Air Reading: A Randomized Evaluation of a Virtual Tutoring Model

Amanda J. Neitzel, PhD, Nathan Storey, PhD

September, 2024

JOHNS HOPKINS
UNIVERSITY

Air Reading: A Randomized Evaluation of a Virtual Tutoring Model

Amanda J. Neitzel, PhD, Nathan Storey, PhD

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

Contents

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 semester, addressing individual reading gaps. Small groups of up to four students are paired with the same highly-qualified tutor in 40-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 implemented in the spring of the 2023-24 school year in a project funded by Accelerate. This report focused on findings from the randomized study that examined the impact of the program as implemented in rural district within Texas.

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

During the 2023-24 school year, six elementary schools in a district in Texas took part in a randomized controlled trial to evaluate the impact of Air Reading on reading outcomes for first through sixth grade students. The research was designed to follow procedures that could qualify it for "meets standards without reservation" in the What Works

Clearinghouse (What Works Clearinghouse, 2022). At all schools students were identified as needing additional reading intervention. These students were stratified based on scheduling slots, and then randomly assigned within the strata to participate in Air Reading (treatment) or continue with their business-as-usual classroom instruction (control). Following an intent to treat (ITT) approach, students were included in their assigned group, regardless of whether or not they received tutoring.

Participants

Details about study participants are presented below.

6 schools　　　　　418 1st – 6th graders　　　　　Rural district in Texas

Demographic snapshot of student participants

Race / Ethnicity

4%
21%
27%
48%

■ Black　■ Hispanic　■ White　■ Other

Characteristics

Students with disabilities

English learners

Economically disadvantaged

0%　20%　40%　60%　80%　100%
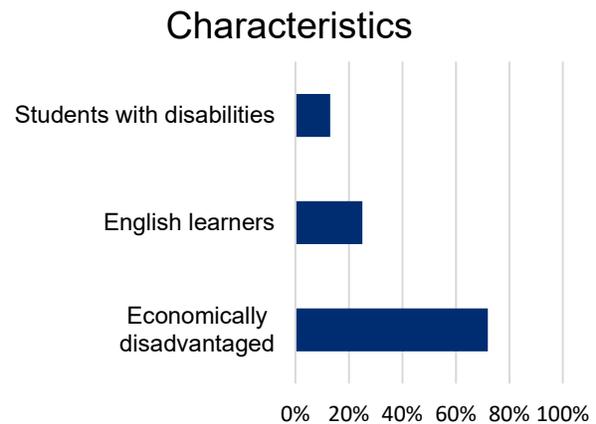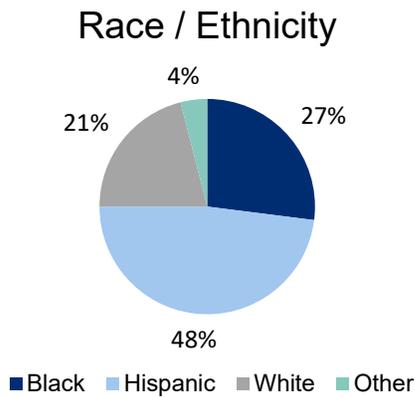
The study took place in a rural town located in north central Texas. Demographics of the analytic sample are presented in Table 1.

Table 1
Demographics of analytic sample

		All students	Treatment	Comparison
n		381	225	156
Race/Ethnicity N (%)	Black/African American	104 (27.3)	67 (29.8)	37 (23.7)
	Hispanic	183 (48.0)	108 (48.0)	75 (48.1)
	Other	15 (3.9)	9 (4.0)	6 (3.8)
	White	79 (20.7)	41 (18.2)	38 (24.4)
Bilingual N (%)		95 (24.9)	55 (24.4)	40 (25.6)
Special Education N (%)		50 (13.1)	33 (14.7)	17 (10.9)
Economic Disadvantage N (%)		276 (72.4)	165 (73.3)	111 (71.2)
Grade N (%)	1	112 (29.4)	68 (30.2)	44 (28.2)
	2	143 (37.5)	85 (37.8)	58 (37.2)
	3	56 (14.7)	33 (14.7)	23 (14.7)
	4	27 (7.1)	13 (5.8)	14 (9.0)
	5	21 (5.5)	13 (5.8)	8 (5.1)
	6	22 (5.8)	13 (5.8)	9 (5.8)

Random assignment of students was successful, as shown in Table 1, in that student characteristiscs, including academic achievement, were similar across groups. The overall attrition rate was 8.85%, while the differential attrition rate was 2.86%. This indicates low attrition in this study.

