AMIRA
LEARNING

Amira Learning in Louisiana Public Schools

Amira Learning | Every Child Deserves the Chance to Become a Reader
5214F Diamond Heights Blvd, #3255, San Francisco, CA 94131
wwww.AmiraLearning.com | 866-883-7323 | info@amiralearning.com

THE EDTECH COLLECTIVE
Instructure Partner Ecosystem

AMIRA LEARNING® IN LOUISIANA PUBLIC SCHOOLS

ESSA Level II Study (2023–24)

Prepared for:
Amira Learning®

Prepared by Instructure:
Meetal Shah, Ph.D., Lead Researcher
Andrew Scanlan, M.Ed., Senior Researcher
Avery Wall, Data Scientist

June 2025

EXECUTIVE SUMMARY

Amira Learning® contracted with Instructure, a third-party edtech research company, to examine the impact of *Amira Learning's* AI-powered reading platform (henceforth *Amira*) on elementary school students' literacy outcomes. Using the Every Student Succeeds Act (ESSA) standards as guidance in developing a study design, findings in this report align with Level II requirements (Moderate Evidence).

Study Sample and Methodology

This study used a quasi-experimental design to align with ESSA Level II evidence standards. It included a matched analysis sample of 79,084 elementary school (Kindergarten–Grade 5) students (39,542 treatment, 39,542 comparison) from across 12 school districts in Louisiana. The sample was predominantly African American and White (38%, respectively), followed by Hispanic (18%), multi-racial (4%), and Asian (2%). In terms of socioeconomic status (SES), this sample was classified as 75% economically disadvantaged. Ten percent of the sample has English language learner (ELL) designation, 14% of the sample has special education designation, and 50% of the sample identified as female.

Researchers analyzed *Amira's* implementation data—including total ***session time*** (minutes) and the number of passages read—along with demographic data from the 2023–24 school year and standardized assessment results to assess *Amira's* impact on student outcomes. The analysis included Dynamic Indicators of Basic Early Literacy Skills (DIBELS®) composite scores from fall 2023 and spring 2024 for the K–3 sample, as well as Louisiana Educational Assessment Program (LEAP) ELA scores from spring 2023 and 2024 for the Grades 4–5 sample.

For impact analysis, researchers created within-grade matched samples and conducted baseline equivalence testing. All analyses met What Works Clearinghouse (WWC) Version 5.0 baseline equivalence standards (What Works Clearinghouse, 2022). Analyses also included descriptive statistics and multi-level models to examine the association between *Amira* usage and students' spring 2024 DIBELS and LEAP performance (while controlling for fall 2023 and spring 2023 performance, respectively). Researchers also included student-level covariates to control for potential selection bias.

Main Research Findings

Main Research Findings	
+	There was a statistically significant, **positive** association between the total number of **minutes** spent on *Amira* and **DIBELS** scores for students in grades K–3.
+	There was a statistically significant, **positive** association between the total number of **minutes** spent on *Amira* and **LEAP** ELA scores for students in grade 4.
+	There was a statistically significant, **positive** association between the total number of **passages** read in *Amira* and **LEAP** ELA scores for students in grades 4 and 5.
+	Grades K–3 students who used *Amira* had higher spring 2024 DIBELS scores than non-users. This result was statistically significant across all grade-level samples.
+	Grades 4–5 students who used *Amira* had higher spring 2024 LEAP ELA scores than non-users. This result was statistically significant across both grade-level samples.

Conclusions

Given the positive findings, this study provides results to satisfy ESSA evidence requirements for Level II (Moderate Evidence).

TABLE OF CONTENTS

INTRODUCTION 5

STUDY DESIGN AND METHODS 6

IMPLEMENTATION 8

DIBELS ® OUTCOME FINDINGS FOR K-3 STUDENTS 10

LEAP OUTCOME FINDINGS FOR GRADES 4 AND 5 STUDENTS 16

CONCLUSIONS 20

REFERENCES 21

APPENDIX A. KINDERGARTEN: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS 22

APPENDIX B. GRADE 1: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS 25

APPENDIX C. GRADE 2: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS 28

APPENDIX D. GRADE 3: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS 31

APPENDIX E. GRADE 4: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS 34

APPENDIX F. GRADE 5: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS 38

INTRODUCTION

Amira Learning recognizes that teachers and families of early elementary school students often do not have resources to address their individual reading needs. *Amira's* AI-powered reading platform with automated screeners, practice, and embedded assessments, provides teachers and parents with the appropriate tools to identify specific learning needs (including learning difficulties) in a timely manner, engage students in productive struggle through targeted practice, and generate appropriate reading interventions after assessing students. As part of their ongoing efforts to demonstrate the effectiveness of their solution, *Amira Learning* contracted with Instructure, a third-party edtech research company, to examine the impact of *Amira* on elementary school students' literacy outcomes. Using the Every Student Succeeds Act (ESSA) standards as guidance in developing a study design, findings in this report align with Level II requirements (Moderate Evidence). The following research questions guided this study:

Implementation

1) What was the nature of implementation of *Amira* in the 2023-24 school year among Grades K–5 students?
    a) Overall, how many students accessed *Amira*?
    b) On average, how:
        i) much time (in minutes) did students spend on *Amira*,
        ii) many passages did students read in *Amira* in total?

Student Outcomes

2) What was the association between Amira use and students' DIBELS scores (Grade K–3) or Louisiana Educational Assessment Program scores (LEAP; Grades 4–5)? Did students who:
    a) spent more time on *Amira* have better literacy outcomes?
    b) read more passages in *Amira* have better literacy outcomes?
3) Did students who practiced reading in *Amira* have better literacy outcomes than a matched sample of students who did not have access to *Amira*? What was the magnitude of this difference?

This report details the study design and methods, implementation, findings, and conclusions.

STUDY DESIGN AND METHODS

This section of the report briefly describes the study participants, measures, and analysis methods.

Study Design

This study used a quasi-experimental design to align with ESSA Level II evidence standards. It included students who participated in *Amira* during the 2023–24 school year and a matched sample of students who did not use *Amira*.

Setting and Participants

This study included a matched analysis sample of 79,084 elementary school (Kindergarten–Grade 5) students (39,542 treatment, 39,542 comparison) from across 12 school districts in Louisiana.

Based on student demographic data provided by the district, the sample was predominantly African American and White (38%, respectively), followed by Hispanic (18%), multi-racial (4%), and Asian (2%). In terms of socioeconomic status (SES), this sample was classified as 75% economically disadvantaged. Ten percent of the sample has English language learner (ELL) designation, 14% of the sample has special education designation, and 50% of the sample identified as female. The sample was evenly distributed across grades: Kindergarten (15%), Grade 1 (19%), Grade 2 (19%), Grade 3 (18%), Grade 4 (15%), and Grade 5 (14%).

Measures

Researchers analyzed *Amira's* implementation data— including total ***session time*** (minutes) and the number of passages read—along with demographic data from the 2023–24 school year and standardized assessment results to assess *Amira's* impact on student literacy outcomes. The analysis included Dynamic Indicators of Basic Early Literacy Skills (DIBELS®) composite scores from fall 2023 and spring 2024 for the K–3 sample, as well as Louisiana Educational Assessment Program (LEAP) ELA scores from spring 2023 and 2024 for the Grades 4–5 sample. Since both assessment scores are not vertically scaled, researchers conducted the analysis separately by grade-level.

Background on usage metrics. In grades K–3, the number of passages read is a flawed usage metric due to the high variability in activity lengths and the strong negative correlation between student ability and session duration. Younger students, especially in kindergarten and 1st grade, often engage in shorter foundational reading activities, while those who can read connected text encounter passages ranging from 20 to over 200 words based on their level. As a result, lower-performing students tend to have a higher count simply because their activities are shorter. A such the time spent on the platform (***session time***) is a more appropriate metric for measuring usage in these early grades. In grades 4 and 5, previous internal studies have shown a correlation between student ability and time spent per passage is no longer significant. Students who are stronger readers generally read slightly longer texts, but they also read faster and more fluently. For this reason, we examined both time on platform and total passages read as usage metrics in grades 4 and 5.

