WestEd
WestEd.org

Districtwide Pilot Study of *UFLI Foundations*

Nicholas Gage

March 2023

In partnership with the University of Florida Literacy Institute (UFLI)

Gage, N. (2023). *Districtwide pilot study of* UFLI Foundations. WestEd.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Districtwide Pilot Study of *UFLI Foundations*

The development of strong foundational skills is essential for literacy success. Foundational skills for literacy include letter knowledge, phonemic awareness, accuracy and automaticity with grapheme–phoneme correspondences, decoding and encoding skills, irregular word reading, and the ability to apply all of these skills to read and write connected text. The University of Florida Literacy Institute (UFLI) developed *UFLI Foundations: An Explicit and Systematic Phonics Program* to address these skills. The following is a preliminary report of data from a yearlong, districtwide pilot study of *UFLI Foundations* in Alachua County Public Schools, a medium-sized district in north central Florida.

UFLI Foundations

UFLI Foundations was designed to support beginning readers as they develop foundational reading skills. The program follows a carefully planned scope and sequence that addresses 128 concepts, including grapheme–phoneme correspondences, common orthographic patterns, and basic morphemes. The program begins with a series of ten "Getting Ready" lessons to prepare students with basic information about phoneme production and letter formation. These are followed by 138 detailed core lesson plans that address each concept on the scope and sequence, as well as alternative plans for review of key concepts.

Each core lesson includes eight steps: phonemic awareness, visual drill, auditory drill, blending drill, new concept introduction, word work, irregular words, and connected text. Lessons are

designed to be taught across 2 days, and 1 day per week is dedicated to review and progress monitoring. Each weekly progress monitoring assessment is intended to be used to plan small-group supplemental support for the following week.

Methods

WestEd conducted an external evaluation of UFLI using data collected by the partner school district, Alachua County Public Schools, and the UFLI. This study was designed to examine the impact of *UFLI Foundations* on student reading skills. The specific research question was as follows: Do kindergarten and 1st grade students receiving a full year of *UFLI Foundations* instruction demonstrate greater gains in reading skills than do students who did not receive *UFLI Foundations* instruction?

Study Design

This study was designed to meet established standards for high-quality research, including the What Works Clearinghouse (WWC) evidence standards with reservations. As such, this study is consistent with the Every Student Succeeds Act (ESSA) guidance for Tier 2 Moderate Evidence (Every Student Succeeds Act, 2015), which is defined as a study that (a) meets WWC Standards with reservations under version 3.0, (b) indicates a statistically significant positive effect, (c) evaluates at least 350 students, and (d) includes at least two educational sites. We used a quasi-experimental design (QED) with pre- and posttests and baseline equivalence. The comparison group, which received business-as-usual (BAU) instruction, was created using students in the same grade levels during the prior year. Table 1 presents the treatment condition by year for each grade.

Table 1. Treatment and Comparison Groups in the UFLI Study

Grade	Year	Condition
Kindergarten	2020/21	Comparison
	2021/22	Treatment
1st Grade	2020/21	Comparison
	2021/22	Treatment

We modeled each grade level separately to ensure independence of observations between grade levels. Student reading skills were measured at the beginning and end of each school year using a series of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) reading measures. The study design is presented in Table 2.

Table 2. Quasi-Experimental Design of UFLI Study

Condition	Design sequence
Treatment group	DIBELS Reading Pretest Scores Treatment DIBELS Reading Posttest Scores
Comparison group	DIBELS Reading Pretest Scores DIBELS Reading Posttest Scores

Measures

Reading skills were measured using the DIBELS® 8th Edition Assessment. The DIBELS assessment was administered in each year at the beginning of the school year in September (pretest) and again at the end of the school year in March/April (posttest). DIBELS is a widely used measure of reading skills and has established evidence of reliability and validity (University of Oregon, 2018). The DIBELS assessment includes several subtests and an aggregated composite score. The following subtests, each measuring a specific reading component skill, were included in this study: Letter Naming Fluency (LNF), Phonemic Segmentation Fluency (PSF), Nonsense Word Fluency–Correct Letter Sounds (NWF–CLS), Oral Reading Fluency (ORF), and the total aggregate score.

Data Collection

Teachers and literacy coaches collected and entered DIBELS data at the beginning, middle, and end of the school year. Alachua County Public Schools provided WestEd the data for all K–2 students in the school district for 2 consecutive years at the end of the 2021/22 school year.