Measures

In order to address the research questions, the study team gathered and analyzed data including student achievement data and Air Reading usage data (see Table 2).

Table 2
Research questions with data sources and measures.

Research questions	Student achievement data	Air Reading usage data
1. What is the effect of Air Reading on reading achievement for students performing below grade level, in comparison to 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?	✓	✓

Data sources and measures for the current study included student achievement data and Air Reading usage data as described below.

Student achievement. Student achievement outcomes included the Northwest Evaluation Association's Measures of Academic Progress (MAP) reading scores from the 2023-24 school year. The MAP is a computer-adaptive assessment given to students to assess baseline skills and student growth in reading. It is given up to three times a year, during the fall, winter, and spring. These student achievement data were collected by Air Reading from the school district.

Air Reading Usage. Usage data related to student participation in tutoring were collected from Air Reading and shared with the researchers for the 2023-24 school year. These data included one continuous usage measure: number of sessions completed. The usage metric was also operationalized as a categorical variable with two levels: *high dosage*, defined as students who completed 40 or more tutoring sessions, and *low dosage*, encompassing students who completed between 0 to 39 sessions. This classification aimed to delineate meaningful differences in engagement levels among participants[1].

Analytical Approach

To compare student achievement in reading for Air Reading students and control students, we used hierarchical linear modeling, where students were nested within the schools where they were enrolled. For the student achievement analyses, we controlled for students' baseline achievement, as indicated by the pretest measure. For each outcome measure, treatment and control students had acceptably similar pretest scores (What Works Clearinghouse, 2022), outlined in Table 3.

Table 3
Baseline Equivalence on Fall NWEA MAP (Pretest) for Treatment and Control Students

---

[1] In conducting sensitivity analyses, we explored alternative categorizations of the dosage variable, including three-category and four-category models, as well as varying cutoff thresholds. These alternative models yielded results consistent with those reported herein. For the sake of parsimony and clarity, we present findings based on the two-category dosage model and include results for the three-category model in Appendix A.

Analytic Sample	All students	Treatment			Comparison			Effect
	n	n	M	SD	n	M	SD	Size
NWEA MAP	381	156	157.63	17.87	225	156.98	16.80	+0.037

Notes. 1. Effect sizes are experimental minus control means divided by the control group standard deviation. 2. MAP = Northwest Evaluation Association's Measures of Academic Progress.

We used hierarchical linear models to adjust for remaining differences between groups after randomization when comparing achievement between Air Reading and comparison students. We used the following model to estimate program impacts, which adjusted for prior achievement, socioeconomic status, race/ethnicity, special education status, English language learner status, grade level, and blocking variables used in random assignment:

Students are the level-1 units and schools are the level-2 units. The HLM can be generally written as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(Pretest)_{ij} + \gamma_{20}(Treatment)_{ij} + \gamma_{K0}W_{ij} + \gamma_{M0}V_{ij} + u_{0j} + r_{ij}$$

where Y_{ij} represents the student outcome (student achievement on NWEA MAP) for student i in school j, γ_{00} is the covariate-adjusted grand mean for the control group, γ_{10} is the regression coefficient of the pretest, $(Pretest)_{ij}$ is the pretest score, γ_{20} is the average treatment effect, $(Treatment)_{ij}$ is the binary treatment indicator at level 1, γ_{K0} is a vector of regression coefficients for student demographic and grade level covariates, W_{ij} is a vector of student demographic and grade level covariates, γ_{M0} is a vector of regression coefficients for blocking variables used in randomization, V_{ij} is a vector of blocking variables used in randomization, u_{0j} is the random school effect, and r_{ij} is the random student error.

We also examined whether there appeared to be differential effects of Air Reading for students based on demographic characteristics. To do so, we added interaction terms between the treatment indicator and student-level covariates of interest. This allows us to determine whether students with specific characteristics who received Air Reading outperformed similar, comparison peers. For the dosage analysis, we adapted the model above to analyze the association between implementation dosage and student outcomes by replacing the treatment indicator with a categorical indicator of treatment dosage (high/medium/low). All covariates were grand mean centered to facilitate interpretation of the intercept. All analyses were conducted using R statistical software (R Core Team, 2024) and the *lmerTest* package (Kuznetsova et al., 2017).