Data Analysis

Amira and the Louisiana Department of Education uploaded de-identified data from the 2023—24 school year through a secure file transfer protocol. Researchers characterized usage (i.e., the total number of minutes and passages read) using descriptive statistics and establishing usage groups in terms of tertiles (total minutes) and quartiles (total passages read). Researchers used multilevel modeling (MLM) to examine how *Amira* impacts student literacy outcomes. The analyses included district-level random effects and student-level covariates to control for potential selection bias (i.e., baseline achievement, sex, race/ethnicity, and special education designation). In addition, researchers calculated standardized effect sizes to determine the magnitude of changes in treatment students' literacy outcomes.

Baseline Equivalence

To ensure the validity of the study's findings and adhere to ESSA Level II standards, researchers assessed the equivalence of student demographic characteristics and assessment scores between treatment and comparisons groups. The appendices include additional baseline equivalence details for each grade-level sample.

IMPLEMENTATION

This section examines how students used *Amira* during the 2023–24 school year. Researchers analyzed the total amount of time students spent in the platform (frequently referred to as *session time*) and the total number of passages read to understand the extent of student engagement.

What was the nature of implementation of *Amira* in the 2023-24 school year among Grades K–5 students?

- **a) Overall, how many students accessed *Amira*?**
- **b) On average, how:**
  - **i) much time (in minutes) did students spend on *Amira*,**
  - **ii) many passages did students read in Amira in total?**

The total amount of time (in minutes) that students spent, and the total number of passages read in *Amira* varied across grades. Tables 1 and 2 include the variation in usage by grade level and usage metric.

Table 1. *Amira* average total session time (minutes) spent by grade level

Grade	*n*	Average (# of Minutes)	SD	Min.	Max.
Kindergarten	5,765	252	252	1	2,755
Grade 1	7,382	336	303	2	3,125
Grade 2	7,454	290	287	2	3,012
Grade 3	7,156	226	221	2	2,237
Grade 4	6,088	229	219	2	1,256
Grade 5	5,697	190	203	3	1,664

Table 2. *Amira* average total passages read by grade level

Grade	*n*	Average (# of Passages)	SD	Min.	Max.
Grade 4	6,088	39	39	1	485
Grade 5	5,697	34	41	1	444

Table 3. Number of students at that met or exceeded *Amira's* recommended dosage

Grade	Number (%) of students meeting or exceeding session time of 20 mins/week	Number (%) of students meeting or exceeding 5 passages/week
Kindergarten	592 (10%)	963 (17%)

Grade	Number (%) of students meeting or exceeding session time of 20 mins/week	Number (%) of students meeting or exceeding 5 passages/week
Grade 1	1,386 (19%)	1,226 (17%)
Grade 2	1,111 (15%)	548 (7%)
Grade 3	575 (8%)	242 (3%)
Grade 4	467 (8%)	179 (3%)
Grade 5	290 (5%)	166 (3%)

DIBELS ® OUTCOME FINDINGS FOR K–3 STUDENTS

Researchers created a matched sample of *Amira* users and non-users based on students' fall 2023 scores, district, gender, race, socioeconomic status, ELL designation, and special education designation. For *Amira* users, researchers categorized usage groups by dividing total minutes spent on the platform (***session time***) into tertiles. As such, the specifications for usage groups differed by grade-level sample. To address outcome questions, researchers employed a two-level multilevel modeling analysis, with students nested within districts. The models examined the impact of using *Amira* on students' spring 2024 DIBELS scores, controlling for fall 2023 DIBELS scores and statistically significant demographic variables (e.g., gender, race, socioeconomic status, ELL designation, and special education designation). Researchers conducted these analyses in three parts: 1) correlative analyses focusing solely on *Amira* users, 2) comparative analyses comparing matched samples of *Amira* users and non-users, and 3) comparative analyses examining students in the highest Amira usage group versus non-users, provided baseline equivalence was established.

To allow for better interpretability of results, marginal means charts are presented below. The vertical lines at the top of each bar represent a 95% confidence interval. Additional information on these analyses and findings can be found in Appendices A–D.

K

What was the association between Amira use and kindergarten students' DIBELS scores?

Total time spent in Amira (minutes). Results showed multiple statistically significant, positive associations between the total number of minutes spent in *Amira* and DIBELS scores. Kindergarten students who spent:

- 97–271 minutes in *Amira* (moderate use) had significantly higher DIBELS scores than students who spent 96 or fewer minutes (low use; Hedges' g = 0.09, p = .001).
- more than 271 minutes in *Amira* (high use) had significantly higher DIBELS scores than students who spent 96 or fewer minutes (low use; Hedges' g = 0.19, p < .001).

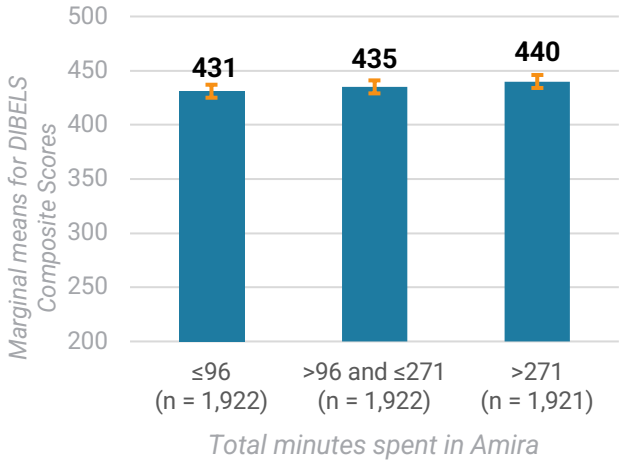


Figure 1. Multi-level models examining the association between total number of minutes and DIBELS scores (Kindergarten).

Did Kindergarten students who practiced reading in *Amira* have better literacy outcomes than a matched sample of students who did not have access to *Amira*?

The matched kindergarten sample demonstrated baseline equivalence (Hedges' g = -0.01; p = .771). The sample of high use students (272–2,755 minutes) and nonusers also met baseline equivalence standards (Hedges' g = 0.05; p = .098). Consequently, researchers then analyzed (a) matched Kindergarten Amira users vs. non-users and (b) high-use Amira users vs. non-users.

Overall, *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant (g = 0.11; p < .001); Figure 2). A Hedges' g value of 0.11 means that if an average Kindergarten student (one who scores right in the middle, at the 50th percentile) had used *Amira*, they would be expected to perform at the **54th percentile.**

High-use *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant (g = 0.21; p < .001); Figure 2). A Hedges' g value of 0.21 means that if an average Kindergarten student (one who scores right in the middle, at the 50th percentile) had used *Amira* at this level, they would be expected to perform at the **58th percentile.**

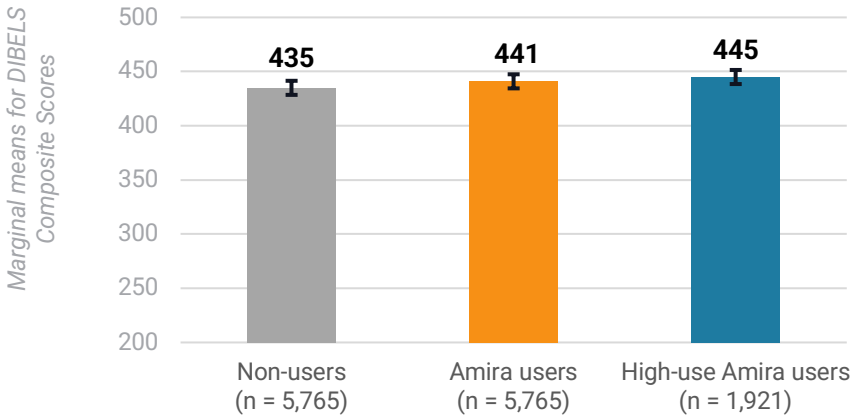


Figure 2. Adjusted mean spring 2024 DIBELS scores for Kindergarten non-users, all Amira users, and high-use Amira users.

1

What was the association between Amira use and Grade 1 students' DIBELS scores?

Total time spent in Amira (minutes). Results showed multiple statistically significant, positive associations between the total number of minutes spent in *Amira* and DIBELS scores. Grade 1 students who spent:

- 147–392 minutes in *Amira* (moderate use) had significantly higher DIBELS scores than students who spent 146 or fewer minutes (low use; Hedges' g = 0.08, p < .001).
- more than 392 minutes in *Amira* (high use) had significantly higher DIBELS scores than students who spent 146 or fewer minutes (low use; Hedges' g = 0.19, p < .001).

