AN EVALUATION AND TRAINING INSTITUTE REPORT

Amira Learning EISP Program Evaluation Student Impact Memo

2022-2023

Submitted October 2023

EVALUATION AND TRAINING INSTITUTE *A non-profit center for research & evaluation*
100 Corporate Pointe, Suite 387, Culver City, California 90230 || 310-473-8367

INTRODUCTION

Software vendor-specific Impact Memos are designed to help program stakeholders understand the effectiveness of the individual programs participating in Utah's Early Intervention Software Program (EISP). This memo begins with an overview of *Amira Learning* enrollment and usage recommendations and is followed up by two main analyses for the 2022-2023 school year: (1) program implementation, which includes average program usage and the extent to which students met *Amira*'s recommended use criteria; and (2) program impacts, analyses developed to study the impact *Amira Learning* had on students' literacy achievement. Following a presentation of the analyses we summarize the key findings and study limitations.

Program Enrollment and Usage Recommendations

We track software enrollment numbers to understand the reach of each individual vendor across the state. In 2022-2023, *Amira* was used in 17 Local Education Agencies (LEAs), 142 schools and by 24,127 Utah students. As outlined in **Table 1**, enrollment was evenly distributed across grades first through third.

Table 1. 2022-2023 Program Enrollment by Grade

Kindergarten	First Grade	Second Grade	Third Grade
179	7,922	8,010	8,016

Amira provided recommendations for the amount of time that students should use the software program in order to have an impact on literacy achievement. These recommendations included both a range of minutes per week and a total number of weeks in the program. Recommended

weekly use was 30 minutes per week with a total of 30 suggested weeks across all grades (**Table 2**).

Table 2. 2022- 2023 Minimum Usage Recommendations

Kindergarten	First Grade	Second Grade	Third Grade	Suggested Minimum Weeks
30 min/week	30 min/week	30 min/week	30 min/week	30 weeks

PROGRAM IMPLEMENTATION

Studying program implementation prior to measuring the program impact provided a better understanding of the way the program was ultimately used by students. Namely, students must use the program long enough to influence the outcomes under study. Critical to successful program implementation was the amount of time and how consistently a student used the *Amira* software during the school year. In this section we answer the research question: *To what extent did students use the software program as intended?*

For descriptive purposes, **Table 3** shows straight averages for three different program use measurements, (1) average weekly minutes of use, (2) average total minutes of use, and (3) average number of weeks of use through the end of the school year.

Table 3. 2022-2023 Average Program Use by Grade

Grade	N	Ave Weekly Min	Ave Total Min.	Ave Weeks of Use
K	179	16	150	8
1	7,922	18	314	15
2	8,010	20	354	16
3	8,016	19	332	15
Total	24,127	19	332	15

Note. K-3 Data source: vendor usage data prior to merging with Acadience Reading and state SIS data.

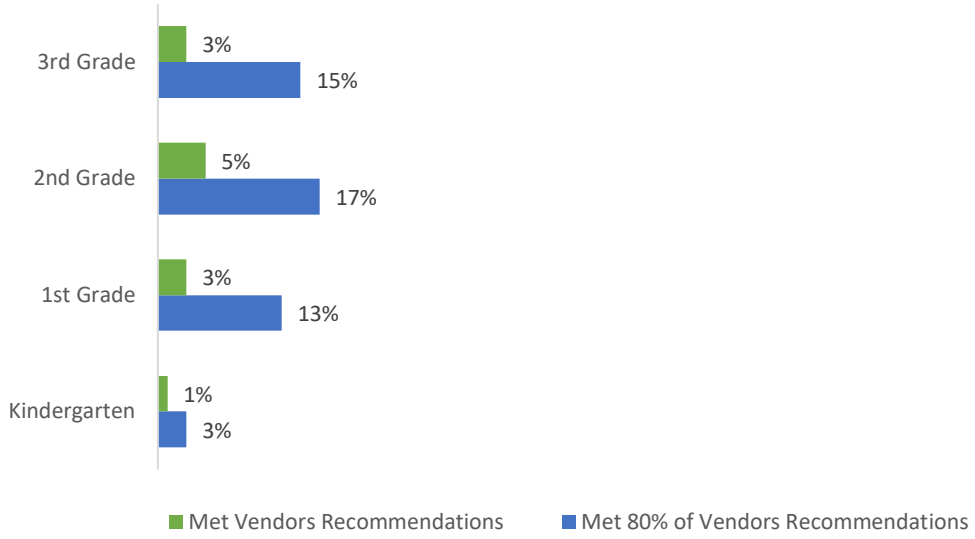
The data presented above represent all students who engaged with the *Amira* program and should be interpreted as the grade-level averages, not as a measure for meeting recommended program use.

