances after randomization were addressed. The general analytic model was specified as follows:

$$Y_i = \beta_0 + \beta_1(Treatment)_i + \beta_2(Pretest)_i + \beta_\mathrm{T}[Assessment]_i + \beta_K[Demographics]_i + \beta_\mathrm{L}[RandomizationBlocks]_i + e_i$$

where Y_i represents the student outcome (e.g. student achievement in reading) for student i, β_0 is the covariate-adjusted grand mean for the control group, β_1 is the average treatment effect, $(Treatment)_i$ is the binary treatment indicator at the student level, β_2 is the regression coefficient of the pretest, $(Pretest)_i$ is the pretest score, β_T is a vector of regression coefficients for the assessment type at pretest and posttest, $[\text{Assessment}]_i$ is a vector of dummy variables denoting the specific assessment at pretest and posttest, β_K is a vector of regression coefficients for student covariates, $[\text{Demographics}]_i$ is a vector of student demographic and grade level covariates, β_L, is a vector of regression coefficients for blocking variables used in randomization, $[\text{RandomizationBlocks}]_i$ is vector of blocking variables used in randomization (specifically school*grade level), and e_i is the residual for student i. All covariates were grand-mean centered to facilitate interpretation. Pretest and outcome scores were standardized to z-scores to allow for pooled analysis across different assessments (LEAP, TPRI, DIBELS). Dummy variables for assessment type were included to control for structural differences in scoring and scaling. This approach ensured comparability while preserving the integrity of site-specific data. All covariates were grand mean centered to facilitate interpretation of the intercept. All analyses were conducted using R statistical software (R Core Team, 2025).

We analyzed at the student level given that students' treatment condition was determined at the student level for the RCT. Robust standard errors were used to address heteroskedasticity and any clustering of data.

We examined impacts by the following subgroups: district, gender, ethnicity, English learner status, grade level, and socioeconomic status. The model from the impact analysis was adapted to analyze differential program impacts on student subgroups by adding interaction terms between student characteristics and the treatment indicator.

To assess the role of dosage, we adapted the primary model by replacing the treatment indicator with measures of attendance. Students were classified as high dosage if they completed at least 56 sessions. Additional analyses modeled dosage as a continuous predictor, measured as the proportion of assigned sessions attended.

To provide a comprehensive interpretation of the program's impact, several metrics were calculated, including effect size, additional months of learning, the improvement index, and tutoring efficiency. The effect size was calculated using Glass' delta, which standardized the difference in means using the standard deviation of the control group (Glass, 1976). The conversion from effect size to additional months of learning was based on annual growth estimates in reading (Hill et al., 2008). Using this benchmark, we translated effect sizes into equivalent months of learning to offer a more intuitive understanding of the program's impact. The improvement index, which quantifies the change in percentile rank for a student at the median of the control group, was calculated following the procedures outlined by the What Works Clearinghouse (What Works Clearinghouse, 2022). This index offers an accessible interpretation of the effect size by illustrating how a student's performance would shift relative to their peers. Finally, the concept of tutoring efficiency, as introduced by Kohlmoos and Steinberg (2024), was applied to quantify the number of hours of tutoring required to achieve one additional month of learning. This metric was crucial for assessing the cost-effectiveness and practical implementation of the tutoring program, offering a clear measure of the return on investment in terms of academic gains.

Procedure

Students were identified as eligible for tutoring by the schools by being flagged for either Tier 2 or Tier 3 intervention based on their fall assessment scores. The list of students was shared with Air Reading. Once eligible students were identified (students identified by their schools as performing below grade level in reading), they were randomized to the tutoring treatment group or the business-as-usual control group. Randomization was conducted by Air Reading using the *IndependentRandomizer* app developed by Amanda J. Neitzel. It allows for transparent and reproducible randomization while maintaining data security. At the conclusion of tutoring, data on student achievement, student demographics, and program dosage was merged and de-identified by Air Reading. The de-identified dataset was shared with the research team for analysis.

RESULTS

This section presents the estimated impacts of Air Reading on student literacy outcomes over the 24-25 school year. We begin with overall program impacts for the full sample as well as variation by student subgroups and dosage levels. We conclude with an analysis of program costs.

Program Impacts

Table 3 presents the estimated impacts of Air Reading on student literacy outcomes. Students assigned to Air Reading significantly outperformed their peers in the control condition by (*B* = 0.28, SE = 0.08, p < .05). This positive and statistically significant effect size of +0.29 standard deviations indicates that the program produced meaningful improvements in reading achievement overall.