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 standardizes 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 is 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 NWEA MAP scores. The list of students was shared with Air Reading, who organized the students by their available tutoring slot. The students were then randomly assigned to treatment or control conditions by Air Reading using the *randomizr* package (Coppock, 2023). At the conclusion of tutoring, data on student achievement, student demographics, and program dosage was merged and deidentified by Air Reading. The deidentified dataset was shared with the research team for analysis.

RESULTS

This section of the report begins with findings related to overall impacts on student achievement. Then, we present results pertaining to impacts for subgroups. We conclude with a presentation of how dosage is associated with outcomes.

Impact on Achievement

What is the effect of Air Reading on reading achievement for students performing below grade level, in comparison to students performing below grade level receiving business-as-usual teaching?

Key Findings

> Air Reading students scored significantly higher on the NWEA MAP reading assessment compared with similar comparison students.

Results of the impact on student reading achievement are reported in Table 4. On the NWEA MAP reading assessment, Air Reading students (M = 174.41) outperformed the control students (M = 172.40). This significant impact reflected an effect size of +0.12 (p = .05). This corresponds to an improvement index of 5, meaning that a typical student in the control group who scored at the 50th percentile would increase their score to the 55th percentile if they participate in the tutoring.

Table 4
Impact of Air Reading on students' academic performance

Outcome	N	Adjusted Control Mean	Adjusted Treatment Mean	Impact Estimate (SE)	p Value	Effect Size
MAP	381	172.40	174.41	2.01 (1.03)	0.05	+0.12*

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. 3. MAP = Northwest Evaluation Association's Measures of Academic Progress. 4. SE = Standard Error. 5. * = significant at the .05 level.

Subgroup Analyses

How do the effects of Air Reading differ by race, ethnicity, English learner status, special education status, economic status, and grade level?

Key Findings

> No statistically significant differences in effects by race, ethnicity, gender, or grade level

> Impacts for first grade students were marginally significant and large

In addition to examining impacts for all students assigned to Air Reading, we also explored whether impacts varied for different student subgroups, including by race/ethnicity, English learner status, special education status, socioeconomic disadvantage, and grade. While there were no significant differential impacts identified for any of the subgroups, some subgroup-specific impacts were noteworthy. For example, for students in first grade there was a marginally significant (p < .01) impact with an effect size of +0.20. This corresponds to an improvement index of 8, meaning that a typical student in the control group who scored at the 50th percentile would increase their score to the 58th percentile if they participate in the tutoring. Detailed results, including effect sizes for each subgroup, are presented in Table 5.

Table 5
Impact of Air Reading on students' academic performance by subgroups

Category	Student Subgroup	N	Adjusted Control Mean	Adjusted Treatment Mean
Ethnicity	Black/African American	104	171.08	173.51
	Hispanic	183	171.51	173.46
	White	79	175.38	177.21
Bilingual	Currently Emergent Bilingual	95	172.21	173.21
	Other Non-Emergent Bilingual Student	286	172.46	174.76
SpecialEducation	No	331	172.75	175.03
	Yes	50	170.06	169.99
EconomicDisadvantage	No	105	173.30	177.36
	Yes	276	172.06	173.24
Grade	1	112	163.13	166.58
	2	143	175.94	177.51
	3	56	180.23	180.38
	4	27	171.75	172.35
	5	21	165.88	166.87
	6	22	183.69	188.01

Notes. 1. 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.

Dosage Analyses

To what extent is dosage received associated with better student outcomes?

Key Findings

> Strong implementation with 62% of students receiving acceptable or better dosage of at least 40 sessions over one semester.

> While there was no significant relationship between level of dosage of tutoring session and achievement, the impacts were largest for students who received the full dosage of Air Reading.

Air Reading provided the research team with program dosage information for the study. Across the weeks of the study, treatment students attended an average of 39.8 sessions of tutoring. However, usage varied across students, from a minimum of 0 sessions to a maximum of 63 sessions.

