AIR®
AMERICAN INSTITUTES FOR RESEARCH®

JANUARY 2018

Annual Evaluation Report for the Pennsylvania Dyslexia Screening and Early Literacy Intervention Pilot Program

Pilot Year 2, 2016–17 School Year

Laura Kuchle
Seth Brown
Nicholas Coukoulis

Annual Evaluation Report for the Pennsylvania Dyslexia Screening and Early Literacy Intervention Pilot Program

Pilot Year 2, 2016–17 School Year

JANUARY 2018

Laura Kuchle
Seth Brown
Nicholas Coukoulis

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

Contents

Page

Abbreviations

AIR	American Institutes for Research
CLS	correct letter sounds
CTOPP-2	Comprehensive Test of Phonological Processing, Second Edition
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
ES	effect size
FSF	DIBELS First Sound Fluency
LETRS	Language Essentials for Teachers of Reading and Spelling
LiPS	Lindamood Phoneme Sequencing Program for Reading, Spelling, and Speech
LNF	DIBELS Letter Naming Fluency
M	mean
MD	Mahalanobis distances
MSL	multisensory structured language
NWF	DIBELS Nonsense Word Fluency
ORF	Oral Reading Fluency
OG	Orton-Gillingham
PaTTAN	Pennsylvania Training and Technical Assistance Network
PDE	Pennsylvania Department of Education
PPVT-4	Peabody Picture Vocabulary Test, Fourth Edition
PSF	DIBELS Phoneme Segmentation Fluency
RD	regression discontinuity
RQ	research question
SD	standard deviation
WC	words correct
WRS	Wilson Reading System
WWR	whole words read

Executive Summary

American Institutes for Research (AIR) is conducting the independent evaluation of the implementation and effectiveness of the Pennsylvania Dyslexia Screening and Early Literacy Intervention Pilot Program (Pilot). The 3-year Pilot began in 2015–16 (Year 1) with the kindergarten class of 2015–16 (Cohort 1). In 2016–17 (Year 2), the Pilot was implemented with Cohort 1 students, now in first grade, and a second cohort of kindergarteners (Cohort 2). The Pilot provides two levels of support: (1) a classroom program, which supplements core instruction for all students, with an increased focus on phonemic awareness and multisensory structured language (MSL), and (2) an MSL intervention to provide extra instruction for students identified as needing more support based on early literacy screening in the winter of kindergarten. Both levels of support are meant to affect special education referrals and students' literacy skills, measured by the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Next benchmark assessments (only DIBELS data are available at this time). This report presents key findings from Year 2.

Pilot Program

The Pilot treatment condition includes two levels of support through two distinct treatment components. The classroom program, provided to all students, strengthens core instruction by providing classroom teachers with professional development aligned with the recommendations of the National Early Literacy Panel (2008), the National Reading Panel (2000), and other evidence-based approaches (Adams, 1990; Foorman et al., 2016; Kosanovich & Foorman, 2016; National Research Council, 1998; Shanahan et al., 2010).The participating schools continued the core literacy program already in place (or being adopted) at the beginning of the Pilot. The second level of support is the MSL intervention, which is provided only to students identified as needing more support. Districts were free to implement the MSL intervention approach of their choice, including Orton Gillingham (OG), Sonday, Wilson, and others.

Evaluation Methods and Sample

Different evaluation designs were needed to evaluate each of the Pilot's two treatment components. The effectiveness of the classroom program was evaluated using a school-level matched design, in which the performance of students in the 21 Pilot schools was compared with the performance of students in 21 matched comparison (Comparison) schools identified through Mahalanobis distance matching.[1] Because some Pilot students received the MSL intervention in addition to the classroom program, any benefit of the intervention will influence classroom program analyses. All Pilot and Comparison schools implemented universal screening using DIBELS in the fall, winter, and spring of both Year 1 and Year 2. The DIBELS subtests administered in the spring served as outcome variables. For kindergarten, these were Letter Naming Fluency (LNF), Nonsense Word Fluency (NWF), and Phoneme Segmentation Fluency (PSF). For first grade, these were NWF and Oral Reading Fluency (ORF). NWF has two scoring methods—correct letter sounds (NWF-CLS) and whole words read (NWF-WWR); only the

[1] Comparison schools participated in another funded literacy initiative that was in its fifth year during Year 2 of this Pilot. This other initiative also used universal screening to inform core instruction and identify students to receive supplemental intervention.

correct letter sounds method is recommended in kindergarten, while both are recommended in first grade. ORF yields two scores—words correct in one minute (ORF-WC) and the percentage of attempted words that were read correctly (ORF-Accuracy). Kindergarten students in Pilot schools were assigned to the MSL intervention based on their winter LNF score;[2] students qualified for intervention with a winter LNF score of 39 or below (the Pilot sample's 35th percentile for Cohort 1).[3] Cohorts 1 and 2 had similar levels of MSL intervention participation in kindergarten: 484 of 603 (80.3%) Cohort 1 students and 446 of 567 (78.6%) Cohort 2 students who qualified for the intervention participated through the end of kindergarten.[4] Fewer Cohort 1 students participated in the intervention through the end of first grade (408 of 603, or 67.7%), because of students moving away from participating schools between their kindergarten and first grade years. However, these 408 students made up 83.6% of the qualified Cohort 1 students who remained in the schools and continued on to the first grade. The effectiveness of the MSL intervention was assessed using a regression discontinuity (RD) design, in which Pilot students eligible for the intervention were compared with similarly performing students in the same schools who were not eligible for the intervention (i.e., students who scored 40 or above on their kindergarten winter LNF). All students, regardless of whether or not they were eligible for the intervention, continued to be assessed three times per year using DIBELS.

Implementation

In spring 2015, the Pennsylvania Department of Education (PDE) provided 4 days of training for kindergarten teachers and designated interventionists for each Pilot district on Language Essentials for Teachers of Reading and Spelling (LETRS) Modules 1–3. In Year 1, PDE focused efforts on intervention training and implementation; interventionists received additional professional development during summer 2015 and the 2015–16 school year. PDE provided classroom program materials and conducted classroom instruction observations, but the data were not analyzed because of concerns about their reliability. However, PDE learned that implementation of the classroom lessons was limited in Year 1. For Year 2, PDE developed and provided 4 days of training for first grade teachers based on the recommendations of the National Early Literacy Panel (2008), the National Reading Panel (2000), and other evidence-based approaches (Adams, 1990; Foorman et al., 2016; Kosanovich & Foorman, 2016; National Research Council, 1998; Shanahan et al., 2010). Kindergarten teachers had access to those sessions as well as a refresher session specifically for kindergarten teachers.