In addition to traditional effect sizes, examining tutoring impact through metrics like tutoring efficiency and the improvement index provides a more nuanced understanding of Air Readings's effectiveness. For Air Reading, this effect size of +0.29 SD translates to approximately 2.8 months of additional literacy learning. The improvement index further contextualizes the program's impact. Air Reading's improvement index showed an increase of 11 percentile points for students in the treatment group relative to the control group. This metric translates the effect size into an easily interpretable gain, suggesting that students who participated in Ignite Reading moved up eleven percentile points, a meaningful shift, particularly in foundational literacy skills for first graders.

These findings suggest that Air Reading generated consistently positive effects on student literacy, with stronger and statistically reliable gains emerging in the full-year implementation

Table 3

Impact of Air Reading on Students' Academic Performance

N	Adjusted control mean	Adjusted treatment mean	Impact estimate (SE)	Effect size	Improvement index	Additional months of learning
359	-0.12	0.17	0.28* (0.08)	+0.29	+11	+2.8

Notes. 1. Effect sizes are experimental minus control means divided by the control group standard deviation. 2. The model also controlled for prior achievement, socioeconomic status, race/ethnicity, special education status, English language learner status, gender, cohort, grade level, and blocking variables used in random assignment. 3. SE = Standard Error. 4. * = significant at the .05 level.

Subgroup Analyses

To examine whether the effects of Air Reading varied for particular groups of students, we extended the main analytic models to include interaction terms between the treatment indicator and key student characteristics. These characteristics included grade level, gender, English learner status, special education status, economic disadvantage, race/ethnicity, and state. Statistically significant interaction terms were interpreted as evidence of differential program impacts.

While there were no significant differential impacts identified for any of the subgroups, some subgroup-specific impacts were noteworthy. For example, for students in Louisiana there was a substantial effect size of +0.32. Impacts were also larger for boys

(ES = +0.36) than girls (ES = +0.22) though this difference was not statistically significant. Detailed results, including effect sizes for each subgroup, are presented in Table 4. Across all subgroups, treatment students outperformed their control peers.

Table 4

Impact of Air Reading on Students' Academic Performance by Subgroups

Category	Student subgroup	Full sample			
		N	Adjusted treatment mean (SE)	Adjusted control mean (SE)	Effect size
Ethnicity	Black	125	0.047 (0.108)	-0.335 (0.125)	+0.39
	Hispanic	179	0.175 (0.101)	-0.052 (0.099)	+0.23
	Other	8	0.820 ** (0.264)	0.267 (0.347)	+0.56
	White	47	0.288 (0.155)	0.110 (0.187)	+0.18
English learner	Yes	90	0.228 (0.157)	-0.029 (0.135)	+0.26
	No	269	0.137 (0.075)	-0.01078	+0.29
Special education	No	313	0.182 ** (0.066)	-0.101 (0.062)	+0.29
	Yes	46	0.006 (0.201)	-0.266 (0.162)	+0.27
Economic disadvantage	Yes	272	0.180 * (0.076)	-0.088 (0.061)	+0.27
	No	87	0.096 (0.102)	-0.229 (0.166)	+0.33
Grade	1	67	0.613 (0.359)	0.304 (0.409)	+0.31
	2	176	-0.077 (0.149)	-0.393 (0.140)	+0.32
	3	69	0.438 (0.252)	0.206 (0.222)	+0.23
	4	47	-0.010 (0.360)	-0.200 (0.296)	+0.19
Gender	Female	198	0.135 (0.079)	-0.086 (0.080)	+0.22
	Male	161	0.191 (0.100)	-0.167 (0.094)	+0.36
State	Louisiana	170	0.023 (0.157)	-0.292 (0.187)	+0.32
	Texas	189	0.283 (0.173)	0.030 (0.153)	+0.26

Notes. 1. Effect sizes are experimental minus control means divided by the control group standard deviation. 2. The model also controlled for prior achievement, socioeconomic status, race/ethnicity, special education status, English language learner status, gender, cohort, grade level, and blocking variables used in random assignment. 3. SE = Standard Error.

Dosage Analyses

Table 5 summarizes the distribution of tutoring dosage across cohorts. On average, treatment students attended 55 sessions, with 56% reaching the high-dosage threshold of 56 or more sessions. Substantial proportions of treatment students in both cohorts fell below the high-dosage cutoff, underscoring variability in attendance.