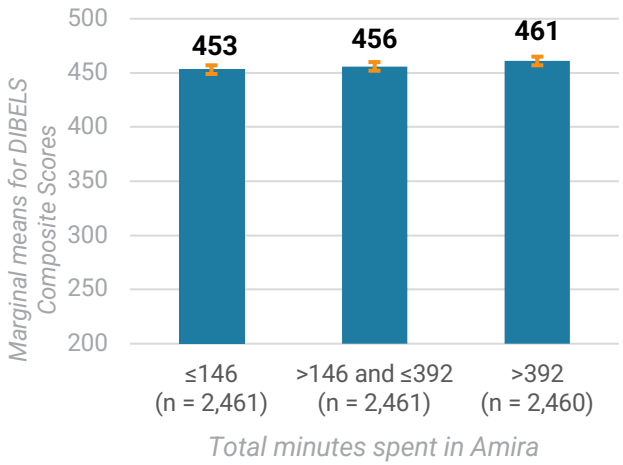


Figure 3. Multi-level models examining the association between total number of minutes and DIBELS scores (Grade 1).

Did Grade 1 students who practiced reading in *Amira* have better literacy outcomes than a matched sample of students who did not have access to *Amira*?

The matched Grade 1 sample demonstrated baseline equivalence (Hedges' $g = 0.14$; $p < .001$). The sample of high use students (393–3,125 minutes) and nonusers also met baseline equivalence standards (Hedges' $g = 0.07$; $p = .037$). Consequently, researchers then analyzed (a) matched Kindergarten Amira users vs. non-users and (b) high-use Amira users vs. non-users.

Overall, *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant ($g = 0.10$; $p < .001$); Figure 4). A Hedges' g value of 0.10 means that if an average Grade 1 student (one who scores right in the middle, at the 50th percentile) had used *Amira*, they would be expected to perform at the **54th percentile.**

High-use *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant ($g = 0.22$; $p < .001$); Figure 4). A Hedges' g value of 0.22 means that if an average Grade 1 student (one who scores right in the middle, at the 50th percentile) had used *Amira* at this level, they would be expected to perform at the **59th percentile.**

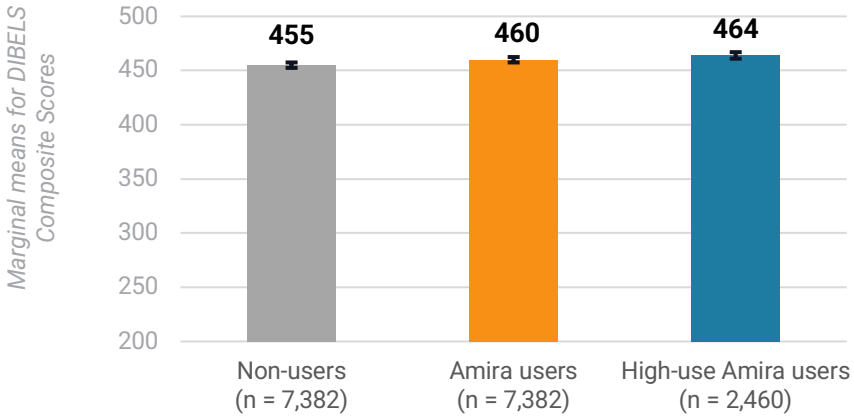


Figure 4. Adjusted mean spring 2024 DIBELS scores for Grade 1 non-users, all Amira users, and high-use Amira users.

2

What was the association between Amira use and Grade 2 students' DIBELS scores? *Total time spent in Amira (minutes).* Results showed one statistically significant, positive association between the total number of minutes spent in *Amira* and DIBELS scores. Grade 2 students who spent:

- more than 318 minutes in *Amira* (high use) had significantly higher DIBELS scores than students who spent 113 or fewer minutes (low use; Hedges' g = 0.06, p = .001).

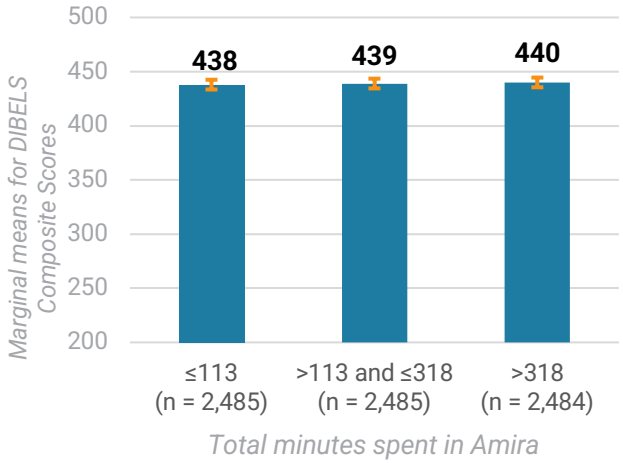


Figure 5. Multi-level models examining the association between total number of minutes and DIBELS scores (Grade 2).

Did Grade 2 students who practiced reading in *Amira* have better literacy outcomes than a matched sample of students who did not have access to *Amira*?

The matched Grade 2 sample demonstrated baseline equivalence (Hedges' g = 0.02; p = .303). The sample of high use students (319–3,012 minutes) and nonusers also met baseline equivalence standards (Hedges' g = 0.09; p < .001). Consequently, researchers then analyzed (a) matched Kindergarten Amira users vs. non-users and (b) high-use Amira users vs. non-users.

Overall, *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant (g = 0.08; p < .001); Figure 6). A Hedges' g value of 0.08 means that if an average Grade 2 student (one who scores right in the middle, at the 50th percentile) had used *Amira*, they would be expected to perform at the **53rd percentile.**

High-use *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant (g = 0.12; p < .001); Figure 6). A Hedges' g value of 0.12 means that if an average Grade 2 student (one who scores right in the middle, at the 50th percentile) had used *Amira* at this level, they would be expected to perform at the **55th percentile.**

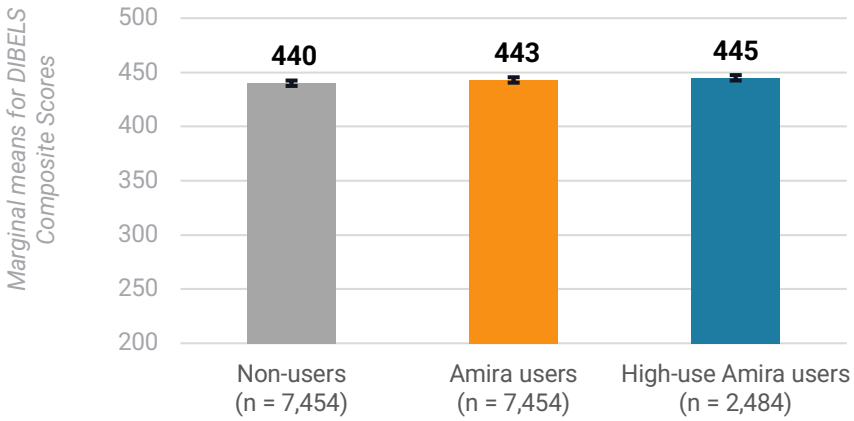


Figure 6. Adjusted mean spring 2024 DIBELS scores for Grade 2 non-users, all Amira users, and high-use Amira users.

3

What was the association between Amira use and Grade 3 students' DIBELS scores?

Total time spent in Amira (minutes). Results showed one statistically significant, positive association between the total number of minutes spent in *Amira* and DIBELS scores. Grade 3 students who spent:

- more than 250 minutes in *Amira* (high use) had significantly higher DIBELS scores than students who spent 92 or fewer minutes (low use; Hedges' g = 0.04, p = .013).