Across both years and cohorts, the majority of intervention students received OG as their MSL intervention (in Year 2, 53% for Cohort 1 and 64% for Cohort 2). Other students received similar MSL intervention programs—most commonly Sonday. OG training was provided by the Compass Reading Center (Compass), which also assessed trainees' knowledge and implementation of OG components. All of the OG interventionists trained by Compass showed growth in knowledge and high adherence to the OG components. For the 48 OG interventionists trained for Year 1, the test scores improved from pretest (mean [*M*] = 36.7%, standard deviation

[2] LNF is the kindergarten winter DIBELS subtest most predictive of future reading fluency (see Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009).
[3] The same cut score was applied to Cohort 2; again, 35% of Pilot students qualified for the MSL intervention.
[4] The most common reason for nonparticipation was parents opting out. Other reasons include students moving or, rarely, being deemed unable to participate (e.g., individualized education program team decision that the MSL intervention was inappropriate for a nonverbal student).

[*SD*] = 11.5) to posttest (*M* = 91.3%, *SD* = 9.6). The 35 additional OG interventionists trained for Year 2 also improved from pretest (*M* = 29.8%, *SD* = 13.9) to posttest (*M* = 94.7%, *SD* = 5.3). All Compass-trained interventionists maintained a mean fidelity score above 85. For Year 1 trainees, the average fidelity observation score was 96 out of 100, with a range of 88 to 99. For Year 2 trainees, the average score was 96, with a range of 90 to 99.

Intervention logs (time records) showed that all schools implemented the MSL intervention. However, the majority of students participating in the intervention did not receive the targeted 30 hours by the end of kindergarten and 100 hours by the end of first grade. For Cohort 1, the average was approximately 76 hours by the end of first grade, with only 5% of students receiving the recommended dosage (5% met the kindergarten target in Year 1 for Cohort 1). Cohort 2 intervention students had 26 hours of intervention, on average, with approximately 41% meeting the kindergarten target (a vast improvement compared to Year 1). Means varied considerably by school. Failure to meet target hours is important because larger intervention dosages were associated with higher DIBELS scores (see *Exploratory Analyses* in this Executive Summary and, for more detail, in Chapter 4).

Outcomes for Classroom Program

The main analyses for the classroom program yielded significant findings (*p* < .05) on some measures for both cohorts. Exhibit S1, for Cohort 1, shows significant effects for both spring first grade NWF scoring methods (correct letter sounds, NWF-CLS; and whole words read, NWF-WWR): NWF-CLS had an estimated difference of 6.8 points (effect size [*ES*] = 0.18) and NWF-WWR had an estimated difference of 2.5 (*ES* = 0.17). As seen in Exhibit S2, for Cohort 2, significant effects were seen for two spring kindergarten measures: LNF had an estimated difference of 3.9 (*ES* = 0.23), and NWF-CLS had an estimated difference of 4.9 (*ES* = 0.21).

Exhibit S1. End-of-Year Pilot and Matched Comparison Analyses, Cohort 1

Outcome	Pilot Group Mean	Comparison Group Mean	Estimated Difference	Standard Error	Effect Size	*p*-Value
Kindergarten (2015–16)						
LNF	57.6	54.9	2.7	1.4	0.16	.063
NWF-CLS	44.4	45.8	-1.4	2.1	-0.05	.512
PSF	54.8	55.2	-0.4	2.4	-0.03	.865
First Grade (2016–17)						
NWF-CLS	86.4	79.6	6.8*	2.8	0.18	.015
NWF-WWR	26.4	23.9	2.5*	1.1	0.17	.029
ORF-WC	64.3	67.3	-3.1	2.2	-0.09	.158
ORF-Accuracy	92.6	90.2	2.4	2.1	0.16	.262

Note. Kindergarten sample size = 2,735 students (1,591 Pilot and 1,144 Comparison); first grade sample size = 2,471–2,472 students (1,433 Pilot and 1,038–1,039 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

Exhibit S2. End-of-Year Pilot and Matched Comparison Analyses, Cohort 2

Outcome	Pilot Group Mean	Comparison Group Mean	Estimated Difference	Standard Error	Effect Size	*p*-Value
Kindergarten (2016–17)						
LNF	58.2	54.3	3.9*	1.5	0.23	.008
NWF-CLS	47.3	42.3	4.9*	1.5	0.21	.001
PSF	54.4	55.6	-1.3	1.8	-0.08	.487

Note. Sample size = 2,819–2,820 students (1,519 Pilot and 1,300–1,301 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

Outcomes for MSL Intervention

RD analyses for the MSL intervention focused on the Pilot students within 5 points below the kindergarten winter LNF cut score (35–39) and 5 points above the cut score (40–44).[5] Controlling for the kindergarten winter LNF score, the analysis showed generally positive but nonsignificant effects of the MSL intervention on spring DIBELS scores (spring of kindergarten for Cohort 2, and spring of kindergarten and first grade for Cohort 1). Exhibits S3 (Cohort 1) and S4 (Cohort 2) summarize these results. Additional analyses, however, showed that in kindergarten, and to a lesser extent in first grade, students who received more intervention time tended to have higher spring DIBELS scores, controlling for kindergarten winter LNF scores.

Exhibit S3. Cohort 1 Regression Discontinuity Analyses for Restricted Sample (LNF 35–44)

Outcome Variable	Estimated Effect	Standard Error	Effect Size	*p*-Value
Kindergarten (2015–16)				
LNF	1.6	2.0	0.15	.431
NWF-CLS	3.3	2.8	0.23	.229
PSF	1.5	2.2	0.12	.498
First Grade (2016–17)				
NWF-CLS	8.5	6.3	0.3	.175
NWF-WWR	2.9	2.5	0.2	.241
ORF-WC	-1.1	4.9	-0.0	.817
ORF-Accuracy	0.9	1.9	0.1	.609

Note. Kindergarten sample size = 431 students (186 intervention and 245 nonintervention); first grade sample size = 391 students (165 intervention and 226 nonintervention). The analyses were based on a two-level regression (students within schools), controlling for kindergarten winter LNF. The *p*-values for the estimated impact are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

[5] This restricted sample yields a more conservative estimate of the effect of the MSL intervention compared with the higher statistical power of the full sample, but was preferred because the two groups being compared were more similar at baseline (i.e., closer kindergarten winter LNF scores).