Research Question 1: To what extent did the program students meet *Amira's* recommended use criteria?

Only five percent or less of students met *Amira's* use recommendations (**Figure 1**, green bars). We analyzed *Amira's* usage data using two definitions in order to capture students' program participation. Our goal was to align as closely as possible to *Amira's* stated criteria for use. First, we calculated the percentage of students in each grade who met the total weeks as recommended by *Amira* AND whose average weekly minutes (for those weeks) was at or above the recommended minimum. Throughout this memo we refer to this group of students as "met vendor's recommendation."

Next, we expanded *Amira's* stated criteria for use, to provide a larger sample of students who engaged with the program. To do this we calculated the percent of students who met at least 80% of *Amira's* total week recommendation and averaged at least 80% of the weekly minutes' recommendation. We refer to this group of students as "met 80% of vendor's recommendation." As illustrated in **Figure 1** (blue bars), this adjustment increased the overall percentage of program students by 10-15% across most grades, however it should be noted that the vast majority of students were still not achieving a more relaxed definition of the minimum usage requirement. We included both of these use groups in our impact evaluation.

Figure 1. Percentage of Students Meeting *Amira*'s Recommendations for Use



Note: Met *Amira*'s Recommendations reflects 'Met minimum weeks and *average* weekly minutes'
Met 80% of *Amira*'s Recommendations reflects 'Met 80% of weeks and 80% of *average* weekly minutes'

PROGRAM IMPACTS ON LITERACY ACHIEVEMENT

Research Question 2: What were the program impacts on Acadience literacy scores for *Amira* students compared to a matched control group?

Methods

In order to study *Amira's* impact on Acadience literacy test scores, we needed two samples of students, those who participated in the program (Treatment group) and those who were matched to the treatment students across characteristics that influence learning, such as socio-economic status, demographic information, and beginning-of-year Acadience test scores, but who did not participate in the program (Control group). The students who made up our treatment and control groups, within each grade K-3, comprised our analytic sample (i.e., the sample we used in the analysis).

Sampling. Among the overall treatment sample, we created three subgroups of students to account for different levels of program usage. These subgroups were created to evaluate how different levels of use influenced *Amira's* impact on literacy achievement. We considered three main factors in creating the subgroups for *Amira* students: (1) students who met the minimum weeks and average weekly use recommendations as defined by *Amira,* (2) students who met at least 80% of the recommended weeks and average weekly minutes, and (3) the broadest use group, inclusive of those who used the program in any amount throughout the program year (Intent to Treat).

Matching. We then matched control students who did not participate in the program to the three *Amira* usage groups using Coarsened Exact Matching (CEM). We used CEM to match students on grade, beginning-of-year achievement scores and benchmark levels[1], gender, race, English Language Learner (ELL) status, and poverty status. The baseline characteristics of the treatment and control samples can be found in **Appendix A and B**. The matched samples were statistically well-balanced as indicated by L1 coefficients.

Statistical Modeling of Program Impacts on Acadience Test Scores. Ordinary least squares (OLS) regression models were computed for each analytic sample. The OLS models predicted the differences in treatment and control groups' end-of-year group mean scores, while controlling for students' beginning-of-year (BOY) reading scores and key demographics, gender, race, ELL status, SPED designation, and poverty status. We examined treatment effects for each analytic sample based on their usage and grade.

[1] Students in kindergarten, 2nd and 3rd grade were matched on reading composite scores (BOY Comp) and students in 1st grade were matched on nonsense word fluency, correct letter sounds (NWF-CLS) scores.

Results

Key Takeaway. ***Amira*** **showed significant treatment effects for students in first through third grade among students who met *Amira's* usage requirements, and even demonstrated benefits for those meeting a more lenient usage criteria.**

The following results are broken up into two different usage groups of K-3rd grade students and their matched control counterparts, (1) students who met *Amira's* recommended weeks and average minutes, and (2) students who met 80% of recommended weeks and average minutes. This section is focused on participants who engaged with the *Amira* program most closely aligned to the recommendation. Results for the third usage group (ITT), which included the students whose time with the program fell far below the recommend levels, can be found in **Appendix B**.