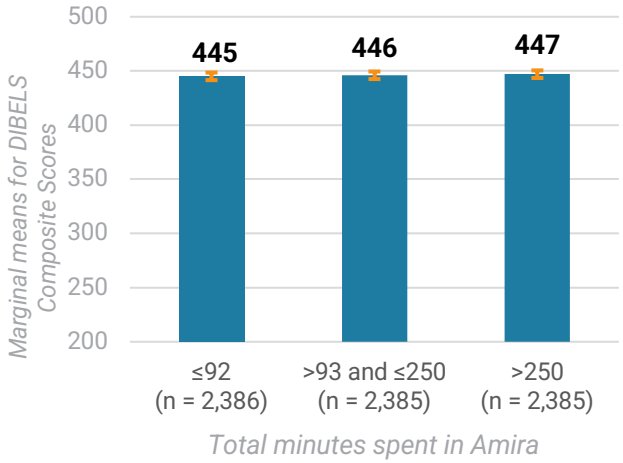


Figure 7. Multi-level models examining the association between total number of minutes and DIBELS scores (Grade 3).

Did Grade 3 students who practiced reading in *Amira* have better literacy outcomes than a matched sample of students who did not have access to *Amira*?

The matched Grade 3 sample demonstrated baseline equivalence (Hedges' g = -0.01; p = .794). The sample of high use students (251–2,237 minutes) and nonusers also met baseline equivalence standards (Hedges' g = -0.07; p = .032). Consequently, researchers then analyzed (a) matched Kindergarten Amira users vs. non-users and (b) high-use Amira users vs. non-users.

Overall, *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant (g = 0.05; p < .001); Figure 8). A Hedges' g value of 0.05 means that if an average Grade 3 student (one who scores right in the middle, at the 50th percentile) had used *Amira*, they would be expected to perform at the **52nd percentile.**

High-use *Amira* users had higher spring 2024 DIBELS scores compared to non-users, and this difference was statistically significant (g = 0.09; p < .001); Figure 8). A Hedges' g value of 0.09 means that if an average Grade 3 student (one who scores right in the middle, at the 50th percentile) had used *Amira* at this level, they would be expected to perform at the **54th percentile.**

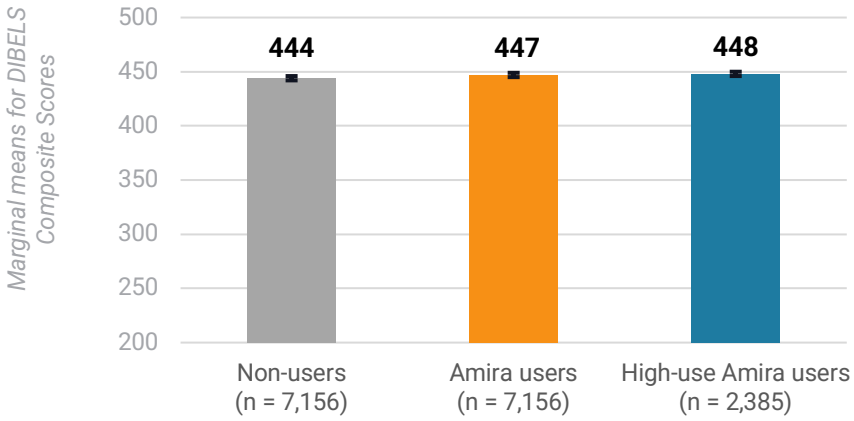


Figure 8. Adjusted mean spring 2024 DIBELS scores for Grade 3 non-users, all Amira users, and high-use Amira users.

LEAP OUTCOME FINDINGS FOR GRADES 4 AND 5 STUDENTS

Researchers created a matched sample of *Amira* users and non-users based on students' spring 2023 LEAP ELA scores, district, gender, race, socioeconomic status, ELL designation, and special education designation. For *Amira* users, researchers categorized usage groups by dividing total minutes spent on the platform (***session time***) into tertiles and the number of passages read into quartiles. As such, the specifications for usage groups differed by grade-level. To address the outcome questions, researchers employed two-level multilevel modeling analyses, with students nested within districts. The models examined the impact of using *Amira* on students' spring 2024 LEAP scores, controlling for spring 2023 LEAP scores and statistically significant demographic variables (e.g., gender, race, socioeconomic status, ELL designation, and special education designation). Researchers conducted these analyses in two parts: 1) correlative analyses focusing solely on *Amira* users, 2) comparative analyses comparing matched samples of *Amira* users and non-users, and 3) comparative analyses examining students in the highest Amira usage group versus non-users, provided baseline equivalence was established.

To allow for better interpretability of results, researchers present marginal means charts below. The vertical lines at the top of each bar represent a 95% confidence interval. Additional information on these analyses and findings can be found in Appendices E and F.

4

What was the association between Amira use and Grade 4 students' LEAP scores?

Total time spent in Amira (minutes). Results showed one statistically significant, positive association between the total number of minutes spent in *Amira* and LEAP scores. Grade 4 students who spent:

- more than 264 minutes in *Amira* (high use) had significantly higher LEAP scores than students who spent 87 or fewer minutes (low use; Hedges' g = 0.04, p = .035).

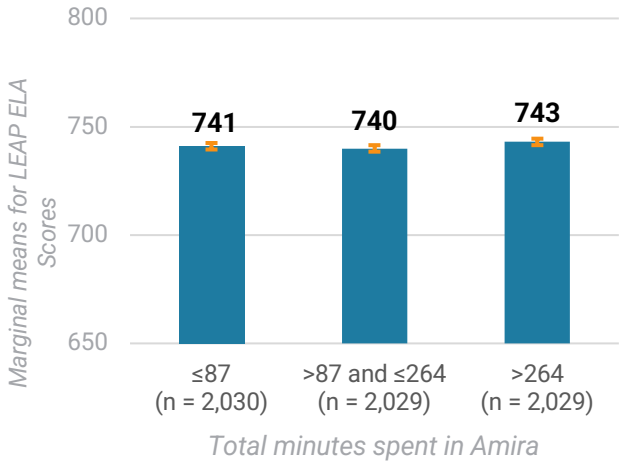


Figure 9. Multi-level models examining the association between total number of minutes and LEAP scores (Grade 4).

Passages read in Amira. Grade 4 students who read more than 58 passages in *Amira* (high use quartile) had significantly higher LEAP scores than students who read 9 or fewer passages (low use quartile; Hedges' g = 0.10, p < .001).

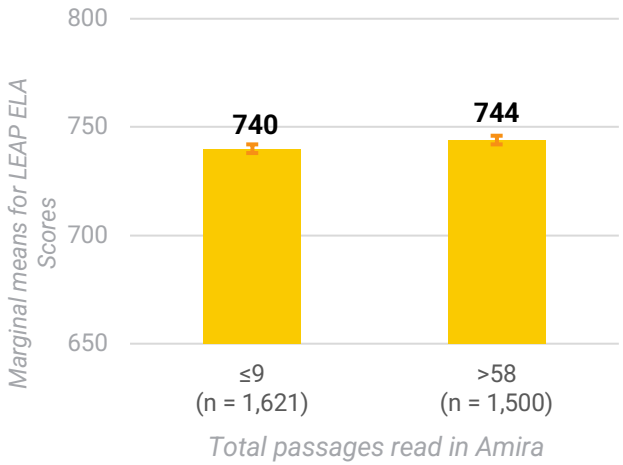


Figure 10. Multi-level models examining the association between total number of passages read and LEAP scores (Grade 4).

Did Grade 4 students who practiced reading in *Amira* have better literacy outcomes than a matched sample of students who did not have access to *Amira*?

The matched Grade 4 sample demonstrated baseline equivalence (Hedges' g = -0.09; p < .001). The sample of high use students (265–2,255 minutes) and nonusers also met baseline equivalence standards (Hedges' g = -0.13; p < .001). Consequently, researchers then analyzed (a) matched Kindergarten Amira users vs. non-users and (b) high-use Amira users vs. non-users.