Exhibit S4. Cohort 2 Regression Discontinuity Analyses for Restricted Sample (LNF 35–44)

Outcome Variable	Estimated Effect	Standard Error	Effect Size	*p*-Value
Kindergarten (2016–17)				
LNF	-3.0	2.2	-0.29	.170
NWF-CLS	0.9	3.2	0.06	.781
PSF	2.5	2.4	0.22	.297

Note. Sample size = 316 students (133 intervention and 183 nonintervention). The analyses were based on a two-level regression (students within schools), controlling for kindergarten winter LNF. The *p*-values for the estimated effect are based on *t* tests. Two-tailed statistical significance at the $p < .05$ level is indicated by an asterisk (*).
Source: District DIBELS data.

Exploratory Analyses

Classroom program analyses that looked separately at the effect of the classroom program on intervention and nonintervention students revealed that for Cohort 1, Pilot students assigned to the intervention outperformed similar Comparison students on kindergarten spring LNF and NWF-CLS and on first grade spring NWF-CLS and NWF-WWR. These results for Cohort 1, along with the null results in the main kindergarten classroom component analyses and MSL intervention analyses, suggest that the MSL intervention had a positive effect on the preliteracy skills of students with the greatest need for support (students who not only qualified for the MSL intervention, but who had scores more than 5 points below the cut score). In contrast, for Cohort 2, both intervention and nonintervention Pilot students outperformed similar Comparison students; taken together with the null Cohort 2 MSL analyses, these results suggest that the classroom component in Cohort 2 was primarily responsible for the effects on students' preliteracy skills. This may suggest that classroom program implementation improved in Year 2. Additional analyses suggest that the growth in DIBELS scores were positively associated with time in MSL intervention (students who received more intervention tended to show greater growth from winter to spring of kindergarten), so more intervention time in Year 3 may yield stronger findings.

Conclusions

Classroom program analyses suggest that both Pilot cohorts outperformed the Comparison sample on some spring 2017 (Year 2) measures. This may be because of improved implementation in Year 2, and is particularly encouraging given the Comparison sample's participation in another literacy initiative, which may result in an underestimation of Pilot program effects compared with typical schools (which may not use universal screening to inform core instruction and identify students to receive supplemental intervention). Although the main MSL intervention analyses yielded no positive effects, exploratory analyses suggest that the intervention may have contributed to improved performance for Cohort 1 Pilot intervention students compared to similar Comparison students.

Chapter 1. Introduction and Overview

Pilot Overview

Act 69 of 2014 amended the Pennsylvania Public School Code of 1949 to establish the Dyslexia Screening and Early Literacy Intervention Pilot Program (Pilot; Pennsylvania Department of Education, 2014b). The Pilot's purpose is to establish methods, as early as possible, to: (a) identify students at risk for reading difficulties, including dyslexia, and (b) provide appropriate supports to these students to improve future reading outcomes and reduce the need for special education in later grades. The long-term goal is for the Pilot to serve as a model for scaling up the multisensory, phonics-based approach to early reading instruction and intervention in kindergarten through second grade.

The Pilot, administered by the Pennsylvania Department of Education (PDE), supports the implementation of evidence-based early literacy screening, instruction, and intervention for Grades K–2 in 21 elementary schools across eight districts during 3 school years. Districts were chosen to represent a variety of locations (East, Central, and West regions) and sizes (three districts have student populations below 3,000, and five districts have student populations between 3,000 and 15,000; Pennsylvania Department of Education, 2014a). The Pilot includes three cohorts, with a new cohort starting kindergarten each school year. The first cohort is made up of students who began kindergarten in the 2015–16 school year; these students (Cohort 1) will be followed for the full 3 years of the project (i.e., through second grade). Cohort 2 students began kindergarten in 2016–17 and will be followed for 2 years (through first grade). Cohort 3 began kindergarten in 2017–18, the final year of the project.

All Pilot schools use Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Next as a universal screener in the fall, winter, and spring of each school year. These data inform instruction and identify students in need of supplemental intervention. The Pilot aims to enhance core instruction by providing classroom teachers with professional development for teaching phonics and phonological awareness, aligned with the recommendations of the National Early Literacy Panel (2008), the National Reading Panel (2000), and other evidence-based approaches (Adams, 1990; Foorman et al., 2016; Kosanovich & Foorman, 2016; National Research Council, 1998; Shanahan et al., 2010), while schools continue using the core literacy program of their choice. Strong core instruction is intended to better support all students, potentially reducing the likelihood of "false positives" (when students who do not truly need intervention qualify for intervention). Enhanced core instruction also serves as an added support to at-risk students receiving supplemental intervention, with the goal of reducing rates of special education referrals, and boosting both short- and long-term reading outcomes. (See Chapter 2 for more information.) Students at risk are offered the multisensory structured language (MSL) intervention, usually Orton-Gillingham (OG), implemented by trained interventionists. PDE recommended that intervention groups should be composed of no more than three students and set targets of 30 intervention hours by the end of kindergarten and a total of 100 hours (including kindergarten hours) by the end of first grade.

Evaluation Overview

PDE contracted American Institutes for Research (AIR) to conduct an independent evaluation of the Pilot using data provided by PDE. The evaluation seeks to answer the following two sets of research questions (RQs), one set related to fidelity of implementation and the other set related to the effectiveness of the approach.

Implementation

1. Do teachers and interventionists receive the training as intended?
2. Do classroom teachers and interventionists implement the program as intended (e.g., do teachers use the classroom program with fidelity; do interventionists provide the MSL intervention for students with fidelity)?

Effectiveness

3. Does the classroom program improve student outcomes (e.g., increased reading assessment scores, reduced number of students identified as being at-risk in reading, reduced number of students referred to special education services)?
4. Does the MSL intervention improve student outcomes (e.g., increased reading assessment scores, reduced number of students referred to special education services)?

Implementation questions were answered using descriptive analyses. Effectiveness questions were assessed using quasi-experimental designs. For more information on the evaluation design, see Chapter 2.

Report Overview

Chapter 2 provides an overview of the study design and timeline. It covers topics such as the overall design, evaluation timelines, data sources, sample descriptions, and analysis methods. Chapter 3 examines implementation of training and instructional programs (RQs 1 and 2) in Years 1 and 2 (2015–16 and 2016–17, respectively). Descriptive summaries of training and program implementation provide a context for interpreting the effectiveness findings. Chapter 4 presents the results regarding the overall effectiveness of the classroom program and intervention on student reading outcomes (RQs 3 and 4) in Year 2. Chapter 5 summarizes and discusses the preliminary findings, identifies the evaluation’s limitations, and previews anticipated future reports.

