RESEARCH REPORT

EVALUATION AND TRAINING INSTITUTE
www.eticonsulting.org

AN EVALUATION AND TRAINING INSTITUTE REPORT

Imagine Learning EISP Program Evaluation Student Impact Memo

2022-2023

imagine learning

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 *Imagine 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 *Imagine Learning's* recommended use criteria; and (2) program impacts, analyses developed to study the impact *Imagine Language & Literacy* 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, *Imagine Language & Literacy* was used in 15 Local Education Agencies (LEAs), 69 schools and by 17,042 Utah students. As outlined in **Table 1**, enrollment was evenly distributed across all grades.

Table 1. 2022-2023 Program Enrollment by Grade

Kindergarten	First Grade	Second Grade	Third Grade
3,528	4,587	4,660	4,267

Imagine Learning 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 varied by grade, from 40 to 50 minutes per week, with a total of 18 suggested weeks across all grades (**Table 2**).

Table 2. 2022- 2023 Minimum Usage Recommendations

Kindergarten	First Grade	Second Grade	Third Grade	Suggested Minimum Weeks
40 min/week	50 min/week	50 min/week	50 min/week	18 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 *Imagine Language & Literacy* 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	3,528	38	850	20
1	4,587	45	1,053	21
2	4,660	41	955	20
3	4,267	37	811	19
Total	17,042	40	923	20

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

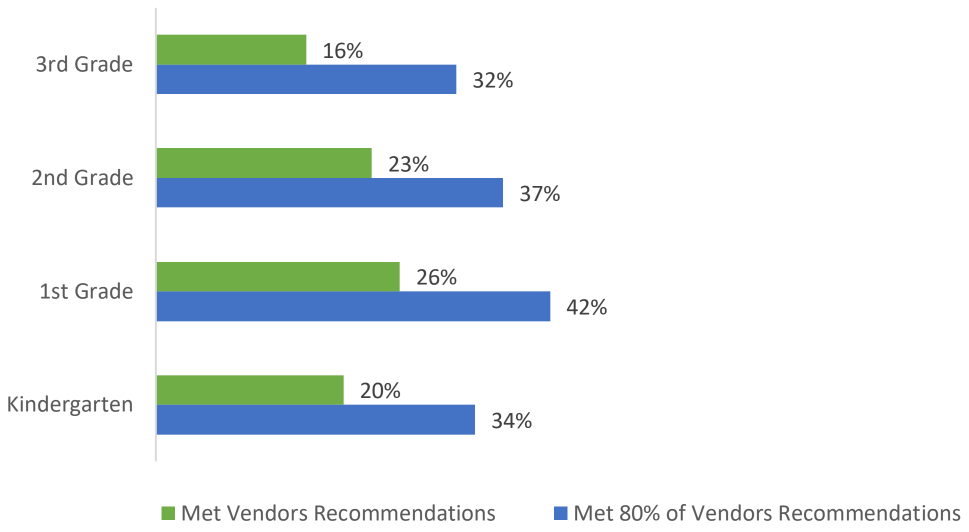
The data presented above represent all students who engaged with the *Imagine Language & Literacy* program and should be interpreted as the grade-level averages, not as a measure for meeting recommended program use.

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

About twenty to thirty percent of students met *Imagine Learning's* use recommendations (**Figure 1**, green bars). We analyzed *Imagine Learning's* usage data using two definitions in order to capture students' program participation. Our goal was to align as closely as possible to *Imagine Learning's* stated criteria for use. First, we calculated the percentage of students in each grade who met the total weeks as recommended by *Imagine Learning* 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 calculated the percent of students who met at least 80% of *Imagine Learning'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." While this expanded *Imagine Learning's* stated criteria for use, it provided a larger sample of students who engaged with the program. As illustrated in **Figure 1** (blue bars), this adjustment increased the overall percentage of program students by about 15% across all grades. We included both of these use groups in our impact evaluation.

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



Note: Met *Imagine Learning*'s Recommendations reflects 'Met minimum weeks and *average* weekly minutes'
Met 80% of *Imagine Learning*'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 *Imagine Learning* students compared to a matched control group?