Files had a unique anonymous identifier for each child and a series of demographic characteristics, including grade, gender, race/ethnicity, and special education status.

Sample

This study was conducted in 21 elementary schools in Alachua County in north central Florida. Although all students in grades K–2 participated in *UFLI Foundations*, DIBELS data were only available for K–1 students in 2020/21. Therefore, we only used data from students in kindergarten and 1st grade so that we could compare their scores on the DIBELS to a comparison group of students in K–1 that did not participate in *UFLI Foundations*. Overall, there were 4,064 students that completed the DIBELS assessment in kindergarten and 4,425 students that completed the DIBELS assessment in 1st grade. However, not every student had a pre- and post-DIBELS (fall and spring, respectively) score. Therefore, we removed all students without pre- and post-DIBELS scores.

Next, we needed to find a one-to-one student match for each treatment group (treatment and control) using the pretest scores and demographics. Unfortunately, an equivalent group could not be identified when using the full sample, likely because the two groups had similar sample sizes and the distribution of scores at pretest were different. Therefore, we focused our study on students below the pretest median score on the DIBELS Composite. The sample median was 306 for kindergarten and 333 for 1st grade. This approach resulted in a final sample of 1,084 kindergarten students and 586 1st grade students. The difference between the sample sizes was due to much more missing demographic data for 1st grade students in 2020/21. Table 3 provides a description of the sample sizes by condition and by grade.

Table 3. Analytic Sample Size by Condition

Condition	Kindergarten	1st grade
UFLI	542 students	293 students
Control	542 students	293 students
Total	1,084 students	586 students

Data Analysis

First, we used multiple imputation to impute missing covariate scores. We did not impute any DIBELS scores, only the demographic variables. Next, we used the pretest scores and

demographic variables, which included race/ethnicity, gender, free and reduced-price lunch status, and special education status, to propensity score match (PSM) students in the *UFLI Foundations* condition to students in the control condition. This approach ensured baseline equivalence on all covariates and pretest measures. PSM methods are designed to reduce bias in treatment effect estimates in experimental design studies that cannot randomly assign to conditions (Leite, 2017).

A propensity score is defined as the conditional probability of treatment assignment based on all available covariates (Rosenbaum & Rubin, 1983) and can be used for one-to-one matching of students in treatment schools to students in BAU schools (control condition). The value of PSM is that a covariate equivalent comparison group can be matched to a treatment group to meet established research standards, including the WWC standards (2020). Furthermore, PSM treatment estimates have been found to be as accurate as those from randomized controlled trial studies (Fortson et al., 2012).

We estimated propensity scores using logistic regression following procedures outlined by Leite (2017). Specifically, treatment was coded as a dichotomous indicator and used as the dependent variable in a logistic regression model with all covariates, including all pretest measures. The propensity score is then the predicted probability of a student being assigned to the treatment or control group based on the model covariates. Thus, this approach reduces selection bias by establishing equivalence on the included model covariates.

Next, we used each student's estimated propensity score to match them using the one-to-one optimal matching method (Rosenbaum, 1989), which minimizes global propensity score distance (i.e., predicted probability of being in the treatment or comparison group) to a student in one of the control schools by finding the smallest average absolute distance across all the matched students. We used the one-to-one optimal matching algorithm with MatchIt (Ho et al., 2011) and optmatch (Hansen et al., 2018) packages in *R* (R Core Team, n.d.). To confirm covariate equivalence, we calculated standardized mean difference effect sizes (g) for pretest scores, by treatment condition, where equivalence was defined as g < .25 standard deviation units (WWC, 2020).

Next, we assessed impacts of *UFLI Foundations* on student outcomes using the following multilevel model:

$$y_{ijk} = \beta_0 + \beta_1 condition_j + \sum_{s}^{8} \pi_s X_{ijk} + \xi_j + \varepsilon_{ij} + e_{ijk}$$

y_{ijk} if the outcome for student *i* in teacher *j* in school *k*; $condition_j$ is set to 1 for *UFLI* and 0 for control; are student-level covariates, including all with equivalence values (*g*) greater than .05

standard deviation units; ξ_j, ε_{ij} and e_{ijk} are school, teacher, and student random effect, respectively. Treatment impacts will be assessed using the estimate of β_1.

Findings

This section presents findings from WestEd's external evaluation of UFLI.