Chapter 2. Evaluation Design and Methodology

Overall Design

Different evaluation designs were needed to evaluate each of the Pilot's two treatment components: the classroom program implemented with all students, and the MSL intervention offered to students identified as needing additional support. The effectiveness of the classroom program (RQ 3) was assessed using a school-level matched design, in which the performance of Pilot students was compared with the performance of students in matched comparison (Comparison) schools. Comparison schools were identified using a Mahalanobis distance (MD) approach (Rubin, 1980). Because some Pilot students received the intervention in addition to the classroom program, any effect of the intervention will also influence the classroom program analyses; that is, classroom program effects reflect the core instruction provided to all students *and* the MSL intervention provided to students who qualified for additional support.[6] The effectiveness of the MSL intervention (RQ 4) was assessed using a regression discontinuity (RD) design, in which Pilot students qualifying for the intervention were compared with similar students in the same schools who did not qualify.[7] RD design is a rigorous quasi-experimental method for estimating the effect of an intervention when program participants are selected using an arbitrary cut point on a continuous measure (Jacob, Zhu, Somers, & Bloom, 2012; see Appendix A for more information on these designs). For both designs, DIBELS Next scores served as outcome variables. In future years, special education referral and eligibility data will be considered. The following sections describe the processes used to identify the Comparison schools (addressing RQ 3) and the Pilot students qualifying for the intervention (addressing RQ 4).

Identifying a Comparison Sample for Evaluating the Effectiveness of the Classroom Program (RQ 3)

The validity of the matched-school quasi-experimental design depended on the quality of the matching process, which enabled the selection of a Comparison sample that was as similar as possible to the Pilot sample on observed characteristics. The matched Comparison schools were identified prior to the 2015–16 school year using historical data provided by PDE. All potential matched schools were located in non-Pilot districts because all Pilot district elementary schools were participating in the Pilot. Schools considered as matched sites were further limited to elementary schools participating in a different literacy initiative that began 3 years before the Pilot and provided substantial funding to participating districts as part of implementation. That initiative included use of DIBELS Next as a universal screener—a prerequisite for being considered as a Comparison school. This allowed for a comparison of preliteracy and reading outcomes between Pilot and Comparison schools. However, this also meant that Comparison

[6] The Comparison schools used a similar approach in which students needing additional supports received both a core program and supplemental intervention. Core and supplemental programs in Comparison schools reflected their "business as usual" practices, without the benefit of the MSL training received by teachers and interventionists in the Pilot schools.

[7] The main analyses examining the MSL intervention were based on an intent-to-treat (i.e., intent to prove intervention) approach. However, students who qualified for the intervention participated only if their parents consented.

schools were not business-as-usual schools, which may result in an underestimation of Pilot program effects compared with typical schools (which may not use universal screening to inform core instruction and identify students to receive supplemental intervention). Furthermore, Comparison schools might be expected to be further along in implementation because the $33 million program was in its fourth year when the Pilot began (and was in its fifth year during Pilot Year 2). Charter schools were eliminated from the potential pool of matched Comparison schools because the Pilot sample did not include any charter schools. MDs between each Pilot school and potential matched schools were calculated using the following variables:

- DIBELS composite score, kindergarten and Grade 1, beginning of year[8]
- Grade 3 Pennsylvania System of School Assessment (PSSA) reading score
- Percentage eligible for free or reduced-price lunch
- Percentage African American
- Percentage Hispanic
- Total enrollment

AIR examined the top five matches (i.e., the five potential Comparison schools with the lowest MDs) for each Pilot school to identify a unique Comparison school for each Pilot school based on:

- Standardized mean differences for all MD[9]
- Title I status
- Urbanicity
- Grades offered

AIR prioritized reading performance over demographic variables, placing a greater emphasis on Grade 1 DIBELS than Kindergarten DIBELS because first grade scores are generally more predictive of future reading performance (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009) and more reflective of school reading practices. Although Title I status, urbanicity, and grades offered were considered, exact matches on these variables were not always possible.

For each Pilot school, AIR identified a unique Comparison school and a backup school in case the recommended match was deemed ineligible; PDE approved the recommended matches. Exhibit B1 in Appendix B compares the school-level Pilot and Comparison sample characteristics at the time of selection. In March 2016, two Comparison schools were dropped because they began implementing OG after the Pilot started; the backup matches replaced these

[8] An alternative DIBELS comparison was needed for one Pilot school that used AIMSweb as its universal screener before 2015–16. For this school, LNF was used for matching because it was the only subtest shared by both AIMSweb and DIBELS in kindergarten.

[9] With the exception of the DIBELS data, all variables were standardized based on statewide data. The study team used the population of Comparison schools for the standard deviation of the DIBELS data because the data were not available statewide, and because the standard deviation was too high in the national sample, which would mask differences within the sample.

schools. In Year 2, another Comparison school closed; the majority of students from this school went to a single school, which now serves as the Comparison school.

Intervention Assignment for Evaluating the Effectiveness of the MSL Intervention (RQ 4)

