Amira Learning in Texas: Impact on 24-25 STAAR

Kelly Boden, PhD

DATE
12/4/2025

Executive Summary

Analysis of 36,123 students across 130 Texas districts demonstrates that using Amira's AI-powered reading tutor predicts significant, measurable improvements in reading achievement. Students with high implementation (30+ min/week) scored 6 percentile points higher on 2025 STAAR Reading assessments (Cohen's d=0.32, p<0.001) and were 82% more likely to meet grade-level standards compared to minimal users. Longitudinal analysis of 24,053 students shows that increased usage drives accelerated learning, with high-usage students gaining 17 additional scale score points year-over-year. Equity-promoting effects were strongest for Black students (+20 scale score points of growth, representing a 25% relative increase) and economically disadvantaged students (+22 points, 30% relative increase), demonstrating Amira's potential to benefit all students.

Students who used Amira more frequently demonstrated significantly higher reading achievement across all grade levels, with particularly strong effects in earlier grades. Grade 3 students with Very High usage scored 9 percentile points higher than Low usage students (58th vs. 49th percentile; d = 0.42, p < 0.001) and had nearly three times the odds of meeting grade-level standards (OR = 2.81), translating to proficiency rates of 63% compared to 49% for Low usage peers. Grade 4 showed similarly robust effects, with Very High usage students demonstrating 7 percentile point gains and 2.45 times the odds of proficiency (65% vs. 53% meeting standards). Longitudinal growth analyses reinforced these cross-sectional findings by demonstrating that higher usage predicted accelerated improvement from students' own baseline performance. There was a clear dose-response pattern with Medium usage students gaining 5 additional scale score points, High usage students gaining 12 points, and Very High usage students gaining 17 points compared to Low usage peers. This provides strong evidence that sustained engagement with Amira drives meaningful learning gains over time.

The evidence is clear, the opportunity is significant, and the path forward is actionable. Districts that invest in systems ensuring consistent, adequate usage, particularly for students facing the greatest barriers to reading success, are likely to see measurable returns in both overall performance and educational equity. Amira's adaptive AI technology offers a scalable solution to provide every student with personalized reading support, helping fulfill the fundamental promise that every child can become a proficient reader.

Table of Contents

Introduction

Amira Learning was created for one purpose—to couple the Science of Reading with AI, giving every child a pathway to the power of reading. Founded by educational researchers and AI scientists, Amira was born in academia and developed through rigorous collaboration with leading reading scientists to ensure that every instructional decision reflects evidence-based best practices. The platform is designed to deliver the explicit, systematic instruction that decades of reading research have shown to be most effective, while leveraging AI to provide the individualized attention and immediate feedback that is rarely feasible in traditional classroom settings. As reading science continues to evolve, Amira rapidly integrates new insights to reflect every key element of evidence-based literacy instruction.

Amira's AI-powered reading tutor uses advanced speech recognition and natural language processing to listen to students read aloud, assess their reading skills in real-time, and provide immediate, targeted support. The platform identifies specific skill gaps across the foundational pillars of reading—phonics, fluency, vocabulary, and comprehension—then delivers personalized micro-interventions during one-on-one tutoring sessions. The AI tutor provides explicit modeling, corrective feedback, and scaffolded practice that adapts dynamically based on each student's demonstrated needs, creating a truly individualized learning experience that mirrors the instructional moves of expert reading teachers. By combining the scalability of technology with the precision of evidence-based instruction, Amira addresses a critical challenge in education: providing every student with consistent access to high-quality, personalized literacy support.

This study evaluates the effectiveness of Amira's platform in improving reading outcomes for elementary students, focusing on grades 3-5 across 130 school districts in Texas during the 2024-2025 school year.

Amira Tutor

Amira Tutor functions as an AI-powered, one-on-one reading coach that delivers adaptive tutoring support across a wide range of skill levels (Amira Learning, 2025). The program uses advanced speech recognition and natural language processing, developed in collaboration with Carnegie Mellon University and Johns Hopkins University, to listen to oral reading, provide real-time feedback, and adjust instruction

to individual needs. Amira Tutor combines artificial intelligence, Science of Reading principles, and neuroscience research to support word recognition, oral reading fluency, and comprehension while students engage with leveled texts drawn from text sets aligned to district curricula and Core Knowledge topics.

