An Evaluation of the Impact of the Lalilo Software Program: Research Report Prepared for Renaissance Learning

Ethan R. Van Norman Ph.D. & Emily R. Forcht M.Ed.

Center for Promoting Research to Practice

Lehigh University

Introduction/Background

Lalilo is a literacy software program distributed by Renaissance Learning. It is generally used with students learning to read (e.g., Kindergarten through Grade 2) as well as students in Grades 3-6 who have yet to master foundational literacy skills and need additional support. Lalilo is an online tool that can operate in school and distance learning environments. Students and teachers can access the Lalilo website through electronic devices such as computers, iPads, and tablets.

Students "travel" through a series of individualized self-paced exercises that focus on any of the following literacy skills: phonological and phonemic awareness, letter and word recognition, comprehension and fluency, vocabulary, writing, and social literacy. The literacy exercises engage students by allowing them to collect badges and unlock stories as they complete and master lessons.

Purpose and Research Questions

The purpose of this report was to evaluate the Lalilo software program. Three research questions guided the current study.

Research Question #1

- Do students in classrooms using Lalilo experience more growth in literacy skills relative to students in classrooms that do not use the program?

Research Question #2

- Do classrooms that follow best practice guidelines for the program confer greater student growth relative to classrooms that do not?

Research Question #3

- To what degree is the intensity in which Lalilo is used (e.g., number of lessons completed / number of minutes program was used) related to the magnitude of growth in literacy skills amongst students in classrooms that use the program?

Dataset Information

There were a total of 25,282 participants in the dataset. Data were collected across the 2022-2023 academic school year. Participants were in Kindergarten (n = 15,688) and Grade One (n = 9,594). There were a total of 1,020 schools represented in the dataset across 27 states. The state most represented in the dataset was Florida (39.30%) followed by California (14.70%) and New Jersey (10.90%). In this sample, 32.20% identified as White, 27.50% as Hispanic, 19.20% as Black/African American, 10.40% as Other, 5.40% as Asian, 3.60% as Multirace, and 0.76% as American Indian/Alaska Native. The sample was about evenly split between Female (50.1%) and Male (49.9%). Only 2.51% of the sample identified as English Language Learners. A little more than half (54.40%) of the participants were in the intervention group. Subsets of the sample were used for other analyses.

Assessment data was collected three times a year (e.g., fall, winter, and spring). Students in the treatment and control groups were matched based on demographic characteristics including: race/ethnicity, gender, and English learner status. Pre-test scores (fall assessment scores) was also used in the matching process. All matching was conducted by Renaissance Learning.

Research Question #1

Key Variables

- Unified Scaled Score
- Treatment Group

Research Question #2

Key Variables

- Unified Scaled Score
- Percent Fidelity Met Sessions
    - % of weeks the program was used at least 3 times per week
- Percent Fidelity Met Minutes
    - % of weeks the program was used at least 30 minutes per week
- Percent Fidelity Met Total
    - % of weeks criteria for sessions and minutes were both met (3 times + 30 mins)

Research Question #3

Key Variables

- Minutes
    - Total minutes students spent in intervention
- Days
    - Total days student active in intervention
- LO
    - Total learning objectives validated
- Lessons
    - Total lessons completed