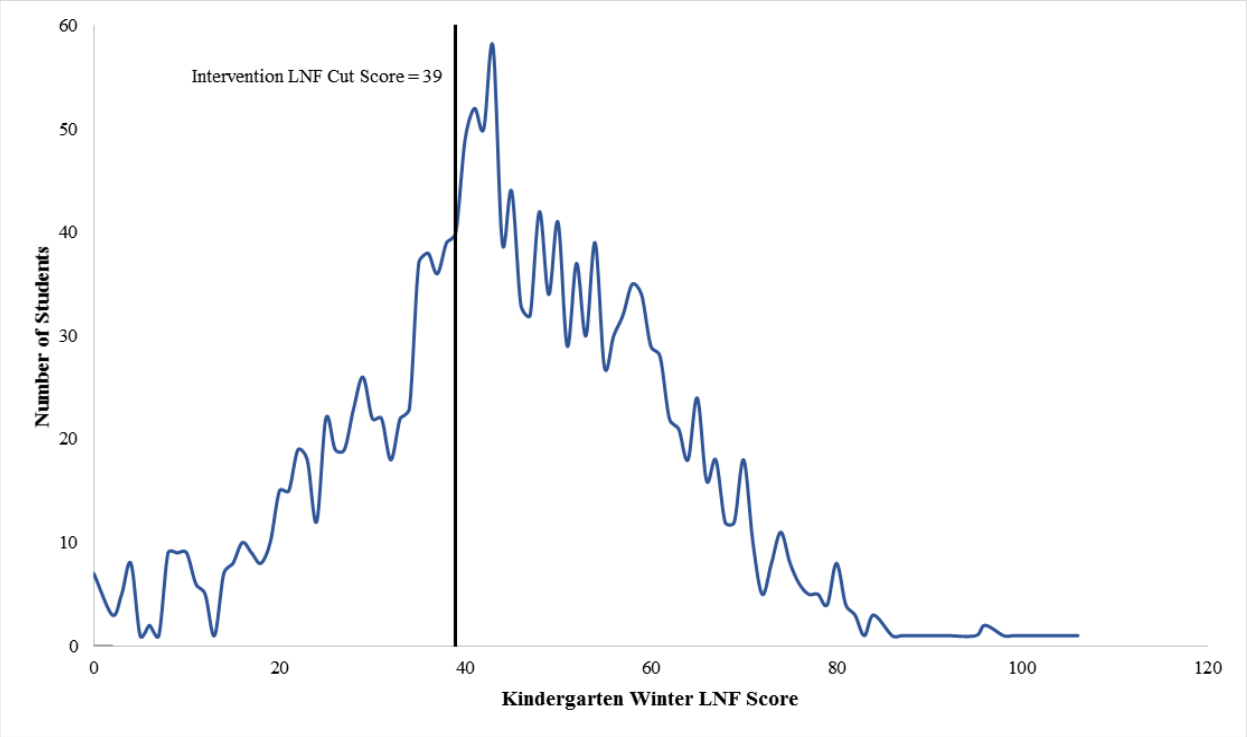
The RD design requires that students be assigned to treatment solely on the basis of a single cut score. Students scoring at or below the cut score qualified for the intervention, while those scoring above did not. RD results were only generalizable near the cut score, so primary RD analyses looked at only those students just above and below the cut score so that the two groups being compared were as similar as possible.

The selection of the measure used to identify students for the intervention was guided by a study by Catts et al. (2009), which examined the predictive power of DIBELS subtests administered in Grades K–3 at four points in the school year (September, December, February, and April). For all subtests administered in kindergarten, predictive analyses used Oral Reading Fluency (ORF) administered in April of third grade as the outcome measure. For the two subtests administered beginning in the fall of kindergarten (Initial Sound Fluency and Letter Naming Fluency [LNF]), the September administration was the least predictive, with predictive power increasing throughout the school year. The authors ascribed this to floor effects, which lessened over time. For the Pilot, DIBELS screening occurred three times a year: September (fall), January (winter), and May (spring). PDE and AIR agreed that intervention assignment should be based on the winter screening, which had lower rates of false positives than the fall screening (i.e., lower rates of incorrectly identifying a student as at risk for poor reading outcomes) and still allowed time for the MSL intervention to take place during the school year. Of the four DIBELS subtests administered in the winter, LNF was selected as the single measure for intervention assignment because it was the most predictive subtest for kindergarteners at the February assessment point of the Catts et al. study.

The RD design requires that the same cut score be used consistently across all Pilot schools. PDE, in consultation with AIR, selected the 35th percentile (across all Pilot schools in the first kindergarten cohort) as the best balance between feasibility of intervention implementation and increasing the likelihood that, even with some parents opting out and attrition (e.g., students moving away), all schools would still have intervention students at the end of the study. For the first kindergarten cohort, the 35th percentile LNF score in January 2016 was 39. The same score was used to determine which students qualified for the intervention in the second kindergarten cohort.[10] Students with scores of 39 or below qualified for the intervention, while students with a score of 40 or higher did not qualify. Because the cut score was identified based on current data, Pilot schools did not know the exact cut score before Cohort 1 screening and therefore could not have altered Cohort 1 students' scores to manipulate who was assigned to the MSL intervention. The distribution of the January 2017 kindergarten LNF scores also suggests that Cohort 2 students' scores were not manipulated to assign students to the MSL intervention. Distributions for both cohorts are presented in Exhibits 2.1 and 2.2.

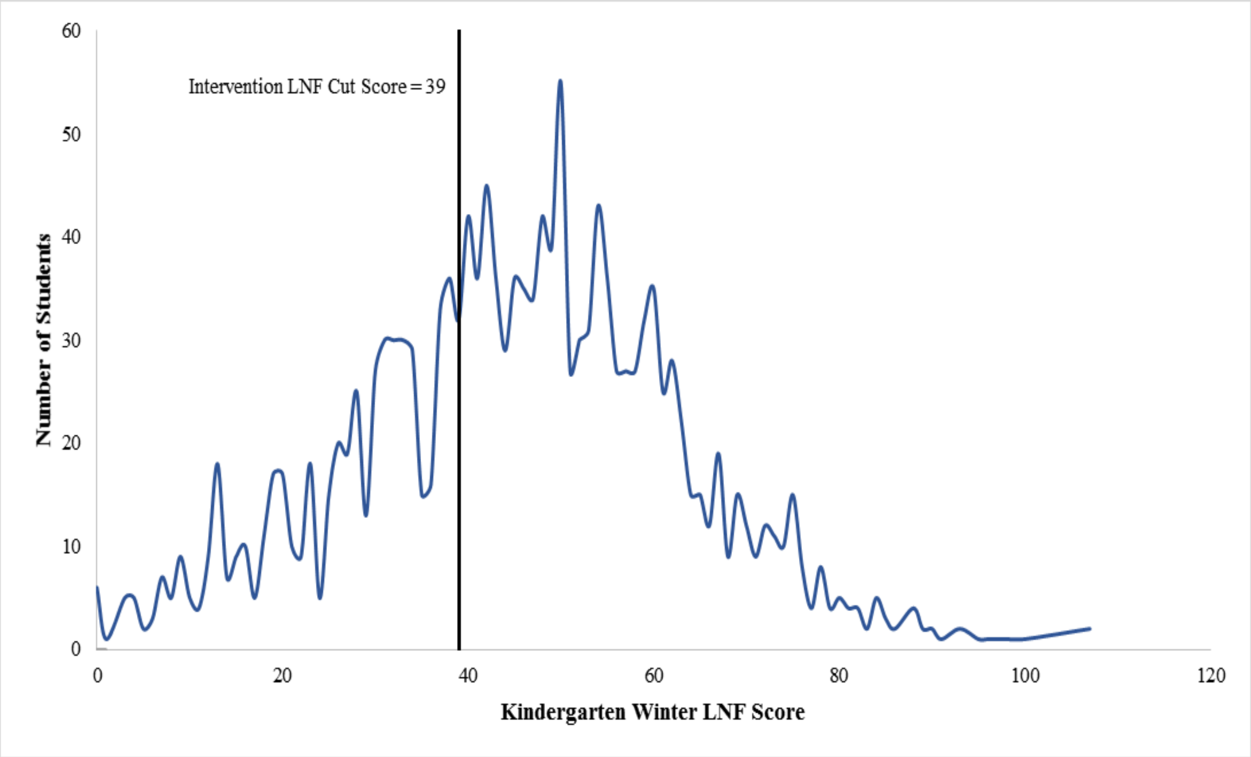
[10] As with the first cohort, 35% of kindergarten students in the second cohort qualified for intervention when using the January LNF cut score of 39.