During each tutoring session, students select from a small set of stories at an appropriate instructional level and then read aloud while Amira Tutor monitors performance and delivers micro-interventions. Sessions include support such as pronunciation help, morphology and word-structure guidance, high-frequency word practice, and decoding prompts, along with fluency support through paced reading, prosody modeling, repeated reading opportunities, and explicit feedback on expression and intonation. For more advanced texts, tutoring sessions may incorporate AI-led Comprehension Conversations consisting of open-ended, text-based questions designed to promote inference making, summarizing, and evidence-based discussion. For Early Readers and Pre-Readers, the Early Reader Skills Scaffold provides a structured sequence of research-based practice activities focused on foundational literacy skills, including phonological awareness and phonics, in place of or in addition to connected-text reading.

Texas STAAR Assessment

The State of Texas Assessments of Academic Readiness (STAAR) is a standardized academic achievement test designed to measure the extent to which students have learned and can apply the knowledge and skills defined in the Texas Essential Knowledge and Skills (TEKS) curriculum standards (Texas Education Agency [TEA], n.d.-a). Every STAAR question is aligned to state academic standards intended to prepare students for postsecondary readiness. The assessment fulfills the requirements of the federal Every Student Succeeds Act, which mandates testing of all students in specific grades and subjects (TEA, n.d.-a).

For elementary students in grades 3–5, STAAR is an online assessment in reading language arts (RLA) administered annually, with Spanish versions available for students in those grades (TEA, n.d.-a). The reading assessments evaluate students' comprehension, analysis, and application of literacy skills aligned to grade-level expectations. In 2023, STAAR underwent a major redesign to better align with classroom instruction, including the addition of more open-response items, cross-subject reading passages, and enhanced writing components (TEA, 2025).

Performance Standards and Scoring

STAAR results are reported using multiple metrics to provide comprehensive information about student performance. Raw scores (the number of items answered correctly) are converted to scaled scores, which typically range from approximately 1400 to 2400 depending on the grade and subject. Scaled scores allow for comparisons across different administrations and forms of the assessment within the same grade and subject. From the scaled scores, students receive both a performance level classification and a percentile rank (TEA, n.d.-a).

Student performance on STAAR is categorized into four performance levels: Masters Grade Level, Meets Grade Level, Approaches Grade Level, and Did Not Meet Grade Level. Students who achieve Approaches Grade Level or higher have passed the assessment (TEA, n.d.-b). The Texas Education Agency defines these categories as follows:

- **Masters Grade Level:** Performance indicates students are expected to succeed in the next grade or course with little or no academic intervention and demonstrate the ability to think critically and apply knowledge and skills in varied contexts.
- **Meets Grade Level:** Performance indicates students have a high likelihood of success in the next grade or course but may still require short-term, targeted academic intervention.
- **Approaches Grade Level:** Performance indicates basic proficiency and represents the minimum passing threshold; students at this level may need ongoing support to succeed in subsequent grades.
- **Did Not Meet Grade Level:** Performance indicates students did not demonstrate sufficient understanding and are likely to struggle in the next grade without substantial intervention (TEA, n.d.-b).

Percentile ranks provide additional context by indicating the percentage of students in the statewide reference group who scored at or below a given student's performance. For example, a student at the 65th percentile performed as well as or better than 65% of students statewide. Percentile ranks allow educators and parents to understand how an individual student's performance compares to that of peers across Texas, independent of whether the student met specific performance standards (TEA, n.d.-a).

Methodology

Study Design

This quasi-experimental study employed a dosage-response design to examine the relationship between Amira tutoring usage and reading achievement for 36,123 students in grades 3-5 across 130 Texas school districts during the 2024-25 school year. Students were categorized into four usage groups based on average weekly minutes of Amira tutoring over a 25-week period: Low (<10 min/week), Medium (10-19 min/week), High (20-29 min/week), and Very High (30+ min/week). This approach allowed us to investigate whether varying levels of engagement with Amira's AI tutor predicted differential performance on the Texas STAAR Reading assessment.