Methods

In order to study *Imagine Language & Literacy'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 EISP program (Control group). The students who made up our treatment and control groups, within each grade K-3, were considered 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 the program's impact on literacy achievement. We considered three main factors in creating the subgroups for *Imagine Learning* students: (1) students who met the minimum weeks and average weekly use recommendations as defined by *Imagine Learning,* (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 *Imagine Learning* 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 matched treatment 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. **EISP students using *Imagine Language & Literacy* as the vendor recommended in grades K-2 achieved higher predicted literacy mean scores compared to students not using the program, however, only kindergarten achieved large treatment effect sizes.**

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 *Imagine Learning'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 *Imagine Language & Literacy* 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 recommended 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 predicted means, mean score differences and effect sizes of matched program students who met *Imagine Learning's* recommendations across both average weekly minutes and total weeks. Results are shown for kindergarten through second grade only due to non-significance at third grade. Kindergarten students exhibited the highest mean score differences between the treatment and control groups, with treatment students scoring 22 points higher than their control counterparts, on average. Kindergarteners had the largest effect size (g= 0.55) which fell within the range for large treatment effects. First and second grade students also performed better than the control group, with a difference of 3 points and 6 points, respectively. Both first grade (g = 0.13) and second grade students (g = 0.11), had an effect size that fell within the benchmark for medium treatment effects.

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

Grade	Ctrl		Tr		Dif.	ES
	Mean	SE	Mean	SE		
Kindergarten	146.21	0.56	168.44	1.59	22.23	**0.55**
First Grade	86.07	0.28	89.50	0.82	3.43	*0.13*
Second Grade	293.53	0.61	299.84	1.75	6.30	*0.11*
Third Grade	NS					

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 *Imagine Learning's* recommended use criteria. Similar to higher use students, kindergarten had the highest predicted mean score differences between the treatment and control groups, with the effect size of 0.47 exceeding the threshold set for large impact (g = 0.30 or greater). In first and second grade, effect sizes were within the medium range. Again, we did not find significant differences in third grade.

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

Grade	Ctrl		Tr		Dif.	ES
	Mean	SE	Mean	SE		
Kindergarten	147.76	0.53	166.56	1.20	18.81	**0.47**
First Grade	85.45	0.29	88.45	0.65	3.00	*0.11*
Second Grade	295.06	0.62	300.24	1.39	5.18	0.09

Grade	Ctrl	Tr	Dif.	ES
Third Grade	NS			

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, *Imagine Learning* students' predicted end-of-year literacy scores were within the "at or above benchmark" range for their grade, which signifies at least an 80-90 percent likelihood of achieving subsequent reading outcomes (Dynamic Measurement Group, 2021). The benchmark ranges, for kindergarten, first, and second 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 *Imagine Learning,* and 2) the impact the program had on Acadience test scores of the students that were served.

We identified positive literacy achievement outcomes for kindergartener, first and second graders who used the program as intended, as compared to matched groups of similar students who did not use the *Imagine Language & Literacy* program as part of the EISP. These gains were more pronounced in kindergarten, where the effects size far exceeded the large threshold for significance. First and second grade effect sizes fell within the medium range, despite positive differences between treatment and control students. The program had statistically non-significant findings in third grade, regardless of usage. We found that only 16% of third grade students met the vendors recommended use. It is possible that a decrease in usage contributed to the non-significant findings.

The implementation study for *Imagine Learning's* program year found that approximately 16-42% (depending on grade) used the program as intended on both aspects of the recommendation: average weekly minutes and total weeks.

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. The results of the evaluation underscore the importance of supporting students' literacy development and creating opportunities for our youngest learners. From an overall perspective, most students served by *Imagine Learning* outperformed the students who were not. Further, the students who were able to engage with the software as it was intended by *Imagine Learning* also showed greater end-of-year literacy scores relative to those participating more casually in the program.

Several recommendations surfaced from our findings:

- The percentage of students who met the recommended use criteria is somewhat lower than previous school years, across all grades and could be increased. We recommend that *Imagine Learning* identify and meet with LEAs who have usage below the recommended levels in order to cultivate ways to improve student engagement with the software.

- *Imagine Learning's* program is most impactful for kindergarteners. Continue to explore the ways in which program participation can support advanced literacy skills for students in the grades that follow.

- We also recommend that future evaluations continue to investigate the ways in which *Imagine Learning* 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.

With intentional effort behind accountability and improving consistency of use, more and more students will benefit from the *Imagine Language & Literacy* program.