Overall, *Amira* users had higher spring 2024 LEAP scores compared to non-users, and this difference was statistically significant (g = 0.03; p = .032); Figure 11). A Hedges' g value of 0.03 means that if an average Grade 4 student (one who scores right in the middle, at the 50th percentile) had used *Amira*, they would be expected to perform at the **51st percentile.**

High-use *Amira* users had higher spring 2024 LEAP scores compared to non-users, and this difference was statistically significant (g = 0.07; p < .001); Figure 11). A Hedges' g value of 0.07 means that if an average Grade 4 student (one who scores right in the middle, at the 50th percentile) had used *Amira* at this level, they would be expected to perform at the **53rd percentile.**

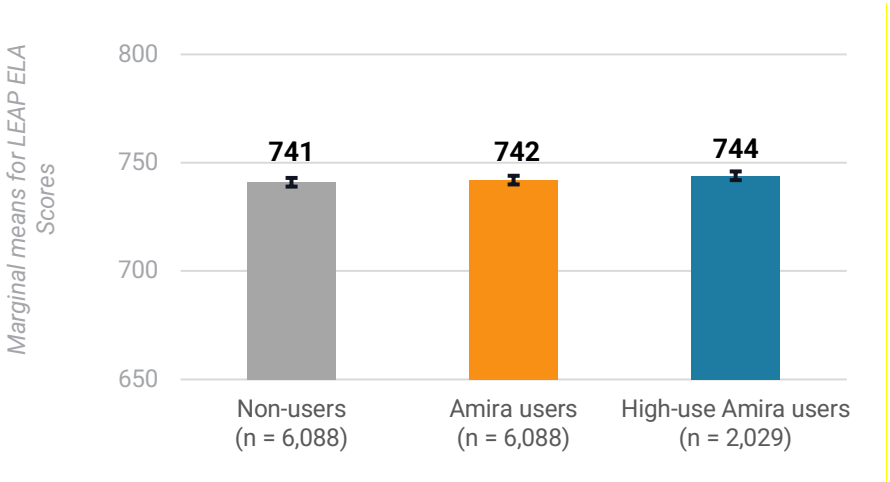


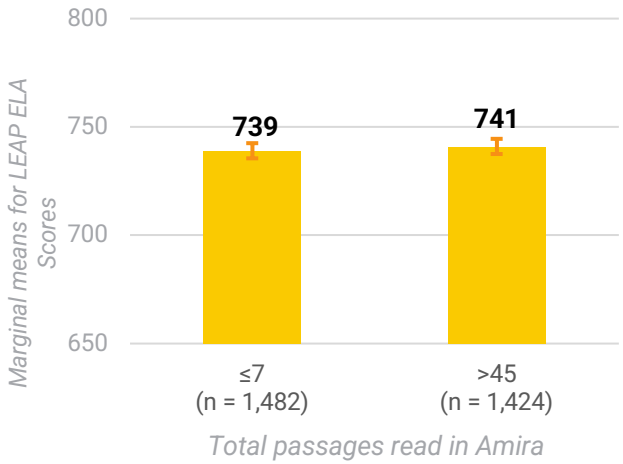
Figure 11. Adjusted mean spring 2024 LEAP scores for Grade 4 non-users, all Amira users, and high-use Amira users.

5

What was the association between Amira use and Grade 5 students' LEAP scores?

Total time spent in Amira (minutes). None of the associations between the total number of minutes spent in *Amira* and LEAP scores were statistically significant.

Passages read in Amira. Grade 5 students who read more than 45 passages in *Amira* (high use quartile) had significantly higher LEAP scores than students who read 7 or fewer passages (low use quartile; Hedges' $g = 0.07$, $p = .003$).



Did Grade 5 students who practiced reading in *Amira* have better literacy outcomes than a matched sample of students who did not have access to *Amira*?

The matched Grade 5 sample demonstrated baseline equivalence (Hedges' $g = -0.01$; $p = .733$). The sample of high use students (187–1,664 minutes) and nonusers also met baseline equivalence standards (Hedges' $g = 0.09$; $p = .001$). Consequently, researchers then analyzed (a) matched Kindergarten Amira users vs. non-users and (b) high-use Amira users vs. non-users.

Overall, *Amira* users had higher spring 2024 LEAP scores compared to non-users, and this difference was statistically significant ($g = 0.04$; $p = .005$); Figure 12). A Hedges' g value of 0.04 means that if an average Grade 5 student (one who scores right in the middle, at the 50th percentile) had used *Amira*, they would be expected to perform at the **52nd percentile.**

High-use *Amira* users had higher spring 2024 LEAP scores compared to non-users, and this difference was statistically significant ($g = 0.06$; $p < .001$); Figure 12). A Hedges' g value of 0.06 means that if an average Grade 5 student (one who scores right in the middle, at the 50th percentile) had used *Amira* at this level, they would be expected to perform at the **52nd percentile.**

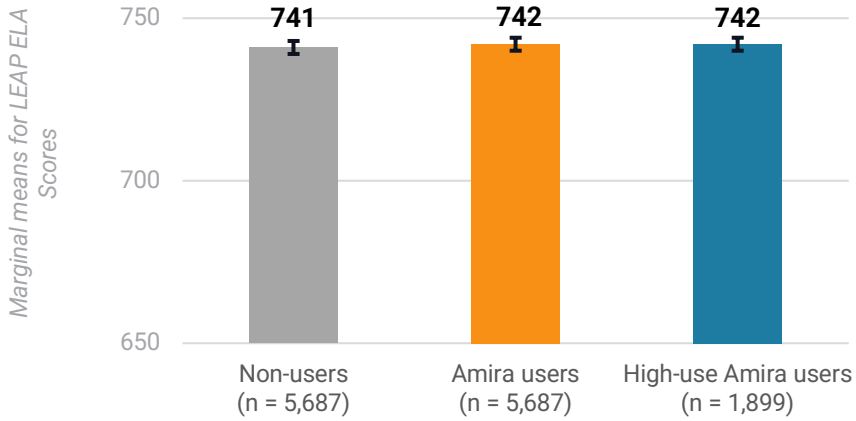


Figure 12. Adjusted mean spring 2024 LEAP scores for Grade 5 non-users, all Amira users, and high-use Amira users.

CONCLUSIONS

In conclusion, the study found a consistently positive and statistically significant association between the time spent on *Amira* (***session time***) and students' achievements as measured by DIBELS and LEAP assessments.

Overall, researchers found modest Hedges' *g* values and impact findings were consistently positive and statistically significant. Since the user group was not modified in terms of dosage for the main comparative analyses, these findings are reflective of real-world implementation. Moving forward, *Amira Learning* could consider conducting a randomized controlled trial (RCT) to further validate these results and/or investigate the reasons behind the lower-than-expected usage.

Given the positive findings, this study provides results to satisfy ESSA evidence requirements for Level II (Moderate Evidence). Specifically, this study met the following, minimum criteria for Level II:

- ✓ Proper design and implementation
- ✓ Baseline equivalence for treatment and comparison groups
- ✓ Statistical controls through covariates
- ✓ At least 350 students in the analysis sample
- ✓ Representative, multi-site study
- ✓ At least one statistically significant, positive effect of the intervention on outcomes

REFERENCES

Every Student Succeeds Act, Pub. L. No. 114-95 (2015). https://www.govinfo.gov/app/details/PLAW-114publ95.

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

APPENDIX A. KINDERGARTEN: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS

Table A1. Student demographics by group for matched sample

Characteristic	*Amira* students (*n* = 5,765)		Non–users (*n* = 5,765)		Total sample (*n* = 11,530)	
	Percent	*n*	Percent	*n*	Percent	*n*
Race $\chi^2(6) = 31.60$, $p < .001$						
Asian	1%	80	2%	129	2%	209
Black or African American	38%	2,197	37%	2,151	38%	4,348
Hispanic	21%	1,210	20%	1,144	20%	2,354
Two or more races	4%	235	4%	233	4%	468
White	35%	2,016	36%	2,062	35%	4,078
Socioeconomic Status (low income flag) $\chi^2(1) = 0.07$, $p = .791$						
Yes	77%	4,432	77%	4,420	77%	8,852
No	23%	1,333	23%	1,345	23%	2,678
Gender $\chi^2(1) = 0.48$, $p = .491$						
Female	50%	2,882	51%	2,919	50%	5,801
Male	50%	2,883	49%	2,846	50%	5,729
English Language Learner $\chi^2(1) = 0.78$, $p = .378$						
Yes	15%	892	15%	858	15%	1750
No	85%	4,873	85%	4,907	85%	9,780
Special Education Status $\chi^2(1) = 1.23$, $p = .268$						
Yes	11%	645	11%	608	11%	1253
No	89%	5,120	89%	5,157	89%	10,277