To determine if the mean score differences could be interpreted as meaningful, we examined their effect sizes. Effect sizes show the magnitude of the difference between two groups on an outcome and are often interpreted as meaningful if they reach a certain minimum threshold. We adapted a set of effect size benchmarks based on categories from Kraft (2020) that were adjusted for early literacy outcome measures: less than 0.10 is *small*, 0.10 to less than .30 is *medium*, and .30 or greater is *large* (M. Kraft, personal communication, October 13, 2023). Effect sizes for all grades and usage groups are referenced in **Appendix C**.

Table 4 presents the model predicted mean scores, mean score differences and effect sizes of matched program students who met *Amira's* recommendations across both average weekly minutes and total weeks. As shown below, all treatment students in grades 1st- 3rd exhibited higher predicted mean scores than their matched control counterparts. Second grade students exhibited the highest mean score differences, with treatment students scoring 21 points higher

than their control counterparts, on average. Among this highest use analytic sample, first and second graders had effect sizes within the large effect size range (g= 0.41 and g=0.45, respectively), while third grade students had an effect size within the medium range (g=0.25). Results are shown for first, second, and third grade only, as kindergarten had an insufficient sample size.

Table 4. MRU Sample Predicted End-of-Year Acadience Reading Composite Mean Scores

Grade	Ctrl		Tr		Dif.	ES
	Mean	SE	Mean	SE		
Kindergarten	ISS					
First Grade	76.37	0.44	87.16	1.94	10.79	**0.41**
Second Grade	284.92	0.71	305.89	3.27	20.97	**0.45**
Third Grade	384.06	1.03	403.87	4.77	19.81	*0.25*

Note. Model covariates were gender, White, special education, low-income, ELL, and BOY reading score. All data points displayed were statistically significant at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater.

Table 5 presents the predicted means, mean score differences and effect sizes of students in the 80% analytic sample. These were program students who met at least 80% of *Amira's* recommended use criteria. Second and third grade had the highest predicted mean score differences between the treatment and control groups, with a difference of 17 and 22 points, respectively. Among the 80% analytic sample, second and third grade students met or exceeded the effect size benchmark for large treatment effects (g=0.30 and g=0.34, respectively). First grade treatment students performed better than control students, by about 4 points, with an effect size (g= 0.15) that fell within the medium effect size range. Again, results are shown for first, second and third grade only, as kindergarten had an insufficient sample size.

Table 5. MRU80 Sample Predicted End-of-Year Acadience Reading Composite Mean Scores

Grade	Ctrl		Tr		Dif.	ES
	Mean	SE	Mean	SE		
Kindergarten	ISS					
First Grade	75.21	0.40	79.42	0.99	4.21	*0.15*
Second Grade	270.00	0.69	287.05	1.74	17.05	**0.30**
Third Grade	380.88	0.90	403.32	2.29	22.44	**0.34**

Note. Model covariates were gender, White, special education, low-income, ELL, and BOY reading score. All data points displayed were statistically significant at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater.

While mean score differences and effect sizes emphasize the effectiveness of the program when compared to a group of non-program students, they do not tell us if the students achieved the goal of reading at grade level. Acadience Reading benchmark categories can be used to further interpret mean scores for this purpose. Generally speaking, *Amira* students' predicted end-of-year literacy scores were within the "at benchmark" range for their grade, which signifies an 80-90 percent likelihood of achieving subsequent reading outcomes (Dynamic Measurement Group, 2021). The benchmark ranges, by grade, are presented in **Appendix D.**

SUMMARY & DISCUSSION

Our evaluation explored two main components of the most recent EISP program year: 1) the success of implementation and the extent to which students were able to engage with the software program as it was intended by *Amira*, and 2) the program's impact on Acadience test scores of the students that were served.

Implementation. The implementation study for *Amira's* program year found that five percent or less of students used the program as intended on both aspects of the recommendation: average

weekly minutes and total weeks. Average minutes of use fell well below the usage recommendations in all grades, with students using the program an average of 10 minutes less per week than what *Amira* recommends (30 weekly minutes). The average number of weeks students used the program (15 weeks) was also below *Amira's* minimum recommended weeks of use (30 weeks).