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	1,135	51%	65%	8%	33%	10%	36.58
	1	1,800	44%	77%	10%	24%	6%	42.65
	2	1,628	49%	72%	10%	28%	8%	206.43
	3	1,249	46%	74%	13%	28%	8%	286.85
Matched MRU 80 Treatment Sample	K	1,121	51%	65%	7%	33%	10%	36.78
	1	1,720	44%	80%	10%	23%	5%	41.74
	2	1,600	49%	73%	10%	28%	7%	207.56
	3	1,227	45%	75%	13%	28%	7%	288.27

Note: The matched sample had a multivariate L1 score of 0.00000000000002221. 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= 0.000000000000065), White (L1= 0.000000000000031), SPED (L1 = 0.0000000000000084), Low-Income (L1= 0.000000000000026), and ELL (L1= 0.0000000000000092).

Table A2. MRU Matched Treatment Balance

	Grade	N	Female	Caucasian	SPED	Low-Income	ELL	BOY Score
Total Treatment Sample	K	662	51%	61%	8%	36%	12%	34.53
	1	1,122	43%	78%	10%	24%	6%	42.96
	2	1,030	49%	69%	10%	30%	8%	204.59
	3	642	43%	73%	15%	28%	6%	283.65
Matched MRU Treatment Sample	K	655	51%	62%	7%	36%	11%	34.67
	1	1,076	43%	80%	10%	23%	5%	42.12
	2	1,010	49%	71%	10%	30%	7%	205.98
	3	630	43%	74%	14%	27%	5%	285.40

Note: The matched sample had a multivariate L1 score of 0.00000000000000871. 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= 0.000000000000025), White (L1= 0.00000000000001), SPED (L1 = 0.000000000000023), Low-Income (L1= 0.0000000000000077), and ELL (L1= 0.0000000000000095).

APPENDIX B

Table B1. 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	144.10	0.51	152.58	0.73	8.48	*0.21*
First Grade	78.88	0.30	80.47	0.43	1.59	0.06
Second Grade	NS					
Third Grade	388.03	0.74	384.23	1.05	-3.80	-0.06

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 *Imagine Learning* software in any amount (including very low usage levels) over the course of the program year. Kindergarten had the highest predicted mean score differences between the treatment and control groups, with a medium effect size of 0.21. First grade students also exhibited predicted mean scores higher than control students, however the effect size was considered small in magnitude. Third grade, conversely, shows control students performing better than treatment students by about 4 points, however the effect size of -0.06 indicates that this difference is not considered substantive.

Table B2. ITT Matched Treatment Balance

	Grade	N	Female	Caucasian	SPED	Low-Income	ELL	BOY Score
Total Treatment Sample	K	3,082	49%	72%	8%	29%	7%	34.04
	1	4,093	48%	75%	13%	28%	6%	36.79
	2	4,296	51%	75%	14%	28%	6%	179.80
	3	3,700	48%	76%	16%	28%	7%	261.67
Matched ITT Treatment Sample	K	3,051	49%	73%	8%	29%	7%	34.14
	1	3,914	48%	78%	12%	27%	5%	35.92
	2	4,225	51%	77%	13%	28%	6%	180.60
	3	3,628	48%	77%	15%	28%	6%	262.80

Note: The matched sample had a multivariate L1 score of 0.00000000000002281. 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= 0.000000000000079), White (L1= 0.000000000000034), SPED (L1 = 0.000000000000021), Low-Income (L1= 0.000000000000024), and ELL (L1= 0.0000000000000057).

APPENDIX C

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. Imagine Learning Effect Sizes by Grade and Usage Level