Sample

The sample consisted of 36,123 students in grades 3-5 who used Amira's AI-powered reading tutor during the 2024-25 school year and completed the Spring 2025 Texas STAAR Reading assessment. Students were distributed across 130 unique school districts in Texas, with relatively balanced enrollment across grade levels: 12,070 third graders (33.4%), 11,945 fourth graders (33.1%), and 12,108 fifth graders (33.5%). The sample reflected Texas's diverse student population, with 47% Hispanic, 26% Black, 15% White, 8% Asian, and 4% students of other races. Approximately 36% of students were economically disadvantaged (eligible for free/reduced lunch), 31% were classified as emergent bilingual, 17% received special education services, and 57% were identified as at-risk for academic difficulties. For the longitudinal growth analyses, a subset of 24,053 students in grades 4-5 had both Spring 2024 and Spring 2025 STAAR scores available, enabling year-over-year comparisons. This large, diverse sample provides robust representation of elementary students across Texas and allows for meaningful subgroup analyses examining differential effects across demographic groups and performance levels.

Usage Categorization

Student engagement with Amira was quantified as average weekly minutes of tutoring over the 25-week implementation period. Total minutes of Amira usage were summed across the school year and divided by 25 weeks to calculate each

student's average weekly usage. Students were then classified into four mutually exclusive usage categories:

- Low usage: <10 minutes per week on average (n=14,549; 40.3% of sample)
- Medium usage: 10-19 minutes per week on average (n=8,359; 23.1% of sample)
- High usage: 20-29 minutes per week on average (n=5,769; 16.0% of sample)
- Very High usage: 30+ minutes per week on average (n=7,446; 20.6% of sample)

These categories were selected to reflect meaningful differences in implementation intensity while maintaining adequate sample sizes for statistical analyses. Low usage served as the reference category in all regression models, representing minimal engagement with the platform.

Analytical Approach

We conducted three complementary analyses to provide evidence of Amira's effectiveness across diverse student populations. All analyses employed multilevel mixed-effects models to account for the hierarchical structure of the data, with students (Level 1) nested within schools (Level 2) and districts (Level 3). Random intercepts were specified at the school level to account for clustering effects and between-school variability in baseline achievement. All models included fixed effects for student-level covariates: race/ethnicity (Black, Hispanic, White, Asian, Other Race/Ethnicity), economic disadvantage status, emergent bilingual classification, special education status, and at-risk designation. Usage category (Low, Medium, High, Very High) was entered as a categorical predictor to examine dosage-response relationships, with Low usage serving as the reference category.

Cross-Sectional Achievement Analysis.

Linear mixed-effects models predicted 2025 STAAR Reading percentile ranks as the continuous outcome variable, with random intercepts at the school level and fixed effects for usage category, grade level, and student demographics. Models were estimated using restricted maximum likelihood (REML). We calculated estimated marginal means (EMMs) for each usage category, controlling for all covariates at their sample means. Cohen's d effect sizes were computed using the pooled standard deviation across usage groups.

Proficiency Analysis.

Generalized linear mixed-effects models with a logit link function predicted the probability of achieving Meets or Masters performance on the 2025 STAAR assessment (binary outcome: 1 = Meets/Masters, 0 = Does Not Meet). Models were estimated using the Laplace approximation for maximum likelihood estimation. Results are reported as odds ratios (OR) with 95% confidence intervals, representing the multiplicative increase in odds of proficiency for each usage category relative to Low usage. Predicted probabilities were calculated from the fitted models and converted to percentage point differences in proficiency rates. Grade-specific models were estimated separately for each grade level.

Longitudinal Growth Analysis.

For the subset of 24,053 students in grades 4-5 with both Spring 2024 and Spring 2025 STAAR scores, we estimated linear mixed-effects models predicting scaled score growth (calculated as 2025 scaled score minus 2024 scaled score). This within-student change score approach controls for individual baseline achievement and strengthens causal inference by examining whether higher usage predicted greater improvement from students' own starting points. Subgroup analyses examined whether growth effects varied across demographic groups by estimating separate models for each subgroup (Black, Hispanic, economically disadvantaged, emergent bilingual, special education, at-risk students). Estimated marginal means were calculated for each usage category within each subgroup, and relative percentage increases were computed by dividing the absolute gain by the Low usage group mean.