Figure 1
Air Reading students' dosage of tutoring received

Tutoring Sessions Attended

38%

62%

■ Acceptable Dosage (40+)
■ Not Enough Dosage (0-39)
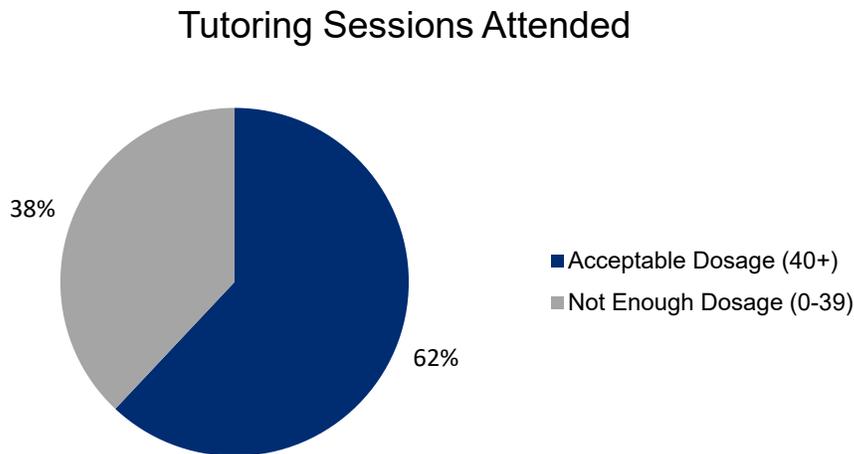
As shown above in Figure 1, most students (62%) received a high dose of sessions, with few students (38%) receiving a low dose of tutoring sessions.

Table 6 presents the results estimating the relationship between level of dosage and NWEA MAP achievement. For this analysis, we categorized students into two dosage groups: high dosage (40 or more sessions) and low dosage (0-39 sessions), with the control group included as a reference. The results revealed that while there were no statistically significant differences in impacts between the high and low dosage groups, the magnitude of the effects differed substantially. Specifically, students who received the high dosage of Air Reading demonstrated a large and statistically significant improvement in reading achievement compared to the control group (ES = +0.17). This corresponds to an improvement index of 7, meaning that a typical student in the control group who scored at the 50th percentile would increase their score to the 57th percentile if they participate in the tutoring. In contrast, the low dosage group showed no significant effect relative to the control group (ES = +0.02).

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 (40+ sessions)	97	225	175.38	172.4	+0.17*
Low (0-39 sessions)	59	225	173.82	172.83	+0.02

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. 3. * = significant at the .05 level.

DISCUSSION

The purpose of this study was to evaluate the effectiveness of Air Reading, a live virtual tutoring initiative aimed at improving reading skills among elementary students, specifically those performing below grade level. The randomized controlled trial design ensured robust and reliable findings by comparing the reading achievement of students receiving the Air Reading intervention to those receiving the standard classroom instruction. This discussion will summarize the key findings, contextualize them within existing literature, discuss the implications for educational practice and policy, and address the study's limitations and suggestions for future research.

Summary of Findings

The results of this evaluation indicate that Air Reading had a positive impact on student reading achievement. Students who participated in Air Reading scored significantly higher on the NWEA MAP reading assessment compared to their peers in the control group, with an effect size of +0.12. This effect size corresponds to an estimated additional 1.5 months of learning and a 5-point improvement in percentile rank, as shown in Table 7. This suggests that the program was effective in enhancing the reading skills of students who were performing below grade level.

Table 7
Interpretation of Air Reading impacts across multiple metrics

Analysis	Effect Size	Additional Months of Learning	Tutoring Efficiency	Improvement Index
Overall	+0.12	1.55	25.8 hours	5
High Dosage (40+ sessions)	+0.17	2.20	18.2 hours	7

Notes. 1. Effect sizes are experimental minus control means divided by the control group standard deviation. 2. Tutoring efficiency provides the number of hours needed to achieve one additional month of

learning. 3. Improvement index describes the change in percentile rank for the median student in the control group to the median student in the treatment group.

Subgroup analyses revealed that the program's effectiveness was consistent across different demographic groups, including race/ethnicity, English learner status, special education status, and economic disadvantage. Additionally, the dosage analysis suggested that while there was no statistically significant relationship between the number of sessions attended and reading outcomes, students who received the full dosage of 40 or more sessions exhibited the largest positive effect size. Notably, students who received the high dosage of tutoring exhibited a larger effect size of +0.17, translating to 2 additional months of learning and a 7-point percentile gain. This finding highlights the need for consistent and adequate exposure to the intervention.

Contextualization within Existing Literature

The findings of this study contribute to the existing literature on the effectiveness of small-group tutoring interventions for improving academic outcomes. Previous research has demonstrated that targeted, intensive tutoring can significantly boost the academic performance of struggling students (Neitzel et al., 2022; Nickow et al., 2024). The success of Air Reading, as evidenced by this study, supports the notion that structured, high-quality tutoring delivered by trained instructors can produce meaningful gains in student achievement.

The findings from the Air Reading program offer important insights into the potential of virtual tutoring as a tool for improving elementary reading skills. While this effect size is somewhat lower than the robust impacts typically reported in studies of traditional in-person tutoring, these findings should be considered in the context of recent studies on similar virtual tutoring models.