Students receiving *UFLI Foundations* performed much higher than control students

First, we examined the pre- and posttest scores by treatment condition for the baseline equivalence propensity score–matched students. Table 4 presents the descriptive statistics. Based on pretest scores, the students were equivalent at baseline, with $g < .05$ standard deviations. Put differently, all of the students were essentially the same with regard to their reading ability at the beginning of the school year. Based on posttest scores, the students receiving *UFLI Foundations* performed much higher at posttest than did students in the control condition. The effect size, controlling for pretest, was $g = 1.20$ for kindergarten students and $g = 1.42$ for 1st grade when using the pretest standard deviation in the effect size calculation.

These descriptive findings suggest that UFLI had a meaningful impact on students' ability to read. These impacts were true for students in both kindergarten and 1st grade.

Table 4. Descriptive Statistics by Condition and Pre- and Posttest

	Pretest Mean score	**Pretest** Standard deviation	**Posttest** Mean score	**Posttest** Standard deviation
Kindergarten control	274	18.9	397	24.4
Kindergarten UFLI	275	19.5	421	32.1
1st grade control	320	8.6	427	22.0
1st grade UFLI	321	8.2	440	29.9

Next, we estimated the multilevel models (see Tables 5 and 6 for results). These models provide a more refined estimate of *UFLI Foundations*' impact by controlling for student characteristics.

First, we estimated empty, or intercept-only models to estimate the intraclass correlation coefficients (ICC). As can be seen in Table 6, the largest ICC was between kindergarten teachers, suggesting that about 13 percent of the variance was between the teachers. All other ICC values were less than .10, suggesting little difference between levels. Next, we examined the primary treatment effect, defined as the covariate adjusted impact of *UFLI Foundations* on students' reading ability. As can be seen in Table 5, the coefficient for both kindergarten and 1st grade was statistically significant. When adjusted for student characteristics, the effect size increased to $g = 1.44$ for kindergarten and $g = 2.04$ for 1st grade. Cohen (1988) defined large effect sizes as $g > 0.80$; therefore, the findings suggest that *UFLI Foundations* had a large effect on reading performance. With regard to covariates, we should note that there were only Black and White students in the reduced 1st grade sample. We found that in kindergarten, Black students scored lower than White students, while Asian students scored higher. Students with disabilities also performed significantly lower than their peers without disabilities.

The findings suggest that when controlling for these student differences, students receiving UFLI Foundations, including students with disabilities, had better reading performance than did students not receiving UFLI Foundations.

Table 5. Three-Level Multilevel Model Results

	Kindergarten estimate	Kindergarten std. error	Kindergarten P	1st grade estimate	1st grade std. error	1st grade P
(Intercept)	216.05	12.00	0.000	-155.81	35.50	0.000
UFLI	22.23	1.56	0.000	12.84	1.76	0.000
Black	-5.33	1.84	0.004	-1.37	0.90	0.126
Hispanic	-0.58	2.40	0.808	No data	No data	No data
Two or more Races	-1.33	2.59	0.608	No data	No data	No data
Asian	11.57	4.22	0.006	No data	No data	No data
Free and reduced-price lunch	-0.25	3.44	0.942	0.36	3.85	0.925
Female	2.42	1.45	0.095	-1.89	1.75	0.279
Special education	-12.26	2.20	0.000	-2.32	2.33	0.320
Pretest	0.67	0.04	0.000	1.83	0.11	0.000

Table 6. Three-Level Multilevel Model Results: Random Effects

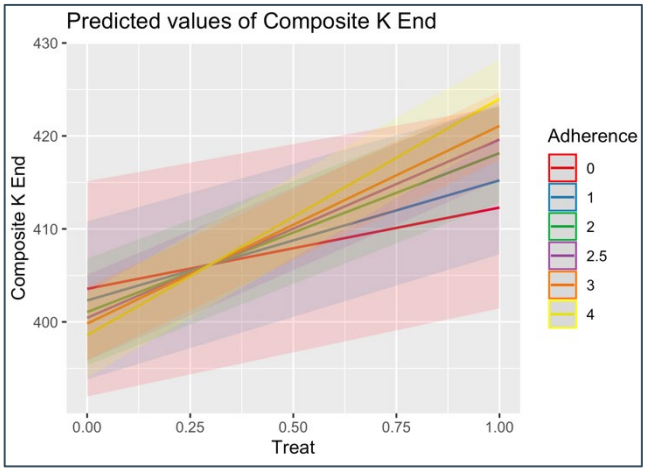
	Kindergarten estimate	1st grade estimate
Teacher	131.38	25.23
School	33.24	10.04
Residual	818.06	696.68
Teacher ICC*	0.13	0.03
School ICC*	0.03	0.01

*ICC is intraclass correlation coefficient.