Impact. Among students who used the program as *Amira* intended, we identified positive literacy achievement outcomes for first, second and third graders compared to matched groups of control students who did not use the program. Additionally, *Amira* students finished the year with literacy scores within the expected "at benchmark" range. These positive program impacts are based on an analysis of students who are using the program according to Amira's guidelines, which underscores the importance of increasing the number of students who are using the program as intended by the vendor.

Limitations. We recognize the potential long-term effects of the pandemic are not fully understood. As a result of the initial covid-19 disruption, it is possible that some students may be navigating greater learning loss than others and are still working to recover from the disruption. We know that all students in our sample may have experienced the initial covid year differently, especially when we consider each grade individually. For example, students in third grade during the 2022-23 school year, were in kindergarten in 2019-20 and first grade in 2020-21 when not all schools reopened to in-person instruction. Without a full longitudinal study, we are limited in our understanding of the potential lasting impacts of covid-19 on EISP student achievement. That said, we are aware that these events and circumstances can impact the engagement and outcomes

with the EISP across the school year. We acknowledge that we were unable to control for all possible scenarios in our analysis.

Recommendations. Students served by *Amira* outperformed the students who were not. Further, the students who were able to use with the software as it was intended by *Amira* also showed greater end-of-year literacy scores relative to those participating below the recommended usage levels in the program. With intentional effort behind accountability and improving consistency of use, more and more students will benefit from the *Amira* program. Several recommendations surfaced from our findings:

- The percentage of students who met the recommended use criteria is notably low (<10%) across all grades and could be increased. We recommend that *Amira* identify and meet with LEAs who have usage below the recommended levels in order to cultivate ways to improve student engagement with the software.

- As a new vendor to the EISP program, we suggest that *Amira* emphasize the importance of consistent program use and encourage LEAs to meet the usage recommendations each week throughout the duration of the school year.

- Our data suggest that *Amira's* program is most impactful for second and third graders. Continue to explore the ways in which program participation can support advanced literacy skills for students in the kindergarten and first grade.

- We also recommend that future evaluations continue to investigate the ways in which *Amira* impacts students of all reading abilities, so that the state can make informed decisions about the most optimal way to support a population of students with diverse learning needs.

REFERENCES

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.

Dynamic Measurement Group, Inc. (2021). *Acadience Reading Benchmark Goals and Composite Score.* https://acadiencelearning.org/wp-content/uploads/2021/11/Acadience-Reading-K-6-Benchmark-Goals-handout_2021_color.pdf

Hill, C. J., Bloom, H. S., Black, A. R. and Lipsey, M. W. (2008). *Empirical Benchmarks for Interpreting Effect Sizes in Research.* Child Development Perspectives, 2: 172–177. doi: 10.1111/j.1750-8606.2008.00061

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2008. *Matching for Causal Inference without Balance Checking*. http://gking.harvard.edu/files/abs/cem-abs.shtml.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189x20912798

APPENDIX A

Table A1. MRU 80 Matched Treatment Balance

	Grade	N	Female	Caucasian	SPED	Low-Income	ELL	BOY Score
Total Treatment Sample	K	ISS						
	1	787	51%	70%	14%	41%	8%	33.83
	2	1113	54%	69%	10%	36%	9%	181.04
	3	858	52%	62%	12%	43%	18%	255.72
Matched MRU 80 Treatment Sample	K	ISS						
	1	775	51%	73%	13%	40%	7%	32.59
	2	1076	54%	71%	10%	35%	7%	182.42
	3	825	52%	64%	12%	42%	16%	257.69

Note: Kindergarten had an insufficient sample size. The matched sample had a multivariate L1 score of 0.000000000000009404. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.000000000000021), White (L1= 0.000000000000011), SPED (L1 = 0.000000000000019), Low-Income (L1= 0.000000000000029), and ELL (L1= 0.0000000000000061).