Table A2. Baseline equivalence analysis of fall 2023 DIBELS scores

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Treatment Condition (Hedges' *g* = -0.01)	-0.32	1.09	-0.29	.771
Gender	-1.61	0.82	-1.97	.049
Race	-1.05	0.26	-4.01	<.001
SES	-24.16	1.08	-22.46	<.001
ELL	-25.10	1.22	-20.57	<.001
Special education	-8.09	1.33	-6.10	<.001
District-level random effects	50.06	22.09	184.60	<.001

Table A3. Descriptive statistics for the Amira usage categories

Usage categories: total minutes spent on Amira		*n*	Mean	SD
Tertile 1	1-96	1,922	48	26
Tertile 2	97-271	1,922	175	50
Tertile 3	272-2,755	1,921	533	247

Overall Association between *Amira* Usage and Kindergarten Students' Spring 2024 DIBELS Scores

Table A4. Students' spring 2024 DIBELS scores by time spent on *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Moderate Use vs. Low Use (Hedges' g = 0.09)	4.11	1.18	3.48	.001
High Use vs. Low Use (Hedges' g = 0.19)	8.98	1.20	7.49	<.001
Fall 2023 DIBELS scores	0.66	0.01	58.80	<.001
Gender	0.16	0.96	0.16	.870
Race	0.71	0.32	2.23	.026
SES	-9.04	1.30	-6.95	<.001
ELL	29.00	1.48	19.65	<.001
Special Education	-14.30	1.54	-9.29	<.001
District-level random effects	65.50	39.73	174.70	<.001

Difference Between Kindergarten Students who used *Amira* and Students Who Did Not Use the Program

Table A5. Differences between spring 2024 DIBELS scores by condition (any use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Students who used Amira vs. Students who did not use the program (Hedges' *g* = 0.11)	5.22	0.93	5.62	<.001
Fall 2023 DIBELS scores	0.66	0.01	84.69	<.001
Race	1.12	0.22	5.07	<.001
SES	-7.24	0.92	-7.85	<.001
ELL	22.84	1.04	21.91	<.001
Special Education	-15.51	1.10	-14.06	<.001
District-level random effects	121.56	50.96	565.20	<.001

Table A6. Differences between spring 2024 DIBELS scores by condition (high use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High-use Amira students vs. Students who did not use the program (Hedges' *g* = 0.21)	10.32	1.16	8.88	<.001
Fall 2023 DIBELS scores	0.66	0.01	84.62	<.001
Race	1.14	0.22	5.19	<.001
SES	-7.39	0.92	-8.03	<.001
ELL	22.42	1.04	21.53	<.001
Special Education	-15.38	1.10	-13.97	<.001
District-level random effects	122.36	51.30	565.94	<.001

APPENDIX B. GRADE 1: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS

Table B1. Student demographics by group for matched sample

Characteristic	*Amira* students (*n* = 7,382)		Non–users (*n* = 7,382)		Total sample (*n* = 14,764)	
	Percent	*n*	Percent	*n*	Percent	*n*
Race $\chi^2(1) = 0.00, p = 1.00$						
Asian	2%	129	2%	129	2%	258
Black or African American	39%	2,911	39%	2,911	39%	5,822
Hispanic	18%	1,330	18%	1,330	18%	2,660
Two or more races	5%	340	5%	340	5%	680
White	36%	2,640	36%	2,640	36%	5,280
Socioeconomic Status (low income flag) $\chi^2(1) = 0.00, p = .100$						
Yes	77%	5,657	77%	5,657	77%	11,314
No	23%	1,725	23%	1,725	23%	3,450
Gender $\chi^2(1) = 23.13, p < 0.01$						
Female	54%	3,978	50%	3,686	52%	7,664
Male	46%	3,404	50%	3,696	48%	7,100
English Language Learner $\chi^2(1) = 0.00, p = 1$						
Yes	12%	874	12%	874	12%	1748
No	88%	6,508	88%	6,508	88%	13,016
Special Education Status $\chi^2(1) = 0.00, p = 1$						
Yes	14%	1,049	14%	1,049	14%	2,098
No	86%	6,333	86%	6,333	86%	12,666

Table B2. Baseline equivalence analysis of fall 2023 DIBELS scores

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Treatment Condition (Hedges' *g* = 0.14)	4.01	0.65	6.16	<.001
Gender	-2.75	0.46	-6.05	<.001
SES	0.15	0.14	1.04	.297
ELL	-11.84	0.58	-20.26	<.001
Special education	-5.37	0.75	-7.21	<.001
District-level random effects	-11.21	0.66	-17.08	<.001

Table B3. Descriptive statistics for the usage categories for Amira

Usage categories: total minutes spent on Amira		*n*	Mean	SD
Tertile 1	2-146	2,461	72	40
Tertile 2	147-392	2,461	254	70
Tertile 3	393-3,125	2,460	681	271

Overall Association between *Amira* Usage and Grade 1 Students' Spring 2024 DIBELS Scores

Table B4. Students' spring 2024 DIBELS scores by time spent on *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Moderate Use vs. Low Use (Hedges' *g* = 0.08)	3.41	0.74	4.62	<.001
High Use vs. Low Use (Hedges' *g* = 0.19)	7.95	0.77	10.33	<.001
Fall 2023 DIBELS scores	1.10	0.01	103.26	<.001
Gender	-2.52	0.62	-4.06	<.001
Race	0.86	0.19	4.45	<.001
SES	-4.55	0.79	-5.75	<.001
ELL	19.86	1.06	18.81	<.001
Special Education	-9.05	0.88	-10.32	<.001
District-level random effects	24.48	15.11	160.69	<.001

Difference Between Grade 1 Students who used *Amira* and Students Who Did Not Use the Program

Table B5. Differences between spring 2024 DIBELS scores by condition (any use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Students who used Amira vs. Students who did not use the program (Hedges' *g* = 0.10)	4.12	0.62	6.62	<.001
Fall 2023 DIBELS scores	1.07	0.01	138.52	<.001
Race	1.24	0.13	9.18	<.001
SES	-5.19	0.56	-9.27	<.001
ELL	14.32	0.70	20.36	<.001
Special Education	-9.55	0.62	-15.46	<.001
District-level random effects	19.89	8.58	230.84	<.001

Table B6. Differences between spring 2024 DIBELS scores by condition (high use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High-use Amira students vs. Students who did not use the program (Hedges' *g* = 0.22)	8.80	0.79	11.20	<.001
Fall 2023 DIBELS scores	1.07	0.01	138.86	<.001
Gender	-0.98	0.43	-2.28	.023
Race	1.21	0.13	9.03	<.001
SES	-5.22	0.56	-9.36	<.001
ELL	13.99	0.70	19.87	<.001
Special Education	-9.62	0.62	-15.43	<.001
District-level random effects	24.36	10.42	288.05	<.001

APPENDIX C. GRADE 2: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS

Table C1. Student demographics by group for matched sample

Characteristic	*Amira* students (*n* = 7,454)		Non–users (*n* = 7,454)		Total sample (*n* = 14,908)	
	Percent	*n*	Percent	*n*	Percent	*n*
Race $\chi^2(1) = 11.27, p = .080$						
Asian	2%	131	2%	152	2%	283
Black or African American	38%	2,844	38%	2,832	38%	5,676
Hispanic	18%	1,340	17%	1,300	18%	2,640
Two or more races	4%	312	4%	300	4%	612
White	37%	2,793	38%	2,835	38%	5,628
Socioeconomic Status (low income flag) $\chi^2(1) = 0.06, p = .805$						
Yes	75%	5,598	75%	5,611	75%	11,209
No	25%	1,856	25%	1,843	25%	3,699
Gender $\chi^2(1) = 0.02, p = .896$						
Female	50%	3,706	50%	3,714	50%	7,420
Male	50%	3,748	50%	3,740	50%	7,488
English Language Learner $\chi^2(1) = 0.00, p = 1.00$						
Yes	11%	812	11%	812	11%	1624
No	89%	6,642	89%	6,642	89%	13,284
Special Education Status $\chi^2(1) = 0.12, p = .729$						
Yes	15%	1,104	15%	1,089	15%	2,193
No	85%	6,350	85%	6,365	85%	12,715