All analyses were conducted in R version 4.3.1 using the lme4 package for mixed-effects models and the emmeans package for estimated marginal means. Statistical significance was evaluated at $\alpha = 0.05$, and all p-values for multiple comparisons were adjusted using Tukey's method to control family-wise error rate. Effect sizes (Cohen's d) were interpreted using conventional benchmarks: small (d = 0.20), medium (d = 0.50), and large (d = 0.80).

Results

Cross-Sectional Achievement Analysis.

Multilevel mixed-effects models revealed a significant dose-response relationship between Amira tutoring usage and 2025 STAAR Reading percentile ranks, controlling for school clustering, grade level, and student demographics. Students with Very High usage (30+ min/week) scored 6 percentile points higher than students with Low usage (<10 min/week), representing a small-to-medium effect size (Cohen's d = 0.32, $p < 0.001$). The estimated marginal means showed a consistent upward trend across usage categories: Low usage students scored at the 53rd percentile, Medium usage at the 54th percentile, High usage at the 56th percentile, and Very High usage at the 59th percentile. This pattern held after controlling for race/ethnicity, economic disadvantage, emergent bilingual status, special education status, and at-risk designation, suggesting that higher engagement with the Amira tutor was associated with meaningfully higher achievement independent of student background characteristics.

Figure 1. Estimated Marginal Means: 2025 STAAR Percentile by Usage Category

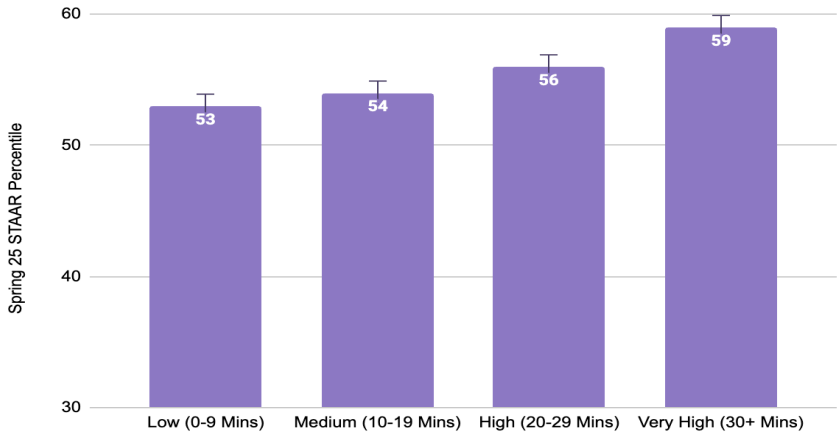


Table 1. Fixed and Random Effects Estimates for STAAR Reading Percentile Ranks

Fixed Effects	Model 1	Model 2

Intercept	46.84*(0.69)	66.58(1.13)
Grade Level		0.04 (0.17)
Usage Category (Low as reference)		
Medium Usage		1.04*(0.40)
High Usage		3.43*(0.46)
Very High Usage		6.40(0.47)
Race/Ethnicity (White as reference)		
Asian		6.42*(0.59)
Black/African American		-7.63*(0.50)
Hispanic		-4.16*(0.45)
Other		-2.07*(0.77)
Student Characteristics		
Economically Disadvantaged		-7.59*(0.34)
Emergent Bilingual		8.06*(0.45)
Special Education		-21.49*(0.39)
At-Risk		-21.59(0.38)
Random Effects		
School-level Intercept	117.8	130.4
Student-level residual	752.5	499.5

Model Fit		
REML Criterion	342,038.1	232,212.0

Note. Standard errors in parentheses. * p < 0.05. Model 1 includes only random school-level intercepts. Model 2 includes all predictors with random school-level intercepts.