Table A2. MRU Matched Treatment Balance

	Grade	N	Female	Caucasian	SPED	Low-Income	ELL	BOY Score
Total Treatment Sample	K	ISS						
	1	211	47%	72%	16%	39%	5%	34.89
	2	313	54%	74%	12%	31%	9%	190.51
	3	205	54%	62%	13%	55%	19%	251.95
Matched MRU Treatment Sample	K	ISS						
	1	208	46%	75%	15%	37%	4%	33.07
	2	297	54%	77%	11%	29%	7%	192.69
	3	194	54%	65%	12%	54%	15%	254.99

Note: Kindergarten had an insufficient sample size. The matched sample had a multivariate L1 score of 0.000000000000002668. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.000000000000013), White (L1= 0.0000000000000093), SPED (L1 = 0.0000000000000026), Low-Income (L1= 0.000000000000012), and ELL (L1= 0.0000000000000012).

APPENDIX B

Table B1. *Amira* Predicted Means of End-of-Year Acadience Reading Composite for Matched ITT Treatment and Control Students

Grade	Ctrl		Tr		Dif.	ES
	Mean	SE	Mean	SE		
Kindergarten	ISS					
First Grade	72.91	0.30	75.14	0.37	2.23	0.08
Second Grade	245.50	0.67	251.50	0.75	6.00	*0.10*
Third Grade	362.57	0.80	369.20	0.89	6.64	*0.10*

Note. Model covariates were gender, White, special education, low-income, ELL, and BOY reading score. All data points displayed were statistically significant at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater.

Table B1 presents the predicted means, mean score differences and effect sizes of students in the ITT analytic sample. These were program students who used the *Amira* software in any amount (including very low usage levels) over the course of the program year. Among the ITT analytic sample, no grade exceeded the effect size threshold for large treatment effects (g= 0.30 or greater).

Table B2. ITT Matched Treatment Balance

	Grade	N	Female	Caucasian	SPED	Low-Income	ELL	BOY Score
Total Treatment Sample	K	ISS						
	1	5760	50%	68%	13%	40%	9%	32.51
	2	6202	49%	67%	14%	40%	11%	158.89
	3	5635	50%	66%	15%	39%	14%	238.96
Matched ITT Treatment Sample	K	ISS						
	1	5470	50%	71%	12%	38%	8%	31.43
	2	6015	50%	69%	14%	39%	10%	159.53
	3	5499	50%	67%	15%	38%	13%	240.22

Note: Kindergarten had an insufficient sample size. The matched sample had a multivariate L1 score of 0.000000000000005123. Lower values indicated less imbalance, and the closer to zero the better the two samples are balanced across covariates. Additionally, all covariates in the matched sample were found to be balanced: Female (L1= 0.000000000000006), White (L1= 0.000000000000012), SPED (L1 = 0.000000000000013), Low-Income (L1= 0.000000000000018), and ELL (L1= 0.00000000000001).

APPENDIX C

Effect sizes describe the magnitude of the difference between two groups on an outcome measure. We adapted a set of effect size benchmarks based on categories from Kraft (2020) that were adjusted for early literacy outcome measures: less than 0.10 is *small*, 0.10 to less than .30 is *medium* and .30 or greater is *large* (M. Kraft, personal communication, October 13, 2023). There are multiple ways to interpret effect sizes, including the use of categories such as small, medium, or large (e.g., Cohen, 1988; Kraft, 2020), or using a minimum threshold (Hill 2008). Variations of both approaches are widely used and accepted, yet both require careful considerations of the research design and key study components (such as sample, measures, etc.). Our effect size interpretation approach uses a categorical range based on effect sizes for similar types of research, studying similar interventions (early literacy programs) and with similar populations (elementary students). Specifically, the range used in the current study represents the benchmarks for early literacy found in a summary of meta-analyses of relevant and similar educational studies, as well as the direct recommendation from the author (Kraft, 2020; M. Kraft, personal communication, October 13, 2023).