Table C2. Baseline equivalence analysis of fall 2023 DIBELS scores

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Treatment Condition (Hedges' *g* = 0.09)	3.00	0.73	4.13	<.001
Gender	-2.81	0.50	-5.67	<.001
Race	0.76	0.16	4.89	<.001
SES	-13.26	0.63	-21.11	<.001
ELL	-5.45	0.83	-6.54	<.001
Special education	-17.68	0.71	-25.02	<.001
District-level random effects	22.33	9.56	281.89	<.001

Table C3. Descriptive statistics for the usage categories for Amira

Usage categories: total minutes spent on Amira		*n*	Mean	SD
Tertile 1	2-113	2,485	58	31
Tertile 2	114-318	2,485	199	57
Tertile 3	319-3012	2,484	612	278

Overall Association between *Amira* Usage and Grade 2 Students' Spring 2024 DIBELS Scores

Table C4. Students' spring 2024 DIBELS scores by time spent on *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Moderate Use vs. Low Use (Hedges' *g* = 0.02)	0.62	0.60	1.03	.303
High Use vs. Low Use (Hedges' *g* = 0.06)	2.25	0.65	3.48	.001
Fall 2023 DIBELS scores	0.88	0.01	109.09	<.001
Race	1.11	0.16	6.95	<.001
SES	1.76	0.64	-2.76	.006
ELL	23.08	0.88	26.26	<.001
Special Education	7.92	0.70	-11.24	<.001
District-level random effects	33.39	20.90	88.15	<.001

Difference Between Grade 2 Students who used *Amira* and Students Who Did Not Use the Program

Table C5. Differences between spring 2024 DIBELS scores by condition (any use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Students who used Amira vs. Students who did not use the program (Hedges' *g* = 0.08)	2.72	0.51	5.34	<.001
Fall 2023 DIBELS scores	0.89	0.01	155.62	<.001
Race	1.06	0.11	9.77	<.001
SES	-2.49	0.44	-5.61	<.001
ELL	14.37	0.58	24.73	<.001
Special Education	-7.91	0.50	-15.96	<.001
District-level random effects	16.61	7.10	257.31	<.001

Table C6. Differences between spring 2024 DIBELS scores by condition (high use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High-use Amira students vs. Students who did not use the program (Hedges' *g* = 0.12)	4.50	0.64	7.04	<.001
Fall 2023 DIBELS scores	0.89	0.01	155.54	<.001
Gender	-0.61	0.34	-1.78	.075
Race	1.07	0.11	9.83	<.001
SES	-2.54	0.44	-5.72	<.001
ELL	14.09	0.58	24.16	<.001
Special Education	-8.01	0.50	-15.96	<.001
District-level random effects	17.30	7.38	269.99	<.001

APPENDIX D. GRADE 3: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS

Table D1. Student demographics by group for matched sample

Characteristic	*Amira* students (*n* = 7,156)		Non–users (*n* = 7,156)		Total sample (*n* = 14,312)	
	Percent	*n*	Percent	*n*	Percent	*n*
Race $\chi^2(1) = 4.65, p = .589$						
Asian	2%	121	2%	145	2%	266
Black or African American	37%	2,666	38%	2,688	37%	5,354
Hispanic	18%	1,306	17%	1,245	18%	2,551
Two or more races	4%	318	4%	315	4%	633
White	38%	2,701	38%	2,710	38%	5,411
Socioeconomic Status (low income flag) $\chi^2(1) = 2.54, p = .111$						
Yes	74%	5,302	75%	5,366	75%	10,668
No	26%	1,854	25%	1,790	25%	3,644
Gender $\chi^2(1) = 0.00, p = .987$						
Female	49%	3,535	49%	3,536	49%	7,071
Male	51%	3,621	51%	3,620	51%	7,241
English Language Learner $\chi^2(1) = 0.00, p = 1.00$						
Yes	9%	618	9%	618	9%	1236
No	91%	6,538	91%	6,538	91%	13,076
Special Education Status $\chi^2(1) = 2.98, p = .084$						
Yes	15%	1,107	15%	1,062	15%	2,169
No	85%	6,049	85%	6,094	85%	12,143

Table D2. Baseline equivalence analysis of fall 2023 DIBELS scores

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Treatment Condition (Hedges' *g* = -0.01)	-0.21	0.79	-0.26	.794
Gender	-4.96	0.56	-8.93	<.001
Race	1.41	0.17	8.06	<.001
SES	-13.27	0.69	-19.11	<.001
ELL	-8.65	1.03	-8.38	<.001
Special education	-24.24	0.78	-30.93	<.001
District-level random effects	31.14	13.20	352.73	<.001

Table D3. Descriptive statistics for the usage categories for Amira

Usage categories: total minutes spent on Amira		*n*	Mean	SD
Tertile 1	2-92	2,386	46	25
Tertile 2	93-250	2,385	159	45
Tertile 3	251-2237	2,385	473	214

Overall Association between *Amira* Usage and Grade 3 Students' Spring 2024 DIBELS Scores

Table D4. Students' spring 2024 DIBELS scores by time spent on *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Moderate Use vs. Low Use (Hedges' *g* = 0.01)	0.35	0.67	0.53	.598
High Use vs. Low Use (Hedges' *g* = 0.04)	1.71	0.69	2.48	.013
Fall 2023 DIBELS scores	0.89	0.01	108.70	<.001
SES	-1.39	0.66	-2.09	.037
ELL	27.35	1.04	26.28	<.001
Special Education	-8.56	0.78	-11.05	<.001
District-level random effects	25.17	13.97	122.45	<.001

Difference Between Grade 3 Students who used *Amira* and Students Who Did Not Use the Program

Table D5. Differences between spring 2024 DIBELS scores by condition (any use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Students who used Amira vs. Students who did not use the program (Hedges' *g* = 0.05)	2.17	0.55	3.97	<.001
Fall 2023 DIBELS scores	0.89	0.01	152.27	<.001
Race	0.29	0.12	2.37	.018
SES	-2.26	0.49	-4.60	<.001
ELL	16.70	0.72	23.13	<.001
Special Education	-10.35	0.56	-18.46	<.001
District-level random effects	10.31	4.61	132.11	<.001

Table D6. Differences between spring 2024 DIBELS scores by condition (high use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High-use Amira students vs. Students who did not use the program (Hedges' *g* = 0.09)	3.81	0.69	5.56	<.001
Fall 2023 DIBELS scores	0.89	0.01	152.41	<.001
Race	0.30	0.12	2.45	.014
SES	-2.20	0.49	-4.47	<.001
ELL	16.44	0.72	22.69	<.001
Special Education	-10.31	0.56	-18.40	<.001
District-level random effects	11.08	4.93	141.81	<.001

APPENDIX E. GRADE 4: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS

Table E1. Student demographics by group for matched sample

Characteristic	*Amira* students (*n* = 6,088)		Non–users (*n* = 6,088)		Total sample (*n* = 12,176)	
	Percent	*n*	Percent	*n*	Percent	*n*
Race $\chi^2(1) = 6.43, p = .377$						
Asian	2%	99	2%	124	2%	223
Black or African American	38%	2,284	38%	2,288	38%	4,572
Hispanic	17%	1,030	16%	972	16%	2,002
Two or more races	4%	226	4%	215	4%	441
White	40%	2,412	40%	2,441	40%	4,853
Socioeconomic Status (low income flag) $\chi^2(1) = 0.49, p = .485$						
Yes	71%	4,340	71%	4,305	71%	8,645
No	29%	1,748	29%	1,783	29%	3,531
Gender $\chi^2(1) = 0.00, p = .986$						
Female	50%	3,015	50%	3,016	50%	6,031
Male	50%	3,073	50%	3,072	50%	6,145
English Language Learner $\chi^2(1) = 0.00, p = 1.00$						
Yes	8%	469	8%	469	8%	938
No	92%	5,619	92%	5,619	92%	11,238
Special Education Status $\chi^2(1) = 0.54, p = .462$						
Yes	15%	929	15%	900	15%	1829
No	85%	5,159	85%	5,188	85%	10,347