Table 5

Air Reading Students' Dosage of Tutoring Received

		Control	Treatment
n		196	193
Dosage N (%)	High	0 (0.0)	90 (55.6)
	Low	0 (0.0)	72 (44.4)
	Control	196 (100.0)	0 (0.0)
Sessions attended mean (SD)		0 (0.0)	54.94 (15.26)

Table 6 reports the relationship between dosage and end-of-year achievement. Students who received high levels of tutoring exposure (56+ sessions) demonstrated significantly higher achievement than their control group peers (ES = +0.32, p < .05). In contrast, those in the low-dosage group showed significant but smaller impacts relative to controls (ES = +0.26, p < .05).

Table 6

Relationship Between Air Reading Dosage and End-of-Year Achievement

Level of dosage	Treatment N	Control N	Adjusted treatment EOY mean	Adjusted control EOY mean	Effect size
High (56+ sessions)	90	196	0.196	-0.118	+0.32
Low (0-55 sessions)	72	196	0.143	-0.118	+0.26

Notes. 1. Effect sizes are experimental minus control means divided by the control group standard deviation. 2. The model also controlled for prior achievement, socioeconomic status, race/ethnicity, special education status, English language learner status, grade level, and blocking variables used in random assignment.

Taken together, these results suggest that dosage may play an important role in mediating program impacts.

Cost Analysis

Cost analyses were conducted examining the implementation of Air Reading in both the Louisiana and Texas school district settings, conducted using the Accelerate cost analysis tool (Accelerate, 2024). Estimates from the cost analyses are seen in Table 8. In addition to the $1,500 per pupil cost charged to schools, provider-borne and school-borne costs related to personnel involved in tutoring delivery and operations, training and support, equipment and materials, and facilities, were estimated. The per-pupil program costs in each district were similar, roughly $1,700, in each site.

These estimates suggest that, based on actual costs to the society (the total per pupil cost of all inputs regardless of whether costs were paid or were in-kind provisions) and to the school (including hidden and opportunity costs), between 14.5 (Louisiana) and 17.8 (Texas) hours of tutoring were necessary to improve student learning by one month (tutoring efficiency). Based on these cost estimates, students received between 37.6 and 42.7 hours of tutoring per $1,000 per pupil. Higher figures here indicate more cost-efficient programs. Finally, these estimates indicate that between 2.1 and 2.9 additional months of learning may be gained on average by investing $1,000 per pupil. This measure of cost effectiveness relies on both the tutoring efficiency and cost efficiency estimates. In this case, Air Reading appeared to be somewhat less efficient in tutoring and slightly less cost effective in the Texas context compared to Louisiana.

Table 8
Cost Analysis of Air Reading in Louisiana and Texas Implementation Sites

	Louisiana school district		Texas school district	
	Cost to society	Cost to school	Cost to society	Cost to school
Metric	Actual price	Actual price	Actual price	Actual price
Tutoring efficiency	14.5	14.5	17.8	17.8
Cost efficiency	37.6	42.7	37.6	42.7
Cost effectiveness	2.6	2.9	2.1	2.4

Considering both analyses, these results suggest that Air Reading, a high dosage tutoring intervention, has the potential, as seen in both Louisiana and Texas contexts, to be both efficient and cost effective. Further evidence of implementation cost estimates in other contexts and representing variations in the tutoring model may provide additional understanding of the efficiency and effectiveness of the Air Reading program. In addition, these estimates focused on actual costs incurred by the provider and school/district. Further examinations may benefit from greater examination of costs estimated in local and national expense contexts.

DISCUSSION

Summary of Key Findings

Across two districts in a full year implementation, Air Reading produced consistently positive impacts on early literacy outcomes. Students assigned to receive tutoring significantly outperformed their control group peers, with an effect size of +0.29 SD.

Subgroup analyses showed that the benefits of Air Reading were broadly distributed across student groups. No systematic differential effects were detected by grade level, race/ethnicity, English learner status, special education status, grade, or economic disadvantage.

Dosage analyses produced a more nuanced picture. Students who reached the high-dosage threshold of 56 or more sessions showed significantly greater achievement gains compared with their control peers, whereas low-dosage students demonstrated smaller, though still significant, achievement gains compared with control peers.

Taken together, these findings indicate that Air Reading can produce meaningful improvements in early literacy, particularly when implemented across a full school year. At the same time, the absence of strong subgroup variation suggests that the program's effects are not narrowly concentrated but instead benefit a broad range of students.

Interpretation and Implications

The findings provide several insights into the conditions under which virtual tutoring may be most effective. First, the stronger impacts observed in the present evaluation of a full year implementation compared with a prior study of a semester-long implementation (Neitzel et al., 2025) highlight the importance of program duration and exposure. Whereas the one-semester implementation yielded only modest gains (ITT ES = +0.12), the full-year implementation produced larger and statistically reliable effects (ITT ES = 0.29). This pattern underscores that sustained exposure is likely necessary for tutoring to produce meaningful literacy gains, particularly in early grades where skill development is cumulative and foundational.