Grade-specific analyses revealed that the relationship between usage and achievement was consistent across all three grade levels, though effect sizes varied by grade. Grade 3 students showed the strongest effects, with Very High usage students scoring 9 percentile points higher than Low usage students (49th vs. 58th percentile; Cohen's d = 0.42, p < 0.001). Grade 4 students demonstrated similar patterns, with a 7 percentile point difference between Very High and Low usage groups (52nd vs. 59th percentile; Cohen's d = 0.37, p < 0.001). Grade 5 students showed smaller but still significant effects, with a 6 percentile point difference (58th vs. 64th percentile; Cohen's d = 0.26, p < 0.001). Notably, Grade 5 students had higher baseline performance across all usage categories, suggesting that even students entering the year with stronger reading skills benefited from increased Amira usage.

Proficiency Analysis.

Generalized linear mixed-effects models with logit link functions examined how Amira usage predicted the probability of students achieving Meets or Masters performance levels on the 2025 STAAR assessment. Across all grades, students with Very High usage had significantly higher odds of meeting grade-level standards compared to Low usage students. The overall effect showed that Very High usage students had 1.82 times the odds of meeting standards (OR = 1.82, 95% CI [1.67, 1.98], p < 0.001), translating to a 14.3 percentage point increase in the probability of proficiency. Proficiency rates increased systematically across usage categories: 32.9% of Low usage students met grade-level standards, compared to 37.2% of Medium usage students, 42.1% of High usage students, and 47.2% of Very High usage students.

Grade-specific logistic regressions revealed the strongest effects in Grade 3, where Very High usage students had nearly three times the odds of meeting grade-level standards compared to Low usage students (OR = 2.81, 95% CI [2.45, 3.23], p < 0.001). This translated to a 14 percentage point increase in the rate of proficiency, from 49% of Low usage students meeting standards to 63% of Very High usage students. Grade

4 showed similarly robust effects (OR = 2.45, 95% CI [2.14, 2.80], p < 0.001), with proficiency rates increasing from 53% to 65% (+12 percentage points). Grade 5 demonstrated moderate effects (OR = 1.76, 95% CI [1.55, 2.00], p < 0.001), with proficiency rates increasing from 57% to 64% (+7 percentage points). The decreasing effect sizes across grades may reflect ceiling effects, as Grade 5 students had higher baseline proficiency rates and therefore less room for improvement.

The consistent relationship between usage and proficiency across all three grade levels provides strong evidence that increased engagement with Amira's AI tutor is associated with meaningful improvements in the likelihood of meeting state reading standards. These findings are particularly notable because the proficiency threshold represents a high-stakes benchmark that determines grade-level advancement and school accountability ratings in Texas.

Table 2. Logistic Regression Results: Probability of Meeting Grade-Level Standards on 2025 STAAR Reading

Grade	Usage Category	Odds Ratio	95% CI	p-value
Grade 3	Low (reference)	1.00	–	–
	Medium	1.38	[1.24, 1.54]	<0.001
	High	1.84	[1.63, 2.07]	<0.001
	Very High	2.81	[2.45, 3.23]	<0.001
Grade 4	Low (reference)	1.00	–	–
	Medium	1.16	[1.03, 1.31]	0.016
	High	1.46	[1.29, 1.66]	<0.001
	Very High	2.45	[2.14, 2.80]	<0.001
Grade 5	Low (reference)	1.00	–	–
	Medium	1.06	[0.95, 1.19]	0.325

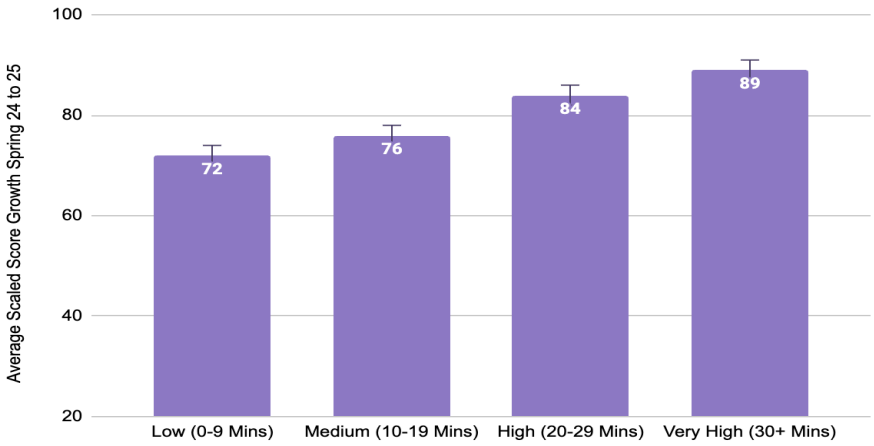
	High	1.46	[1.29, 1.65]	<0.001
	Very High	1.76	[1.55, 2.00]	<0.001