Overall Effect Sizes Across Virtual Tutoring Models

The effect size observed in this study of Air Reading (+0.12, noted above) is consistent with the outcomes of other virtual tutoring programs which have generally reported smaller impacts compared to in-person interventions. For example, the OnYourMark program, which also focuses on early literacy, demonstrated overall positive impacts that were statistically significant, but small (ES = +0.08) (Loeb et al., 2023). The BookNook program, another virtual reading intervention, reported an even smaller overall effect size of +0.05, further indicating the challenges of achieving large-scale impacts through virtual tutoring (Ready et al., 2024) (Consortium for Policy Research in Education, 2023). Against this backdrop, the Air Reading results are notable for their relatively stronger impact.

Dosage and Implementation Fidelity

The success of virtual tutoring programs often hinges on the level of dosage and the

fidelity of implementation. In this study of Air Reading, consistent engagement with the tutoring sessions was crucial to achieving the observed impacts. This pattern is echoed in the BookNook study, where only 20% of students received the recommended dosage of 20 or more sessions, which likely contributed to the modest overall effect size (Ready et al., 2024). These findings highlight a common challenge in virtual tutoring: maintaining high levels of student participation and engagement. Programs like OnYourMark have similarly challenges with implementation fidelity, which has impacted the overall effectiveness of their interventions (Loeb et al., 2023).

Impact of High Dosage

Where virtual tutoring programs have shown the most promise is in their impact on students who receive a higher dosage of tutoring. This result showed that students who completed 40 or more sessions achieved an effect size of +0.17, demonstrating the significant gains that can be made with sustained engagement. These results suggest that virtual tutoring can indeed produce meaningful academic improvements, particularly when students engage consistently with the intervention.

In summary, while the overall impacts of Air Reading may not reach the levels typically seen with in-person tutoring, they are in line with—and in some cases exceed—the results of other virtual tutoring models. These findings underscore the importance of dosage and implementation fidelity in maximizing the effectiveness of virtual tutoring programs. As the landscape of educational interventions continues to evolve, Air Reading's relatively strong performance within the virtual context highlights its potential as a valuable tool for supporting student learning in diverse educational settings.

Implications for Educational Practice and Policy

The findings from this study of Air Reading offer important insights into the potential of virtual tutoring as an effective intervention for improving reading outcomes among elementary students. The scalability of virtual tutoring programs represents a significant advantage, particularly in extending educational support to students in diverse geographic and socioeconomic contexts. Air Reading's impact, translating to an estimated 1.5 additional months of learning for the overall cohort and 2.0 additional months for students who received a high dosage, underscores the potential of virtual tutoring to effectively supplement traditional classroom instruction. This scalability is particularly critical in rural and underserved areas, where access to high-quality educational resources is often limited. The findings suggest that virtual tutoring can be a viable strategy for addressing educational inequities, provided that sufficient engagement and dosage are maintained.

The critical role of dosage in determining the effectiveness of virtual tutoring programs is evident in these results. Students who received a high dosage of tutoring (40 or more sessions) demonstrated significantly greater gains. These outcomes emphasize the necessity of sustained engagement in virtual tutoring to maximize its benefits.

Educational practitioners should consider strategies that facilitate consistent student participation, such as integrating tutoring sessions into the regular school schedule or providing targeted incentives to encourage attendance. Effective implementation of virtual tutoring requires not only a well-structured program but also robust support systems for both students and educators. Moreover, ongoing communication between tutors, teachers, and parents is vital to maintaining student engagement and addressing any challenges that arise.

The evidence from this study of Air Reading helps inform the design and implementation of educational policies aimed at improving literacy outcomes. While virtual tutoring may not always replicate the impact of in-person tutoring, it provides a flexible and scalable alternative that can complement traditional instructional strategies. Policymakers should consider investing in virtual tutoring programs, particularly in contexts where in-person tutoring is not feasible, to help bridge educational gaps. However, it is essential that such investments are accompanied by efforts to ensure high levels of dosage and fidelity in implementation, which are critical for maximizing the effectiveness of these programs.

Limitations and Future Research

While the study provides robust evidence of the effectiveness of Air Reading, several limitations should be considered. First, the study was conducted within a specific rural district in Texas, which may limit the generalizability of the findings to other contexts, such as urban or suburban districts, or regions with different demographic profiles. Future research should aim to replicate this study in diverse geographic and educational settings to assess the broader applicability of Air Reading and to understand how local context may influence program effectiveness.