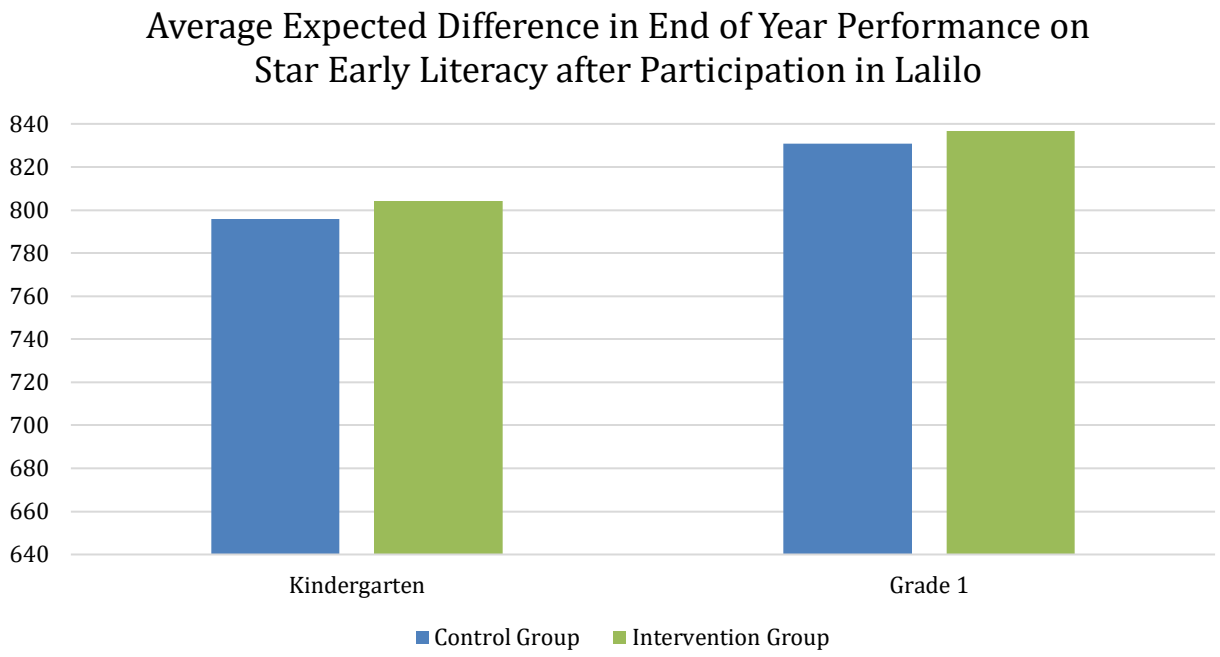
Results

Research Question #1

To address the first question a series of longitudinal multilevel models were estimated. Here unified scaled scores on Star Early Literacy were the primary outcome of interest. First, the average weekly rate of improvement on the unified scaled scores was calculated (i.e., fixed effect for growth) as well as variability around that weekly rate of improvement (i.e., random effect). A series of predictors were then added to the models to determine what student demographic characteristics and whether participation in the intervention explained student growth on Star Early Literacy. To be included in the intervention condition students had to have participated in the Lalilo program for at least 12 weeks. Baseline equivalence data amongst the intervention and control groups is included in the effect size appendix.

The primary interest for research question #1 was the interaction between the weekly rate of growth on Star Early Literacy and participation in Lalilo. For both grades, there was a statistically significant improvement in model fit when the intervention condition predictor was added to the models. **After controlling for student demographic information (e.g., ethnicity, gender, English Learner status), students in classrooms that used Lalilo experienced greater growth in early literacy skills relative to students in classrooms that did not participate in the intervention across the school year.** Kindergarten students that participated in Lalilo grew, on average, at a rate of 3.615 scaled score points per week compared to 3.372 for students that did not participate. In Grade 1, students that participated in Lalilo grew on average 2.922 scaled score points per week compared to 2.754 amongst students that did not participate.

Output of the multilevel models were used to estimate the average difference in performance on Star Early Literacy at the end of the school year between the Lalilo and control groups. For Kindergarten, students who participated in the intervention group ended the year on average **8.51 scaled score points** higher than students who did not participate (804.32 vs. 795.81). This translated to an improvement of nearly 5 percentile points on the end of year assessment. For Grade 1, students that used Lalilo ended the year on average **5.88 scaled score points** higher than students that did not (836.72 vs. 830.84). This translated to an increase of 3 percentile points.



Research Question #2

For the second research question, only students who used Lalilo were included in the analyses. The purpose was to determine whether following general program guidelines (e.g., participation in the intervention for (1) at least three times per week (fidelity met sessions) and/or (2) for at least 30 minutes per week (fidelity met minutes) led to greater weekly rates of improvement in early literacy skills. Similar to the first research question, only students that used Lalilo for 12 weeks were included in the analyses. For Kindergarten and Grade 1, **model fit did improve when the predictor fidelity met total (three times per week + met for 30 minutes per week) was modeled**. As a follow-up, fidelity met total was removed from the analysis and the predictors fidelity met sessions (percentage of weeks met three times) and fidelity met minutes (percentage of weeks met for 30 minutes) were added separately to the model. **The addition of these predictors resulted in a statistically significant improvement in model fit for Kindergarten and Grade 1**. The results suggested that students who had higher percentage of weeks met three times showed higher weekly rates of improvement in early literacy skills.

Research Question #3

For the third research question, only students who participated in the intervention were included in the analyses. The focus of this research question was whether the intensity in which Lalilo was used was related to the magnitude of growth in early literacy skills. Overall, **for both grades, adding the intensity predictors resulted in improved model fit**. However, **the only predictor that had a positive statistically significant impact on weekly rate of improvement were learning objectives validated**. For Grade 1, the number of weeks and minutes students participated in the intervention had a small negative but statistically insignificant impact on weekly rate of improvement. That is, students that used the Lalilo program for more weeks and overall minutes experienced lower growth than students that used it less frequently, on average. The number of total lessons accessed also had a negative association with growth. However, the number of **learning objectives validated** had a positive and statistically significant relationship with growth.