Table C1. *Amira* Effect Sizes by Grade and Usage Level

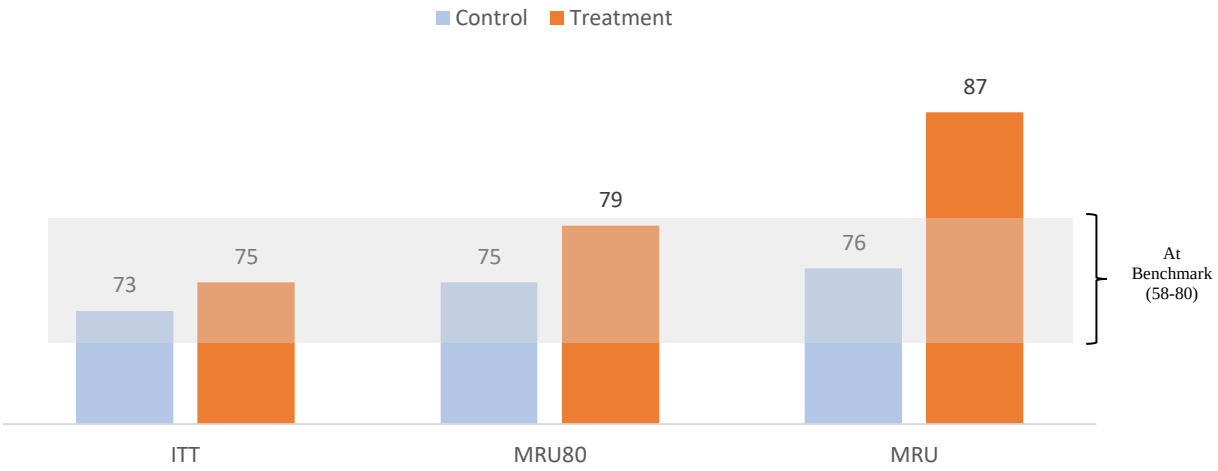
Grade	Intent to Treat	Met 80% of Rec.	Met Rec.
K		ISS	
1	0.08	*0.15*	**0.41**
2	*0.10*	**0.30**	**0.45**
3	*0.10*	**0.34**	*0.25*

Data source: Matched K-3 ITT, MRU80, MRU samples[2]. All effect sizes displayed were statistically significant at p≤ .05. Hedges' g effect size benchmarks are indicated in the table as follows: Small: 0 to < .10; *Medium*, italicized text: .10 < .30, **Large**: bold and underlined text: .30 or greater. Kindergarten had an insufficient sample size.

[2] First Grade- ITT ctrl= 7047.911, tr= 5638; MRU80 ctrl= 4989.714, tr=775; MRU- ctrl= 4372.464, tr=208; Second Grade sample size- ITT ctrl= 7519.188, tr= 6015; MRU80 ctrl= 6927.655, tr=1076; MRU ctrl= 6243.373, tr= 297; Third Grade sample size- ITT ctrl= 6874.1505, tr=5499; MRU80 ctrl=5311.631, tr= 825 MRU ctrl= 4078.163, tr=194

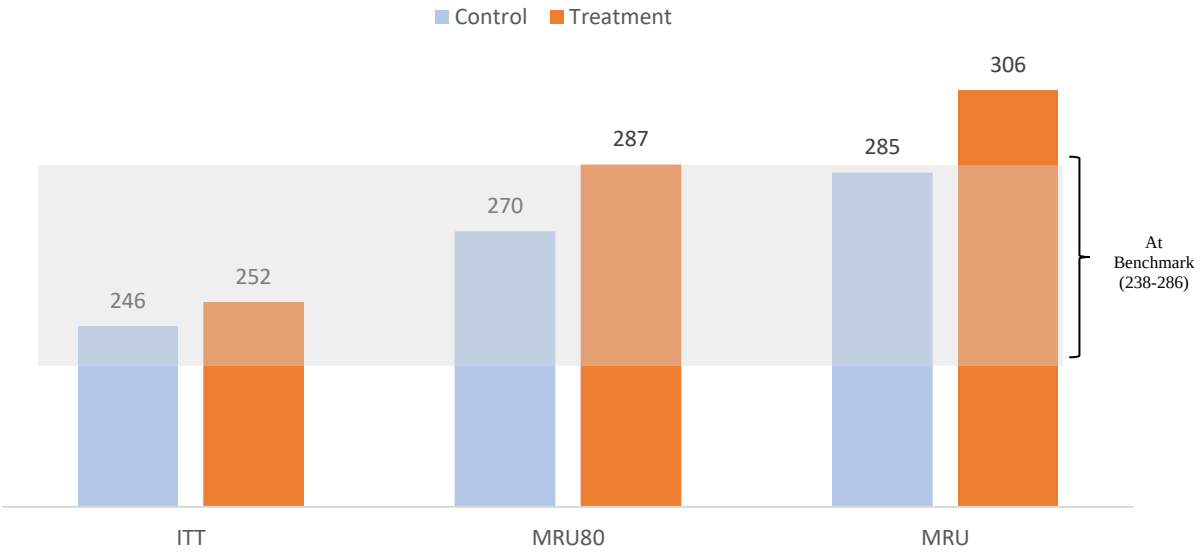
APPENDIX D

Figure D1. *Amira* First Grade Predicted Mean Scores by Usage Level and Matched Sample