Exhibit 2.1. Cohort 1 January 2016 Kindergarten LNF Distribution



Source: Pilot district DIBELS data.

Exhibit 2.2. Cohort 2 January 2017 Kindergarten LNF Distribution

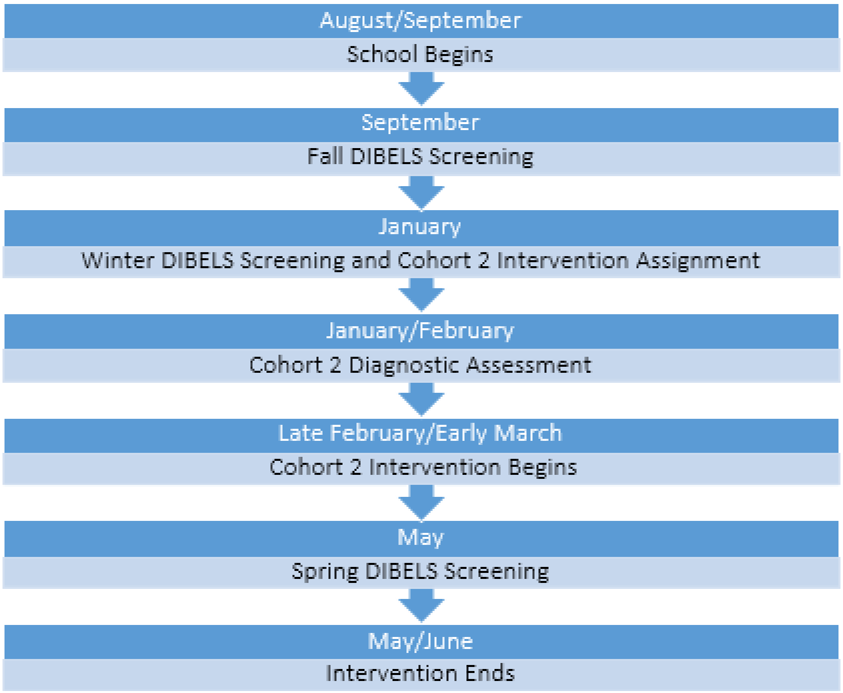


Source: Pilot district DIBELS data.

Evaluation Timeline

Matched Comparison schools were identified in summer 2015. The school year began in August 2015 for all but one Pilot district, which started in September. At this time, Pilot implementation began with kindergarten students at the beginning of the 2015–16 school year. Exhibit 2.3 provides an overview of the implementation timelines for Year 2. (The timeline was similar in Year 1.) Exhibit 2.4 provides an overview of the participation of each cohort, projected to Year 3 of the Pilot.

Exhibit 2.3. Key Year 2 Evaluation Activities, 2016–17 School Year



August/September
School Begins

September
Fall DIBELS Screening

January
Winter DIBELS Screening and Cohort 2 Intervention Assignment

January/February
Cohort 2 Diagnostic Assessment

Late February/Early March
Cohort 2 Intervention Begins

May
Spring DIBELS Screening

May/June
Intervention Ends

Source: Pilot district attendance and DIBELS data, and Pilot school intervention logs and diagnostic data.

Exhibit 2.4. Participation Timelines by Cohort

	2015–16 School Year	2016–17 School Year	2017–18 School Year
Cohort 1	Kindergarten	First grade	Second grade
Cohort 2		Kindergarten	First grade
Cohort 3			Kindergarten

Note. Data from the 2015–16 and 2016–17 school years are the focus of this report. Data from the 2017–18 school year will be included in a future report.

Data Sources

All evaluation data were provided by PDE. Basic student information included roster, demographic, and attendance data sets. Additional data sources are described in the following paragraphs. The first two sources—DIBELS Next and diagnostic assessments—both provided student assessment data. The DIBELS Next measures were used for sample comparisons, MSL intervention assignment, and outcomes analyses. PDE collected diagnostic data to better understand the instructional needs of intervention students, and to compare the students within 5 points (above or below) of the MSL intervention cut score on DIBELS LNF.

DIBELS Next

DIBELS Next was administered three times each year as a universal screener in all Pilot and Comparison schools by trained school staff.[11] The four subtests administered in kindergarten are described in Exhibit 2.5. Each subtest took 1 minute to administer. According to the Center on Response to Intervention (CRTI) Screening Tools Chart,[12] the four subtests used as outcome variables (LNF, Phoneme Segmentation Fluency [PSF], Nonsense Word Fluency [NWF], and ORF) are reliable measures.

Exhibit 2.5. DIBELS Next Subtests Used in Pilot Year 2

Subtest	Administration Dates	Brief Description of Measure
First Sound Fluency (FSF)	Fall and winter of kindergarten	Number of initial sounds a student identifies in words read by the examiner. Used to examine baseline equivalence.
Letter Naming Fluency (LNF)	Fall, winter, and spring of kindergarten; fall of first grade	Number of letters a student correctly names from a sheet of random lower- and uppercase numbers. Used to assign students to the intervention after the kindergarten winter screening. Serves as an outcome variable.
Phoneme Segmentation Fluency (PSF)	Winter and spring of kindergarten; fall of first grade	Number of correct sound segments a student identifies in words read by the examiner. Used as an outcome variable.
Nonsense Word Fluency (NWF)	Winter and spring of kindergarten; fall, winter, and spring of first grade; fall of second grade	Correct letter sounds (CLS) and whole words read (WWR) from a sheet of nonsense words; only CLS is recommended in kindergarten. Used as an outcome variable.
Oral Reading Fluency (ORF)	Winter and spring of first grade; all of second through sixth grade	Number of words read correctly in one minute (words correct; WC) from an unfamiliar passage of grade-level text and percentage of attempted words that were read correctly (accuracy). Used as an outcome variable.