For students in Kindergarten who completed an average amount of total minutes, days, learning objectives validated, and lessons in Lalilo, the average weekly rate of growth was 3.55 and 3.00 for students in Kindergarten and Grade 1, respectively. For students who completed above average amounts of total learning objectives validated and lessons, the average weekly rate of growth was 3.94 for students in Kindergarten and 3.52 for students in Grade 1 (a gain of roughly +.50 scaled score points per week). For students who completed below average amounts of total learning objectives validated and lessons, the average weekly rate of growth was 3.16 for students in Kindergarten and 2.46 for students in Grade 1 (a loss of nearly -.50 scaled score points per week). Above/below average performance was calculated by adding/subtracting one standard deviation of learning objectives validated and lessons completed to the average weekly rate of growth.

Conclusion

Overall, students in schools that used Lalilo grew at a greater rate than students in schools that did not use Lalilo. There appears to be a positive but small relationship between growth and students that participated in the Lalilo program following existing recommendations (e.g., using the program for at least 30 mins per week and using the program at least three times per week). **More specific measures of fidelity or usage benchmarks may shed light on the relationship between adherence to best practice guidelines and student growth in early literacy skills**. Last, the relationship between the intensity in which Lalilo was used and growth in early literacy skills was explored. **Increasing the number of minutes or days spent using Lalilo was not critical to the improvement in weekly growth of early literacy skills. Instead increasing the number of learning objectives validated resulted in the largest improvements.** This suggests that to confer greater growth in early literacy skills, engagement with the Lalilo program needs to be intentional and close monitoring of student progress through lessons is critical.

Appendix

Table 1

Model Fit Criteria

	AIC	BIC	Deviance	AIC	BIC	Deviance
	Research Question #1 - Kindergarten			Research Question #1 – Grade 1		
Model #1	396644.7	396704.1	396630.7	263639.3	263695.9	263625.3
Model #2	396304.5	396431.9	396274.5	263377.3	263498.5	263347.3
Model #3	396247.7	396392.1	396213.7	263341.1	263478.5	263307.1
	Research Question #2 – Kindergarten			Research Question #2 – Grade 1		
Model #1	161916.5	161969.6	161902.5	197153.3	197207.8	197139.3
Model #2	161847.3	161915.6	161829.3	197066.3	197136.4	197048.3
Model #3	161841.5	161925.0	161819.5	197066.2	197151.9	197044.2
	Research Question #3 – Kindergarten			Research Question #3 – Grade 1		
Model #1	275941.4	275998.3	275927.4	283856.0	283913.1	283842.0
Model #2	274375.2	274464.6	274353.2	281865.4	281955.1	281843.4
Model #3	274336.3	274458.2	274306.3	281708.3	281830.7	281678.3

Note. AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion

Table 2

Research Question #1 – Kindergarten

	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed Effects						
Intercept	667.861**	1.076	678.091**	5.669	677.763**	5.695
Weeks	3.450**	0.016	3.449**	0.016	3.372**	0.019
Ethnicity						
Asian	-	-	3.540	5.920	3.883	5.918
Black	-	-	-15.935*	5.673	-15.270*	5.673
Hispanic	-	-	-15.015*	5.649	-14.500*	5.649
Multirace	-	-	-1.660	6.044	-1.255	6.043
Other	-	-	-3.285	6.188	-3.361	6.197
White	-	-	4.300	5.641	4.667	5.641
Male	-	-	-4.837**	0.854	-4.843**	0.854
ELL	-	-	-24.490**	3.976	-24.330**	3.991
Intervention	-	-	-	-	0.995	1.673
Weeks X Intervention	-	-	-	-	0.243**	0.033

Note. ELL = English Language Learner

** <.001, * <.05

Table 3

Research Question #1 – Grade 1

	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed Effects						
Intercept	725.662**	1.411	737.544**	6.258	734.458**	6.325
Weeks	2.862**	0.021	2.861**	0.021	2.754**	0.034
Ethnicity						
Asian	-	-	9.645	6.733	9.742	6.726
Black	-	-	-23.718**	6.408	-23.151**	6.401
Hispanic	-	-	-19.211*	6.264	-18.695*	6.257
Multirace	-	-	-4.554	7.066	-3.975	7.059
Other	-	-	0.722	6.795	0.344	6.788
White	-	-	3.788	6.218	3.564	6.211
Male	-	-	-4.542**	1.208	-4.521**	1.207
ELL	-	-	-14.575**	4.036	-13.202*	4.042
Intervention	-	-	-	-	7.509**	2.022
Weeks X Intervention	-	-	-	-	0.168**	0.043