Note. Results from generalized linear mixed-effects models with logit link function, controlling for school clustering and student demographics (race/ethnicity, economic disadvantage, emergent bilingual status, special education, at-risk designation). Proficiency defined as achieving Meets or Masters performance level. CI = confidence interval. Odds ratios represent multiplicative increase in odds of proficiency relative to Low usage group within each grade.

Longitudinal Growth Analysis.

Multilevel mixed-effects models examining scaled score growth from Spring 2024 to Spring 2025 for students in grades 4 and 5 (n=24,053) revealed significant positive associations between Amira tutoring usage and year-over-year improvement. Students with Very High usage (30+ min/week) showed 17 scale score points more growth than students with Low usage (<10 min/week), representing a small effect size (Cohen's d = 0.20, p < 0.001). The estimated marginal means demonstrated a clear dose-response pattern: Low usage students gained 72 scale score points on average, Medium usage students gained 76 points, High usage students gained 84 points, and Very High usage students gained 89 points.

Figure 2. Estimated Marginal Means: Scaled Score Growth by Usage Category



This longitudinal analysis strengthens causal inference by demonstrating that higher engagement with the AI tutor predicted greater improvement from students' own baseline performance, controlling for demographic characteristics. The finding that students with more tutoring usage showed accelerated growth—rather than simply higher static achievement—provides stronger evidence that Amira is driving learning gains. By examining within-student change over time, this approach controls for unmeasured time-invariant student characteristics and reduces concerns about selection bias that could confound cross-sectional comparisons.

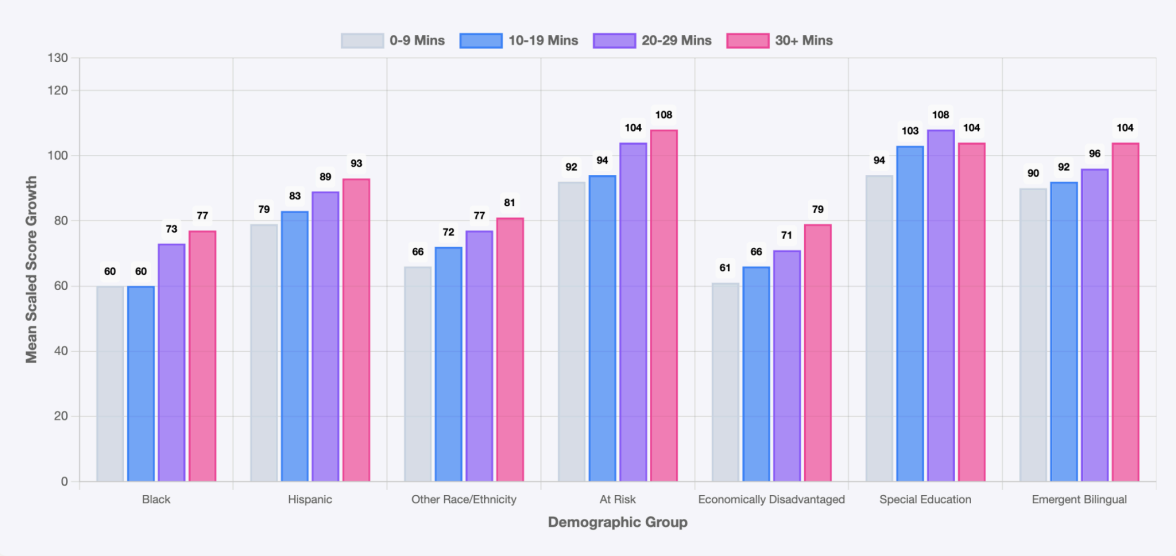
Table 3. Multilevel Model Results: Year-over-Year Growth in STAAR Reading Scaled Scores (Grades 4-5)