Source: DIBELS Next technical manual (Good et al., 2013).

Diagnostic Assessments

PDE required Pilot schools to conduct diagnostic assessments for students qualifying for the intervention (i.e., kindergarten students with a winter LNF of 39 or less) and those who scored within 5 points above the cut score (i.e., kindergarten winter LNF of 40 to 44). These tests were intended to provide information on intervention students' needs and to help compare students just below and above the MSL intervention cut score. Qualified professionals, such as speech-language pathologists reading specialists, and psychologists, individually administered two norm-referenced tests (see Exhibit 2.6).

[11] All Pilot schools (with the exception of one) used DIBELS before the Pilot. The one school that did not use DIBELS before the Pilot previously used AIMSweb, which meant that teachers were familiar with universal screening. Pilot schools had a recalibration day in 2015–16.

[12] http://www.rti4success.org/resources/tools-charts/screening-tools-chart

Exhibit 2.6. Diagnostic Tests

Test	Brief Description
Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4)	Assesses receptive vocabulary by asking examinees to point to the picture that best represents the meaning of a spoken word. Reported as standard score with *M* = 100 and *SD* = 15.
Comprehensive Test of Phonological Processing, Second Edition (CTOPP-2)	Assesses phonological awareness, phonological memory, and rapid naming. For the purposes of the Pilot, four subtests were used. Elision measures the ability to remove phonological segments from spoken words to form other words. Blending measures the ability to synthesize sounds to form words and nonwords. Rapid Color Naming measures the ability to name colors rapidly. Rapid Object Naming measures the ability to rapidly name objects. Subtest scaled scores have a mean of 10 and a standard deviation of 3.

Source: PPVT-4 information from http://images.pearsonassessments.com/images/assets/ppvt-4/2013-PPVT-Tech-RPT.pdf and CTOPP-2 information from http://www.slossonnews.com/CTOPP-2.html.

Fidelity of Training Measures

The majority of interventionists received OG training from the Compass Reading Center (Compass; see Appendix C for a summary of Initial Level Certification OG training requirements and a comparison of the three training levels). For these interventionists, PDE provided pretest/posttest scores and fidelity observation data as evidence of training completion. Pretests and posttests assessed knowledge related to OG implementation. Fidelity observations (described below) assessed trainees' implementation of key components of OG. Trainees must maintain an 85% fidelity average for successful completion. For other intervention programs such as Wilson and Sonday, onsite coaching was provided to support training and strengthen fidelity of implementation. For the classroom program, PDE had sign-in sheets and fidelity data (see below).

Fidelity of Implementation Measures

Classroom Program. For implementation of the classroom lessons, PDE developed an observation protocol and a teacher self-report survey. Although some data were collected with these measures in Year 1, data were not analyzed because of concerns that the protocols were not yielding reliable data. PDE updated both tools for Year 2 and provided data to AIR. The observation protocol focused on reading instructional practices covered in training, including phonological awareness, phonics, oral language, vocabulary, fluency, and comprehension. For each category, observers marked whether or not they observed specific practices and the minutes of instruction. The self-report asked teachers how often, over the course of the past week, they engaged in specific practices in each of these categories, as well as how often they used each of the following: Fundations, Accessing the Code, Neuhaus, and OG.

Intervention Observations. Data on interventionists' implementation of key intervention components were available only for OG interventionists trained by Compass, who were observed by Compass trainers using the Initial Trainee Observation Evaluation form.[13] Scores are the number of 100 possible points awarded for the implementation of various OG components.

Intervention Logs. Interventionists kept logs of each student's intervention minutes per day. They also noted which MSL intervention program they implemented.

Evaluation Sample

Districts interested in participating in the Pilot were required to have all-day kindergarten, and to commit to participating in training and implementation of the classroom and intervention components, as well as study data collections (see previous section). PDE provided financial support for meeting these requirements. Among interested districts, PDE selected eight districts to reflect a variety of locations and sizes. All of the 21 elementary schools within the eight participating districts are participating in the Pilot. (Throughout the report, districts and schools are discussed using pseudonyms to protect their identity.) The Comparison sample consists of 21 schools, each matched to a unique Pilot school (see the *Overall Design* section for a description of matching variables and procedures). To ensure the availability of DIBELS Next data for Comparison school students, the matched schools were selected from a set of districts participating in a different literacy project; Comparison schools are therefore expected to have stronger reading instruction and performance than typical Pennsylvania schools not engaged in a literacy initiative. This Comparison group may reduce the discernable magnitude of the classroom program's effect, compared to the potential effect we may see if we compared the Pilot schools to schools without any form of intervention.

Classroom Program Sample (RQ 3)

The evaluation sample for the classroom program analyses included only those students who (a) participated in DIBELS screening in the fall, winter, and spring of their kindergarten year, and (b) had data for all covariates included in the main model. Exhibit 2.7 shows the key baseline characteristics of the Pilot and Comparison classroom program samples, by cohort. As shown, the two groups were comparable on most measured characteristics, with three exceptions. In Cohort 1, the pilot and matched samples differed on the percentage of students eligible for free or reduced-price lunch (33.3% of Pilot students compared with 47.2% of Comparison students), and the percentage identified as multiracial (7.3% of Pilot students compared with 3.8% of Comparison students). Cohort 2 students differed only on the proportion of special education students (7.7% of Pilot students compared with 13.5% of Comparison students). In both cohorts, the Pilot and Comparison groups had similar baseline DIBELS performance (fall First Sound Fluency [FSF] and fall LNF).

[13] Compass provided OG training in six of eight districts, although in some districts not all interventionists were trained in OG (these districts implemented multiple MSL intervention programs). Exhibit 3.5 summarizes the number of interventionists implementing each program, overall and by district.