Second, the general consistency of results across subgroups points to the broad accessibility of virtual tutoring. The absence of systematic differential effects for English learners, students with disabilities, and economically disadvantaged students suggests that the Air Reading model can support diverse learners. This has important policy implications: tutoring may serve as a scalable intervention that does not exacerbate existing inequities but rather provides equitable opportunities for improvement. The one exception, gender, warrants additional attention. While male students appeared to benefit more in the pooled sample, this pattern was not consistent across cohorts, suggesting it may reflect sampling variability or context-specific dynamics rather than a robust program effect.

Finally, the results highlight the flexibility of virtual tutoring as a delivery model. Schools and districts were able to integrate tutoring into daily schedules with varying levels of intensity and across different grade spans. While logistical barriers, such as session cancellations due to school-level conflicts, remain a challenge, the model demonstrates adaptability that is critical for scaling interventions in diverse educational contexts.

Limitations

While this study provides rigorous evidence on the impacts of Air Reading, several limitations should be acknowledged. First, although attrition was low overall and did not differ meaningfully between treatment and control groups, some student loss between baseline and endline may introduce bias if attrition was systematically related to outcomes.

Second, generalizability is constrained by the study's settings. The study took place in a single Texas district, and in a single Louisiana district. Although both are relatively large systems, results may not extend to districts with different resource levels, demographics, or implementation capacities.

Finally, like most randomized controlled trials, this study focused on short-term achievement outcomes. Longitudinal analyses will be needed to assess the persistence of impacts over time, as well as potential spillover effects on broader measures of literacy development and academic engagement.

Future Directions and Conclusion

The findings from this study add to a growing body of evidence that virtual tutoring can be an effective and scalable approach to improving early literacy. Air Reading produced consistently positive impacts. These outcomes are promising for districts and states seeking to embed tutoring into their instructional recovery and long-term literacy strategies.

At the same time, several avenues for future research remain. First, future work should investigate the persistence of impacts beyond a single school year, including long-term outcomes in later grades such as reading fluency, comprehension, and broader academic achievement. Second, more attention is needed to unpack the mechanisms that influence dosage, including both student-level factors (e.g., absenteeism, engagement) and school-level factors (e.g., scheduling conflicts, staffing). Understanding these dynamics will be critical for maximizing the efficiency and equity of tutoring programs.

Third, there is a need to better understand the counterfactual conditions faced by control students. In many districts, students not assigned to Air Reading may still have received other supports such as small-group instruction, afterschool programs, or interventions provided by classroom teachers. Distinguishing the specific value added by virtual tutoring relative to these alternatives is critical for understanding where tutoring is most effective and for which students it provides the greatest incremental benefit.

For policymakers, the results underscore that virtual tutoring can play a meaningful role in addressing longstanding literacy challenges, but also that sustained exposure and careful implementation are necessary to achieve significant gains. As states such as Massachusetts move to embed tutoring in ongoing budgets, and as districts grapple with sustaining tutoring after ESSER funds expire, the evidence from Air Reading points to both the promise and the challenges of scaling virtual interventions.

In conclusion, Air Reading demonstrates that virtual tutoring can produce significant improvements in early literacy outcomes across diverse student populations. The program's impacts were strongest in the context of a full-year implementation, suggesting that duration and consistency are key to success. As more data become available, and as future research addresses long-term outcomes, cost-effectiveness, and counterfactual supports, virtual tutoring has the potential not only to support pandemic recovery but also to advance the broader goal of ensuring that all students achieve foundational literacy.

REFERENCES

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

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

Kohlmoos, L., & Steinberg, M. P. (2024). *Contextualizing the impact of tutoring on student learning: Efficiency, cost effectiveness, and the known unknowns*. Accelerate. https://accelerate.us/wp-content/uploads/2024/05/Accelerate-Research-Report-Efficiency-and-Cost-Effectiveness-1.pdf

Neitzel, A. J., & Storey, N. (2024). *Air Reading: A randomized evaluation of a virtual tutoring model*. Center for Research and Reform in Education, Johns Hopkins University. https://jscholarship.library.jhu.edu/handle/1774.2/70119

R Core Team. (2025). *R: a language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

What Works Clearinghouse. (2022). *Procedures and Standards Handbook, Version 5.0*. Institute of Education Sciences, US Department of Education.