Note. ELL = English Language Learner

** <.001, * <.05

Table 4

Research Question #2 – Kindergarten

	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed Effects						
Intercept	665.906**	2.581	663.270**	2.868	660.182**	3.167
Weeks	3.615**	0.029	3.436**	0.042	3.410**	0.053
% Fidelity Met Total	-	-	0.178*	0.060	-	-
% Fidelity Met Sessions	-	-	-	-	0.112	0.066
% Fidelity Met Minutes	-	-	-	-	0.110	0.064
Weeks X % Fidelity Met Total	-	-	0.009**	.001	-	-
Weeks X % Fidelity Met Sessions	-	-	-	-	0.008**	0.002
Weeks X % Fidelity Met Minutes	-	-	-	-	-0.000	0.001

Note.

** <.001, * <.05

Table 5

Research Question #2 – Grade 1

	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed Effects						
Intercept	733.582**	2.358	727.390**	2.706	725.031**	2.877
Weeks	2.918**	0.026	2.771**	0.039	2.756**	0.046
% Fidelity Met Total	-	-	0.299**	0.049	-	-
% Fidelity Met Sessions	-	-	-	-	0.358**	0.059
% Fidelity Met Minutes	-	-	-	-	-0.068	0.061
Weeks X % Fidelity Met Total	-	-	0.006**	.001	-	-
Weeks X % Fidelity Met Sessions	-	-	-	-	0.001	0.002
Weeks X % Fidelity Met Minutes	-	-	-	-	0.004*	0.001

Note.

** <.001, * <.05

Table 6

Research Question #3 – Kindergarten

	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed Effects						
Intercept	663.335**	2.281	658.972**	2.500	661.800**	2.510
Weeks	3.612**	0.022	3.611**	0.022	3.460**	0.036
Total Minutes	-	-	-0.018**	0.002	-0.018**	0.002
Total Days	-	-	-0.509**	0.054	-0.494**	0.060
Total Learning Objectives	-	-	-0.014	0.025	-0.058*	0.030
Total Lessons	-	-	1.140**	0.124	1.300**	0.150
Weeks X Total Mins	-	-	-	-	0.000	0.000
Weeks X Total Days	-	-	-	-	-0.001	0.001
Weeks X Learning Objectives	-	-	-	-	0.002**	0.001
Weeks X Lessons	-	-	-	-	-0.008	0.004

Note.

** <.001, * <.05

Table 7

Research Question #3 – Grade 1

	Model 1		Model 2		Model 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Fixed Effects						
Intercept	729.109**	2.068	730.109**	2.184	730.492**	2.813
Weeks	2.900**	0.021	2.895**	0.021	2.791**	0.035
Total Minutes	-	-	-0.054**	0.003	-0.052**	0.003
Total Days	-	-	-0.141*	0.052	-0.025	0.058
Total Learning Objectives	-	-	-0.236**	0.019	-0.265**	0.021
Total Lessons	-	-	1.922**	0.080	1.912**	0.091
Weeks X Total Mins	-	-	-	-	-0.000*	0.000
Weeks X Total Days	-	-	-	-	-0.009**	0.001
Weeks X Learning Objectives	-	-	-	-	0.002**	0.001
Weeks X Lessons	-	-	-	-	-0.000	0.003

Note.

** <.001, * <.05

Appendix

Effect Size Analysis

In addition to evaluating the impact of the Lalilo intervention as described in the main research summary, analyses were conducted to derive a measure of effect consistent with methods that third parties (e.g., What Works Clearing House [WWC]) use. The dataset from research question 1 was used to compute Hedge's g effect sizes consistent with recommendations outlined by the WWC for studies where intervention assignment was clustered (pp. 162 & 170-171). The general formula for Hedge's g when students are clustered within classrooms or schools is provided below:

$$g = \frac{\omega b}{SD_p}\sqrt{\gamma}$$

(p. 170)

Where *b* is the covariate adjusted mean difference on an outcome between students that did and did not participate in the Lalilo intervention. In this analysis the covariate adjusted mean difference is the coefficient from a hierarchical linear model (HLM) where students were nested within schools. The formula also includes a small-sample correction factor ω, a bias correction term Υ and the pooled individual-level standard deviation SD_p, which is defined as:

$$SD_p = \sqrt{\frac{(n_i - 1)SD_i^2 + (n_c - 1)SD_c^2}{n_i + n_c - 2}}$$

(p. 162)

Where n_i is the number of students that participated in the Lalilo intervention, n_c is the number of students that did not participate. In addition SD_i^2 is the variance of the outcome amongst the Lalilo group and SD_c^2 is the variance of the outcome amongst students that did not participate in Lalilo.