Fixed Effects	Model 1	Model 2
Intercept	92.08*(1.94)	252.20(8.36)
Grade Level		-40.06* (1.72)
Usage Category (Low as reference)		
Medium Usage		4.56*(2.44)
High Usage		12.12*(2.78)
Very High Usage		17.36*(2.77)
Race/Ethnicity (White as reference)		
Asian		1.77(3.51)
Black/African American		0.19(2.95)
Hispanic		2.84(2.66)
Other		-2.07(4.70)
Student Characteristics		

Economically Disadvantaged		-5.16*(2.03)
Emergent Bilingual		-6.32*(2.63)
Special Education		-12.41*(2.38)
At-Risk		-19.55(2.25)
Random Effects		
School-level Intercept	588.5	476.0
Student-level residual	11.592.1	11,059.0
Model Fit		
REML Criterion	284,674.4	199,889.6

Subgroup analyses of scaled score growth revealed that all demographic groups benefited from increased Amira tutoring usage, with particularly pronounced effects for students from historically marginalized populations. Black students with Very High usage demonstrated 100 scale score points of growth compared to 80 points for Low usage students, representing a 20-point advantage (25% relative increase). Economically disadvantaged students showed the strongest relative gains, improving from 72 scale score points of growth with Low usage to 94 points with Very High usage (+22 points, 30% relative increase). At-risk students also demonstrated substantial benefits, with growth increasing from 88 points to 101 points (+13 points, 14.5% relative increase). Hispanic students showed a 14-point gain (from 79 to 93 points, 18% relative increase), while students receiving special education services gained 11 additional points (from 94 to 105 points, 12% relative increase). Emergent bilingual students demonstrated a 12-point advantage with Very High usage (from 92 to 104 points, 13% relative increase).

Figure 3. Estimated Marginal Means: Scaled Score Growth by Usage Category and Subgroup



Discussion

This large-scale quasi-experimental study provides compelling evidence that Amira's AI-powered reading tutor relates to meaningful improvements in elementary students' reading achievement. Across three complementary analytic approaches we found consistent dose-response relationships between usage and reading outcomes. Students with Very High usage (30+ minutes per week) demonstrated 6-7 percentile point gains in achievement, nearly doubled their odds of meeting grade-level standards, and showed 14-17 additional scale score points of year-over-year growth compared to students with minimal engagement.

The consistency of effects across multiple outcome measures and analytic methods strengthens confidence in the findings. Cross-sectional analyses revealed not only higher overall achievement but also substantially increased probability of reaching proficiency benchmarks, a particularly meaningful result given the high-stakes nature of these standards in Texas accountability systems. The longitudinal growth analyses provide the strongest evidence of causal impact by demonstrating that increased usage predicted accelerated improvement from students' own baseline performance, effectively controlling for time-invariant student characteristics that might confound cross-sectional comparisons.

Grade-Level Patterns and Implications

Effect sizes varied by grade level, with the strongest impacts observed in Grade 3 (d = 0.42) and progressively smaller effects in Grades 4 (d = 0.37) and 5 (d = 0.26). This pattern likely reflects multiple factors. First, Grade 3 represents a critical transition point from learning to read to reading to learn, when targeted phonics and fluency interventions may be particularly impactful. Second, Grade 5 students demonstrated higher baseline proficiency across all usage categories, suggesting potential ceiling effects that limit the magnitude of observable gains. Third, the decreasing effect sizes may indicate that earlier intervention produces larger returns, consistent with research emphasizing the importance of early literacy development.

Despite smaller effect sizes in upper grades, the impacts remained statistically significant and educationally meaningful. Very High usage Grade 5 students still demonstrated 6 percentile point gains and 76% higher odds of proficiency compared to Low usage peers. These findings suggest that Amira provides value across the elementary grade span.