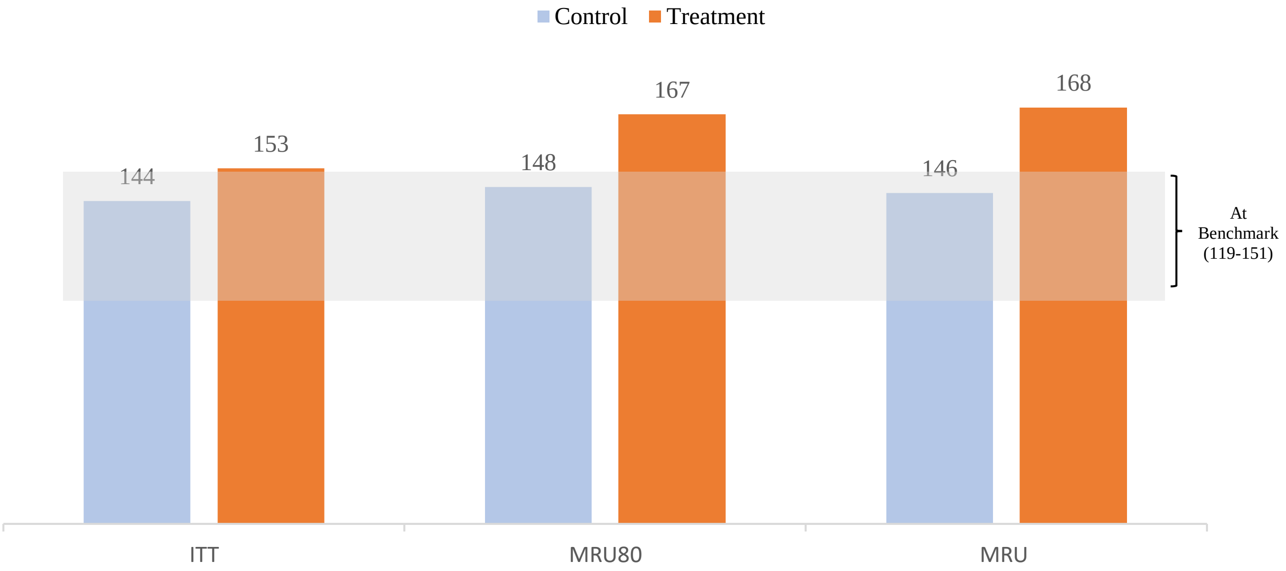
Grade	Intent to Treat	Met 80% of Rec.	Met Rec.
K	*0.21*	**0.47**	**0.55**
1	0.06	*0.11*	*0.13*
2	NS	0.09	*0.11*
3	-0.06	NS	NS

Data source: Matched K-3 ITT, MRU80, MRU samples[2]. All effect sizes displayed were statistically significant at p≤ .05. Bold = Hedges' g exceeds the 0.20 threshold. 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.

[2] Kindergarten sample size –ITT ctrl=6248.93, tr=3051; MRU80 ctrl= 5676.14, tr= 1121; MRU ctrl= 5302.225, tr= 655; First Grade- ITT ctrl= 8079.881, tr= 3914; MRU80 ctrl= 8946.515, tr=1720; MRU- ctrl= 9083.487, tr=1076; Second Grade sample size- ITT ctrl= 8653.468, tr= 4225; MRU80 ctrl= 8101.538, tr=1600; MRU ctrl= 8175.95, tr= 1010; Third Grade sample size- ITT ctrl= 7430.717, tr=3528; MRU80 ctrl=6212.867, tr=1227; MRU ctrl= 5099.85, tr=630.