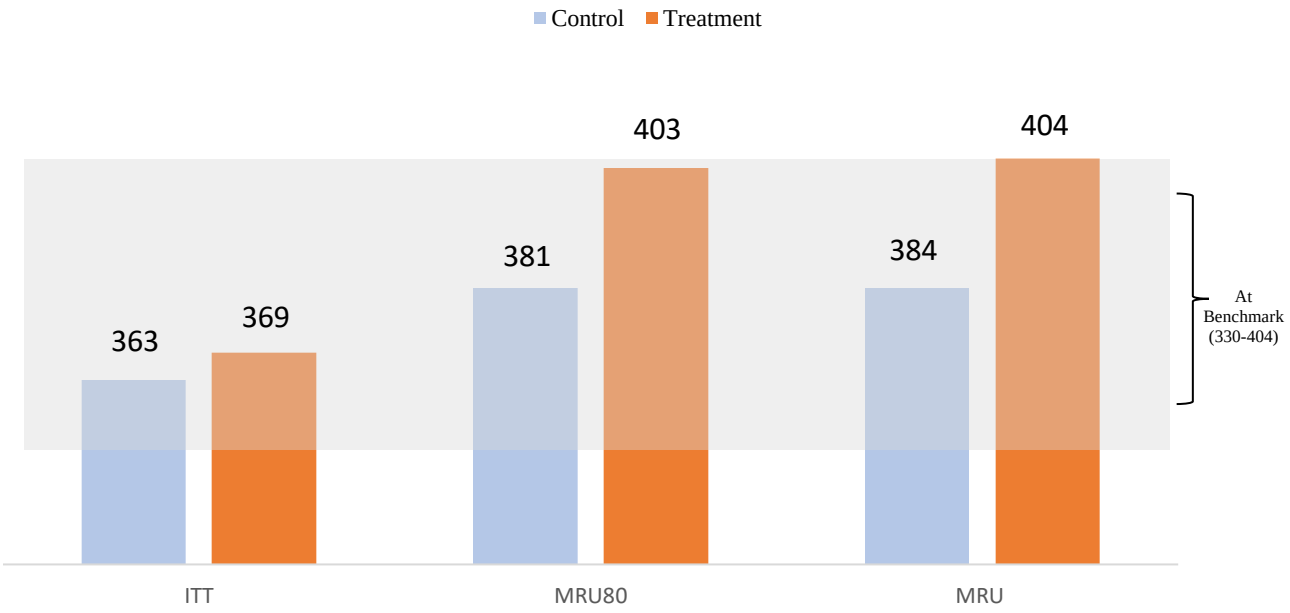
Data source: Matched first ITT, MRU80 and MRU samples. All mean comparisons displayed in the figure were statistically significant at p≤ .05. First grade end-of-year predicted outcomes were measured with the Nonsense Word Fluency- Correct Letter Sounds scale and has a different range than the reading composite scale. Students scoring ***At Benchmark*** *(58-80), or* ***Above Benchmark*** *goal (81 or greater) have the odds in their favor (approximately 80% to 90% overall) of achieving later important reading outcomes.*

Figure D2. *Amira* Second Grade Predicted Mean Scores by Usage Level and Matched Sample



Data source: Matched second grade ITT, MRU80 and MRU samples. All mean comparisons displayed in the table were statistically significant at p≤ .05.

Figure D3. *Amira* Third Grade Predicted Mean Scores by Usage Level and Matched Sample



Data source: Matched third grade ITT, MRU80 and MRU samples. All mean comparisons displayed in the table were statistically significant at p≤ .05.

Evaluation and Training Institute
100 Corporate Pointe, Suite 387
Culver City, CA 90230
www.eticonsulting.org

For more information on the
Evaluation and Training Institute, contact ETI:

Jon Hobbs, Ph.D., President
Phone: 310-473 8367
jhobbs@eticonsulting.org