Students in classrooms implementing *UFLI Foundations* with fidelity performed even better in reading than did those in classrooms not implementing with fidelity.

We also examined teachers' implementation fidelity via ratings by literacy coaches. Each teacher was rated from 0 to 4 on both adherence and dosage, with 4 indicating high fidelity (see Appendix for rubrics). Analyses indicated that implementation fidelity predicted student reading growth (see Figure 1 for a graph for K values for adherence). That is, teachers who implemented *UFLI Foundations* as intended produced greater reading growth in their students. The graph in Figure 1 includes the dependent variable on the y-axis, the kindergarten end-of-year composite score, a treatment indicator on the x-axis (0 is no treatment and 1 is treatment). The lines are color coded by the level of adherence (0–4) with which a teacher implemented UFLI. No adherence (0) is red, an adherence of 1 is blue, 2 is green, 2.5 is purple, 3 is orange, and 4 is yellow. The graph also includes the 95 percent confidence interval for each line using shading of the same color as the line. The steepest line is yellow, suggesting the greatest difference between students' end-of-year composite scores is those in classrooms where the teacher scored a 4 on the adherence measure.

Figure 1. Treatment Effect Moderated by Fidelity of Implementation



An interaction effect graph shows the predicted values of composite K end. The Y axis is labeled "Composite K End" and the X axis is labeled "Treat." Adherence is color coded. See the paragraph above for a thorough description of this line graph.

Conclusion

Data from this districtwide pilot study provide evidence that *UFLI Foundations* is effective for improving foundational literacy skills. Students who received instruction using *UFLI Foundations* in the 2021/22 school year significantly outperformed students who did not receive *UFLI*

Foundations instruction in 2020/21. Teachers with high implementation fidelity produced substantially greater student growth compared with teachers with lower levels of fidelity.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Fortson, K., Verbitsky-Savitz, N., Kopa, E., & Gleason, P. (2012). *Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates* (NCEE technical methods report 2012-4019). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hansen, B. B., Fredrickson, M., Buckner, J., Errickson, J., & Solenberger, P. (2018). *Optmatch: Functions for optimal matching* (R Package Version 0.9-8). https://www.cran.r-project.org/web/packages/optmatch/

Ho, D., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric pre- processing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://www.doi.org/10.18637/jss.v042.i08

Leite, W. (2017). *Practical propensity score methods using R*. Sage Publications.

R Core Team. (n.d.). *The R Project for statistical computing*. https://www.r-project.org/

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

University of Oregon. (2018). 8th edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®). https://dibels.uoregon.edu/assessment/index/material/

What Works Clearinghouse. (2020). *What Works Clearinghouse: Standards handbook version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

Appendix: Rubrics for Rating Implementation Fidelity

Reading coaches measured teachers' fidelity of implementation of UFLI using an adherence rubric and a dosage rubric, both shown below. These rubrics were used to determine how well each teacher implemented UFLI.

Adherence Rubric

Adherence	
4	• Teacher implements **all** lesson steps (8 out of 8) • Teacher implements **all** lesson substeps (e.g., reading and spelling with new concept) • Teacher uses the correct materials during each step • Teacher **fully adheres** to the *UFLI Foundations* scope and sequence
3	• Teacher implements **most** lesson steps (6 out of 8) • Teacher implements **most** lesson substeps (e.g., reading and spelling with new concept) • Teacher uses the correct materials during each step • Teacher **fully adheres** to the *UFLI Foundations* scope and sequence
2	• Teacher implements **some** lesson steps (at least 4 out of 8) • Teacher implements **most** lesson substeps (e.g., reading and spelling with new concept) • Teacher uses the correct materials during most steps
1	• Teacher implements **less than 4** lesson steps • Teacher implements **some** lesson substeps (e.g., reading and spelling with new concept) • Teacher does not use correct materials during most steps
0	• Teacher is not implementing *UFLI Foundations*

Dosage Rubric

Dosage	
4	• During a typical week, teacher implements *UFLI Foundations* **at least 4 times a week**
3	• During a typical week, teacher implements *UFLI Foundations* **at least 3 times a week**
2	• During a typical week, teacher implements *UFLI Foundations* **at least 2 times a week**
1	• During a typical week, teacher implements *UFLI Foundations* **at least 1 time a week**
0	• Teacher is not implementing *UFLI Foundations* at all


WestEd
WestEd.org