Table E2. Baseline equivalence analysis of spring 2023 LEAP scores

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Treatment Condition (Hedges' *g* = -0.09)	-3.85	0.90	-4.30	<.001
Gender	4.20	0.68	6.16	<.001
Race	2.80	0.21	13.05	<.001
SES	-21.34	0.83	-25.71	<.001
ELL	-37.21	1.33	-28.05	<.001
Special education	-29.80	0.97	-30.80	<.001
District-level random effects	54.53	22.91	562.49	<.001

Table E3. Descriptive statistics for the usage categories for Amira

Usage categories: total minutes spent on Amira		*n*	Mean	SD
Tertile 1	2-87	2,030	41	24
Tertile 2	88-264	2,029	168	51
Tertile 3	265-2255	2,029	479	196

Usage categories: total passages read on Amira		*n*	Mean	SD
Quartile 1	1-9	1,621	4	3
Quartile 2	10-26	1,472	17	5
Quartile 3	27-58	1,495	41	9
Quartile 4	59-485	1,500	95	34

Overall Association between *Amira* Usage and Grade 4 Students' Spring 2024 LEAP Scores

Table E4. Students' spring 2024 LEAP scores by time spent on *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Moderate Use vs. Low Use (Hedges' *g* = -0.03)	-1.05	0.66	-1.59	.112
High Use vs. Low Use (Hedges' *g* = 0.04)	1.45	0.69	2.10	.035
Spring 2023 LEAP scores	0.60	0.01	83.38	<.001
Race	0.87	0.17	5.19	<.001
SES	-5.43	0.66	-8.17	<.001

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
ELL	-6.57	1.07	-6.11	<.001
Special Education	-7.02	0.76	-9.20	<.001
District-level random effects	2.79	2.19	11.34	<.001

Table E5. Students' spring 2024 LEAP scores by passages read in *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High Use vs. Low Use (Hedges' g = 0.10)	4.01	0.80	4.99	<.001
Spring 2023 LEAP scores	0.60	0.01	82.43	<.001
Race	0.86	0.17	5.12	<.001
SES	-5.30	0.66	-7.98	<.001
ELL	-7.03	1.08	-6.53	<.001
Special Education	-6.99	0.76	-9.18	<.001
District-level random effects	4.74	3.22	22.89	<.001

Difference Between Grade 4 Students who used *Amira* and Students Who Did Not Use the Program

Table E6. Differences between spring 2024 DIBELS scores by condition (any use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Students who used Amira vs. Students who did not use the program (Hedges' g = 0.03)	1.07	0.50	2.15	.032
Spring 2023 LEAP scores	0.60	0.01	119.91	<.001
Race	0.70	0.12	5.83	<.001
SES	-5.26	0.47	-11.09	<.001
ELL	-5.49	0.76	-7.22	<.001
Special Education	-7.85	0.56	-14.14	<.001
District-level random effects	10.40	4.50	178.05	<.001

Table E7. Differences between spring 2024 LEAP scores by condition (high use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High-use Amira students vs. Students who did not use the program (Hedges' *g* = 0.07)	2.42	0.66	3.67	<.001
Spring 2023 LEAP scores	0.60	0.01	119.89	<.001
Race	0.71	0.12	5.91	<.001
SES	-5.22	0.47	-11.02	<.001
ELL	-5.64	0.76	-7.39	<.001
Special Education	-7.86	0.55	-14.17	<.001
District-level random effects	10.62	4.60	180.11	<.001

APPENDIX F. GRADE 5: ADDITIONAL INFORMATION ON STUDY DESIGN AND METHODS

Table E1. Student demographics by group for matched sample

Characteristic	*Amira* students (*n* = 5,697)		Non–users (*n* = 5,697)		Total sample (*n* = 11,394)	
	Percent	*n*	Percent	*n*	Percent	*n*
Race $\chi^2(1) = 7.48, p = .279$						
Asian	1%	85	2%	99	2%	184
Black or African American	38%	2,175	38%	2,166	38%	4,341
Hispanic	17%	991	16%	929	17%	1920
Two or more races	3%	195	3%	179	3%	374
White	39%	2,225	40%	2,287	40%	4,512
Socioeconomic Status (low income flag) $\chi^2(1) = 0.02, p = .883$						
Yes	73%	4,133	73%	4,140	73%	8,273
No	27%	1,564	27%	1,557	27%	3,121
Gender $\chi^2(1) = 0.05, p = .822$						
Female	49%	2,800	49%	2,812	49%	5,612
Male	51%	2,897	51%	2,885	51%	5,782
English Language Learner $\chi^2(1) = 0.00, p = 1.00$						
Yes	7%	426	7%	426	7%	852
No	93%	5,271	93%	5,271	93%	10,542
Special Education Status $\chi^2(1) = 1.37, p = .242$						
Yes	14%	806	13%	763	14%	1569
No	86%	4,891	87%	4,934	86%	9,825

Table E2. Baseline equivalence analysis of spring 2023 LEAP scores

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Treatment Condition (Hedges' *g* = -0.01)	-0.25	0.72	-0.34	.733
Gender	3.56	0.55	6.52	<.001
Race	2.91	0.17	16.68	<.001
SES	-17.36	0.68	-25.64	<.001
ELL	-33.00	1.08	-30.56	<.001
Special education	-25.97	0.80	-32.54	<.001
District-level random effects	34.34	14.51	427.00	<.001

Table E3. Descriptive statistics for the usage categories for Amira

Usage categories: total minutes spent on Amira		*n*	Mean	SD
Tertile 1	3-69	1,899	34	19
Tertile 2	70-186	1,899	118	33
Tertile 3	187-1664	1,899	416	204

Usage categories: total passages read on Amira		*n*	Mean	SD
Quartile 1	1-7	1,482	4	2
Quartile 2	8-19	1,450	13	3
Quartile 3	20-45	1,341	30	7
Quartile 4	46-444	1,424	92	44

Overall Association between *Amira* Usage and Grade 5 Students' Spring 2024 LEAP Scores

Table E4. Students' spring 2024 LEAP scores by time spent on *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Moderate Use vs. Low Use (Hedges' *g* = 0.01)	0.26	0.58	0.45	.655
High Use vs. Low Use (Hedges' *g* = 0.03)	1.07	0.60	1.79	.073
Spring 2023 LEAP scores	0.66	0.01	81.78	<.001
Gender	2.77	0.47	5.88	<.001
SES	-2.70	0.58	-4.64	<.001
ELL	-6.01	0.99	-6.07	<.001

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Special Education	-8.27	0.71	-11.71	<.001
District-level random effects	19.88	11.00	275.04	<.001

Table E5. Students' spring 2024 LEAP scores by passages read in *Amira*

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High Use vs. Low Use (Hedges' *g* = 0.07)	2.12	0.70	3.06	.002
Spring 2023 LEAP scores	0.66	0.01	80.32	<.001
Gender	2.72	0.47	5.79	<.001
SES	-2.67	0.58	-4.59	<.001
ELL	-6.20	0.99	-6.26	<.001
Special Education	-8.23	0.70	-11.68	<.001
District-level random effects	20.34	11.26	282.57	<.001

Difference Between Grade 5 Students who used *Amira* and Students Who Did Not Use the Program

Table F6. Differences between spring 2024 LEAP scores by condition (any use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
Students who used Amira vs. Students who did not use the program (Hedges' *g* = 0.04)	1.22	0.44	2.78	.005
Spring 2023 LEAP scores	0.68	0.01	121.34	<.001
Gender	2.42	0.33	7.31	<.001
SES	-2.53	0.41	-6.22	<.001
ELL	-6.25	0.68	-9.18	<.001
Special Education	-8.16	0.51	-16.16	<.001
District-level random effects	11.34	4.87	321.48	<.001

Table F7. Differences between spring 2024 LEAP scores by condition (high use vs. no use)

Predictor	Unstd. Beta Coefficient	Standard Error	Test statistic	*p*-value
High-use Amira students vs. Students who did not use the program (Hedges' *g* = 0.06)	1.82	0.56	3.24	.001
Spring 2023 LEAP scores	0.68	0.01	121.15	<.001
Gender	2.41	0.33	7.25	<.001
SES	-2.51	0.41	-6.16	<.001
ELL	-6.35	0.68	-9.29	<.001
Special Education	-8.19	0.51	-16.21	<.001
District-level random effects	11.48582	4.93	322.77	<.001