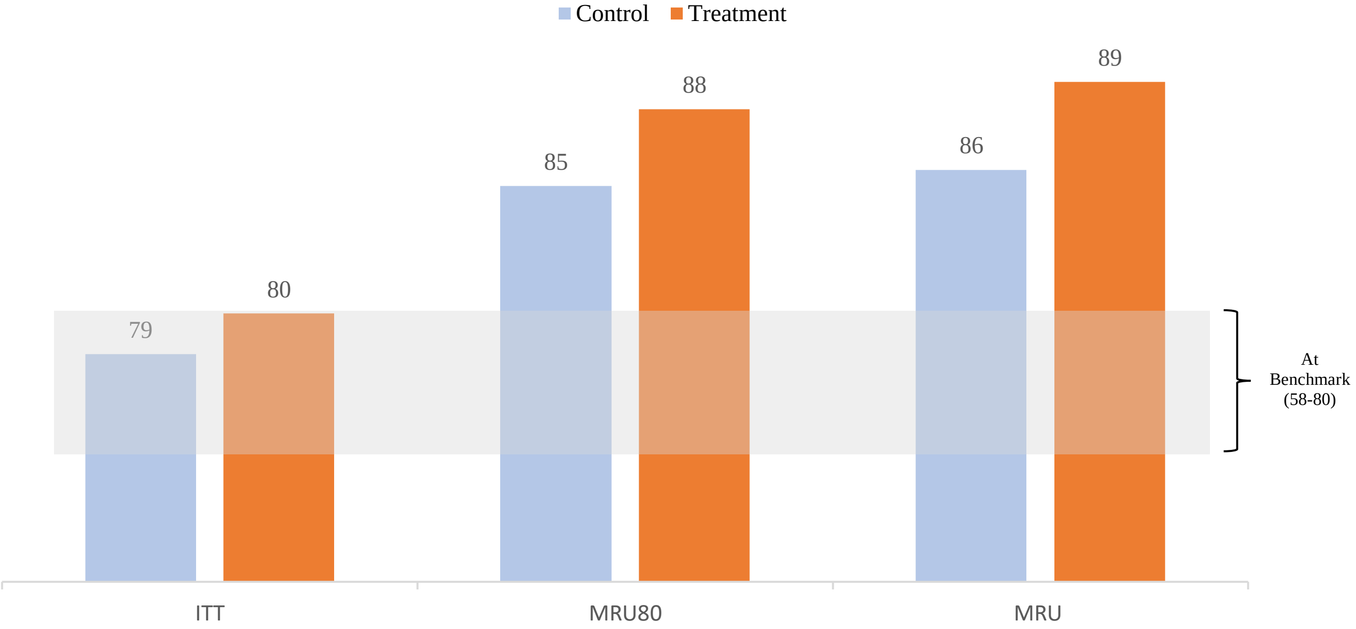
APPENDIX D

Figure D1. Kindergarten Predicted Mean Scores by Usage Level and Matched Sample



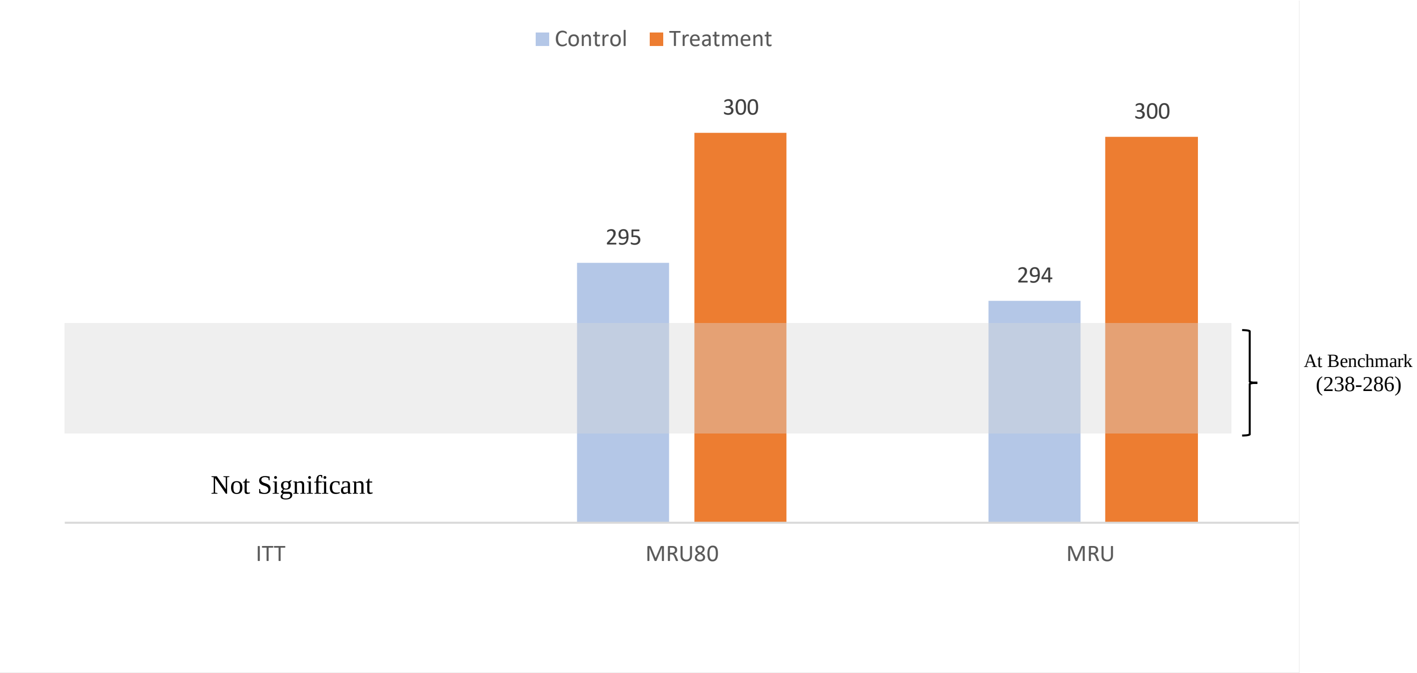
Data source: Matched kindergarten ITT, MRU80 and MRU samples. All mean comparisons displayed in the figure were statistically significant at p≤ .05.

Figure D2. 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 D3. Second Grade Predicted Mean Scores by Usage Level and Matched Sample



Data source: Matched second grade ITT, MRU80 and MRU samples. The MRU80 and MRU 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