The formula for the small-sample correction factor ω is given as:

$$\omega = 1 - \frac{3}{4df - 1}$$

(p. 162)

The degrees of freedom for the correction factor is estimated as:

$$df = \frac{\left[(N-2) - 2\left(\frac{N}{M} - 1\right)\rho_{ICC}\right]^2}{(N-2)(1-\rho_{ICC})^2 + \frac{N}{M}\left(N - 2\frac{N}{M}\right)\rho_{ICC}^2 + 2\left(N - 2\frac{N}{M}\right)\rho_{ICC}(1-\rho_{ICC})}$$

(p. 171)

Where *N* is total number students, *M* is total number of clusters (i.e., schools), and ρ_{ICC} is the intraclass correlation.

Finally, the bias correction term is given as:

$$\sqrt{\gamma} = \sqrt{1 - \frac{2\left(\frac{N}{M} - 1\right)\rho_{ICC}}{N-2}}$$

(p. 170)

Effect sizes were estimated for student growth percentiles (SGP) based upon fall, winter, and spring benchmark assessments as well as end of year unified scaled scores. Separate models were estimated using Kindergarten and Grade 1 data, yielding four effect size estimates.

Baseline Equivalence

As required by the WWC (pg. 52) baseline equivalence between Lalilo and business as usual groups were assessed based upon student demographic characteristics as well as beginning of year performance on unified scaled scores. Cohen's h effect sizes were estimated for demographic characteristics when a statistically significant difference was observed between proportions. The WWC requires that estimates of between group differences be adjusted when absolute effect sizes of demographic and pre-test scores are between 0.05 and 0.25 *SDs* (effect sizes above 0.25 disqualify a study from establishing the baseline equivalence standard). Differences in baseline unified scaled scores were statistically significant for Kindergarten and Grade 1, but the Cohen's d effect size associated with those differences were equal to .01 and .03 respectively.

In Kindergarten, the proportion of Asian, White, and "Other" racial categories differed across groups and the and the proportion of Other and White students required adjustment (i.e., they were included as predictors in the HLM analyses). In grade one, differences in the proportion of American Indian, Latinx, Other, White, and Male students were statistically significant and all effect sizes associated with those differences warranted covariate adjustment.

Table A-1 *Baseline Equivalence by Demographic Categories*

Kindergarten				
Demographic	**Lalilo**	**Control**	**p-value**	**Effect Size**
American Indian	1.04	1.31	0.263	-
Asian	6.82	5.99	0.105	-0.03
Black	12.92	14.22	0.066	-
Latinx	23.78	23.07	0.424	-
Multirace	2.85	2.62	0.522	-
Other	20.04	23.56	<.001	0.09
White	32.54	29.23	<.001	-0.07
Gender	47.31	47.58	0.810	-
Non-English Learners	85.48	84.29	0.107	-
Grade 1				
Demographic	**Lalilo**	**Control**	**p-value**	**Effect Size**
American Indian	1.02	1.55	0.043	0.05
Asian	7.22	6.19	0.075	-
Black	12.06	13.28	0.107	-
Latinx	27.69	32.69	<.001	0.11
Multirace	1.73	2.29	0.080	-
Other	18.28	23.00	<.001	0.12
White	32.00	20.99	<.001	-0.25
Gender	45.57	42.39	0.004	-0.06
Non-English Learners	81.19	80.18	0.266	-

Effect Sizes

Outcomes of the HLM models were used to estimated Hedge's g effect sizes for each outcome for each grade:

Table A-2 *Hedge's g Effect Sizes*

Grade	Student Growth Percentile	End of Year Unified Scaled Score
K	0.12	0.10
1	0.11	0.13

Improvement Index

In addition to Hedge's g the WWC recommends contextualizing mean group differences in practical ways. One such metric is the improvement index (pp. E-110). Here the improvement index corresponds to the percentile points gained on end of year unified scaled score amongst average performing students in the Lalilo intervention group relative to an average performing student in the control group. SGPs are presented on a scale from 1-99 so the average SGP for the Lalilo intervention group was compared to the average SGP from the control group.

Table A-3 *Improvement Indexes based upon Average Group Scores*

	Unified Scaled Score Percentile			Student Growth Percentile		
Grade Level	Lalilo	Control	Difference	Lalilo	Control	Difference
K	66	60	+6	57	53	+4
1	49	41	+8	49	44	+5