Second, although the study highlights the importance of dosage in achieving meaningful academic outcomes, there was considerable variability in the number of sessions completed by students. Moreover, the tutoring was offered only during a single semester, which may have limited the potential for achieving higher dosage levels. This variability complicates the interpretation of the results, particularly regarding the minimum effective dosage required for significant academic gains. Future studies should investigate the optimal dosage needed to maximize the impact of virtual tutoring, possibly through longer implementations spanning multiple semesters or academic years, allowing for greater cumulative exposure to the intervention. Experimental designs that manipulate dosage levels, or more fine-grained analyses that account for session quality as well as quantity, could further clarify the relationship between dosage and outcomes.

Additionally, the current study focused on short-term reading gains as measured by the NWEA MAP assessment. While these results are promising, they do not capture the long-term impact of Air Reading on sustained reading proficiency or broader academic outcomes. Longitudinal studies are needed to examine whether the benefits observed in

this study persist over time and whether they translate into long-term improvements in academic performance and educational attainment.

Finally, the present study did not fully explore the factors contributing to variability in implementation fidelity across different schools and student groups. Understanding the conditions under which virtual tutoring is most effective—such as the role of teacher support, the integration of tutoring within the school day, and the use of data to personalize instruction—remains a critical area for future research. Additionally, more detailed process evaluations could help identify the key components of successful implementation and offer practical guidance for scaling virtual tutoring programs in diverse settings.

While Air Reading has demonstrated its potential as an effective tool for improving reading outcomes, there is a need for further research to refine and optimize the model. Addressing the limitations identified in this study—including exploring longer implementations with greater dosage—will be essential for maximizing the impact of virtual tutoring and ensuring that these programs can effectively support diverse student populations across various educational contexts.

Conclusion

In conclusion, the evaluation of Air Reading demonstrates that live virtual tutoring can significantly improve reading achievement among elementary students performing below grade level. The program's consistent implementation and its potential for scalability and adaptability make it a promising intervention for addressing reading deficiencies in various educational contexts. Policymakers and educators should consider incorporating similar tutoring models into their strategies to support struggling readers and promote equitable educational outcomes. Future research should continue to explore the long-term impact of such programs and identify best practices for implementation and dosage to further enhance their effectiveness.

REFERENCES

Coppock, A. (2023). *randomizr: Easy-to-Use Tools for Common Forms of Random Assignment and Sampling* (Version R package version 1.0.0) [Computer software]. https://CRAN.R-project.org/package=randomizr

Fesler, L., Gu, A., & Chojnacki, G. (2023). *Air Tutors' online tutoring: Math knowledge impacts and participant math perceptions*. Mathematica.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Loeb, S., Novicoff, S., Pollard, C., Robinson, C., & White, S. (2023). *The effects of virtual tutoring on young readers: Results from a randomized controlled trial*. National Student Support Accelerator, Stanford University. https://studentsupportaccelerator.com/sites/default/files/Effects%20of%20Virtual%20Tutoring%20on%20Young%20Readers.pdf

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

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

R Core Team. (2024). *R: a language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ready, D. D., McCormick, S. G., & Shmoys, R. J. (2024). *The effects of in-school virtual tutoring on student reading development: Evidence from a short-cycle randomized controlled trial* (EdWorkingPaper: 24-942). Annenberg Institute at Brown University. https://edworkingpapers.com/ai24-942

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

APPENDIX A: Impact of Dosage Level on Academic Achievement

As noted above in the body of the report, Air Reading provided the research team with program dosage information for the study. In examining the extent to which dosage received was associated with better student outcomes, several different analyses were conducted setting different dosage thresholds. Table 8 presents the results estimating the relationship between level of dosage and NWEA MAP achievement based on three category levels. In this analysis, students completing between 0-36 sessions were determined to have completed "Not Enough" sessions, those completing between 37-47 sessions were identified as having completed an "Acceptable" number, and those completing 48 or more sessions were determined to have received a full dosage. While the magnitude of the impact is largest for students who received the full dosage, post hoc tests revealed no statistically significant differences between impacts for any levels of dosage.

Table 8
Relationship between Air Reading dosage and achievement

Level of Dosage	Treatment N	Control N	Adjusted Treatment Mean	Adjusted Control Mean	Effect Size
Full (48+ sessions)	22	225	176.09	172.4	+0.21
Acceptable (37-47 sessions)	94	225	174.27	172.4	+0.11
Not Enough (0-36 sessions)	40	225	173.82	172.4	+0.08

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.