Equity and Differential Effects

A particularly encouraging finding emerged from subgroup analyses, which revealed that Amira's benefits were consistent across all demographic groups and particularly pronounced for students from historically marginalized populations. Black students with Very High usage demonstrated 20 additional scale score points of growth (25% relative increase), while economically disadvantaged students showed 22 additional points (30% relative increase) compared to their Low usage peers within the same demographic categories.

Notably, emergent bilingual students demonstrated complex patterns across analyses. In cross-sectional models, EB students scored higher than non-EB peers after controlling for race/ethnicity and other demographics—a surprising finding that warrants careful interpretation. This may reflect selection effects (schools with strong bilingual programs), additional support provided to EB students, or the cognitive advantages associated with bilingualism. In longitudinal analyses, however, EB students showed slightly less growth than non-EB students, suggesting that while EB students may start from a relative position of strength (perhaps due to intensive support), they face persistent challenges in accelerating reading development. These nuanced patterns highlight the importance of continued research of AI tutoring approaches to optimally serve multilingual learners.

Limitations and Future Directions

Several limitations should be considered when interpreting these findings. First, this quasi-experimental design cannot definitively establish causality. While the longitudinal growth analyses and extensive covariate adjustment strengthen causal inference, unmeasured factors such as teacher quality, instructional time, or school-level implementation support could partially explain the observed relationships. Students with higher usage may differ from low-usage peers in other impactful ways such as motivation, teacher encouragement, or access to technology.

Second, usage was determined by actual implementation rather than random assignment, which does not eliminate selection bias. Schools and teachers choosing to implement Amira with high fidelity may differ systematically from those with minimal implementation. However, the fact that effects remained significant after controlling for school clustering and comprehensive student demographics provides some reassurance that selection bias does not fully account for the findings.

Third, this study examined only one academic year of implementation. Longer-term follow-up studies are needed to determine whether initial gains are sustained, fade out, or compound over time. Multi-year growth trajectories would provide stronger evidence of lasting impact and help identify optimal implementation strategies across grade levels.

Practical Implications

These findings have several important implications for schools and districts implementing AI-powered reading interventions. The clear dose-response relationship underscores the importance of achieving high-fidelity implementation. Schools should target at least 30 minutes per week of Amira usage to maximize benefits, with particular emphasis on consistent implementation in Grade 3 where effects are strongest.

The substantial improvements in proficiency rates have important implications for school accountability. Districts facing pressure to increase the proportion of students meeting grade-level standards may find that high-dosage Amira implementation provides meaningful support toward those goals, particularly in Grade 3 where proficiency rates increased by 14 percentage points.

Conclusion

This study provides evidence that using Amira's AI-powered reading tutor relates to meaningful, measurable improvements in elementary students' reading achievement. The convergence of findings across cross-sectional achievement, proficiency attainment, and longitudinal growth analyses demonstrates that increased engagement with the platform is associated with better reading outcomes. Critically, these benefits extend across demographic groups and appear particularly pronounced for students from historically marginalized populations, suggesting that AI-powered tutoring has genuine potential to advance educational equity.

As AI-powered educational technologies continue to evolve, studies like this one provide evidence about their real-world effectiveness and help identify both the promise and limitations of these tools for supporting student learning. The combination of scalable technology with the precision of evidence-based reading instruction represent a meaningful step toward ensuring that all students receive the individualized support they need to become proficient readers.

References

Amira Learning (2025). *Amira Implementation Guide*.

Texas 2036. (2025, June 17). STAAR grades 3–8 results released: 5 quick takeaways. https://texas2036.org/posts/staar-grades-3-8-results-released-5-quick-takeaways/

Texas Education Agency. (n.d.-a). STAAR. https://tea.texas.gov/student-assessment/staar

Texas Education Agency. (n.d.-b). STAAR performance standards. https://tea.texas.gov/student-assessment/testing/staar/staar-performance-standards

Texas Education Agency. (2025, June 17). Spring 2025 STAAR® results for grades 3–8. https://tea.texas.gov/about-tea/news-and-multimedia/news-releases/news-2025/texas-education-agency-releases-spring-2025-staarr-results-for-grades-3-8