Exhibit 2.7. Classroom Program Analysis Baseline Sample Characteristics, Cohorts 1 and 2

Variable	Pilot Group Mean	Comparison Group Mean	Estimated Difference	*p*-Value
Cohort 1 Students				
Baseline (fall) FSF	16.4	15.6	0.8	.434
Baseline (fall) LNF	21.9	21.6	0.3	.781
Female (percentage)	49.7	47.5	2.2	.353
Free or reduced-price lunch (percentage)	33.3	47.2	-13.9*	.004
Special education (percentage)	8.6	11.7	-3.1	.089
Race/ethnicity (percentage)				
African American	2.6	5.0	-2.4	.111
Asian, Pacific Islander, Native American	1.8	1.7	0.0	.937
Hispanic	7.0	5.9	1.2	.474
Multiracial	7.3	3.8	3.5*	.018
White	80.8	83.6	-2.7	.470
Cohort 2 Students				
Baseline (fall) FSF	13.8	14.3	-0.5	.591
Baseline (fall) LNF	19.9	19.2	0.7	.403
Female (percentage)	48.7	50.2	-1.5	.513
Free or reduced-price lunch (percentage)	35.7	45.0	-9.3	.076
Special education (percentage)	7.7	13.5	-5.8*	.001
Race/ethnicity (percentage)				
African American	3.5	4.9	-1.4	.334
Asian, Pacific Islander, Native American	2.2	1.8	0.4	.557
Hispanic	5.9	6.6	-0.7	.583
Multiracial	7.4	4.9	2.5	.158
White	80.7	81.8	-1.4	.711

Note. Sample size for Cohort 1 = 2,736 students (1,591 Pilot and 1,145 Comparison); sample size for Cohort 2 = 2,841–2,868 students (1,519–1,524 Pilot and 1,301–1,344 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District demographic and DIBELS data.

Intervention Sample (RQ 4)

After the kindergarten winter screening for each cohort, parents were notified if their child qualified for the intervention (i.e., scored 39 or less on the winter LNF). PDE provided sample parent information and consent form templates to Pilot districts. Depending on district policy, parents provided active or passive consent for their child's participation in the intervention (see Appendix E for an example notification and opt-out form). Cohorts 1 and 2 had similar levels of participation in the intervention in kindergarten: 484 of 603 (80.3%) Cohort 1 students and 446 of 567 (78.6%) Cohort 2 students participated through the end of kindergarten. Fewer Cohort 1 students participated in the intervention through the end of first grade (408 of 603, or 67.7%), because of students moving away from participating schools between their kindergarten and first grade years. However, these 408 students made up 83.6% of the Cohort 1 students who remained in the schools and continued on to the first grade. (See Exhibit C1 in Appendix C for Pilot sample retention and attrition.) Only those students with kindergarten winter DIBELS scores and with spring 2017DIBELS scores were included in outcome analyses; students whose parents did not consent to the MSL intervention were not excluded from the outcome analyses, although these students did not receive the MSL intervention.

Because the MSL intervention students had, by definition, lower winter LNF scores than nonintervention students, the two groups were assumed to be dissimilar in other ways (e.g., other academic measures, socioeconomic status). The study team confirmed this (see Exhibit D1 in Appendix D). However, as the sample was further restricted around the cut score, the groups became more similar. For example, although the Cohort 2 MSL intervention students differed on nearly all baseline measures from nonintervention students, when focusing in on the primary RD analysis sample (i.e., students who had an LNF score ±5 points from the cut score), the Cohort 2 MSL intervention students differed from nonintervention students only in the percentage who were minority, the percentage who were eligible for free or reduced-price lunch, and their fall LNF DIBELs scores. Intervention students were significantly more likely to be minority students and eligible for free or reduced-price lunch, and they scored significantly lower than nonintervention students on the fall LNF measure (see Exhibit D2 in Appendix D). When controlling for kindergarten winter LNF, however, the two groups were similar (no significant estimated differences) on all measured characteristics in both cohorts (Exhibit 2.8).

Exhibit 2.8. Discontinuity Estimates for Baseline Intervention Analysis Sample Characteristics, Main RD Analysis Sample (Kindergarten Winter LNF 35–44), Cohorts 1 and 2

	Cohort 1		**Cohort 2**	
Variable	**Estimated Difference**	***p*-Value**	**Estimated Difference**	***p*-Value**
Student Characteristics (percentage)				
Female	-3.1	.748	-5.1	.645
White	-3.4	.654	-8.1	.268
Free or reduced-price lunch	11.0	.203	8.8	.399
Special education	-8.4	.100	-8.8	.109
Diagnostic Assessments				
PPVT	-0.9	.742	4.4	.180
CTOPP-2 Blending	0.1	.907	0.5	.413
CTOPP-2 Elision	1.0	.079	-0.3	.571
CTOPP-2 Rapid Color Naming	0.2	.654	-1.0	.331
CTOPP-2 Rapid Object Naming	-1.0	.075	0.3	.737

Note. Student characteristics sample size for Cohort 1 = 437–438 students (189–190 intervention and 248 nonintervention students); student characteristics sample size for Cohort 2 = 315 students (130 intervention and 185 nonintervention students); student diagnostic assessments sample size for Cohort 1 = 343–362 students (152–164 intervention and 191–198 nonintervention students); student diagnostic assessments sample size for Cohort 2 = 263–264 students (109–110 intervention and 154 nonintervention students). The analyses were based on a two-level regression (students within schools). The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*). In this analysis, intervention students had a kindergarten winter LNF score between 35 and 39, and nonintervention students had a kindergarten winter LNF score of 40–44.
Source: District demographic and school diagnostic data.

Interventionists

Intervention was implemented by 72 interventionists in Year 1. In Year 2, the number varied from 75 to 127 across months, with 102 at the end of the year. Chapter 3 provides information on their training and the programs they implemented.

Data Analysis

This section briefly describes the models used to analyze the effects of the classroom program and the intervention on end-of-year DIBELS. Appendix F provides further technical details on the statistical model specifications, including power calculations. Analyses for both effectiveness RQs used a two-level analysis (students within schools). For this report, the spring DIBELS measures (LNF, PSF, and NWF for Cohort 2 kindergarten students, and NWF and ORF for Cohort 1 first grade students) served as outcome variables.

Classroom Program Analyses

To examine the effect of the classroom program, the study team compared the spring DIBELS scores of students in the Pilot and Comparison schools, controlling for the matched pair blocks, schools' Title I status, students' free or reduced-price lunch status, special education status, and baseline DIBELS LNF and FSF scores. For each measure, the study team compared the Pilot and Comparison means using a two-level hierarchical linear model, with students nested within schools (see Appendix F for technical details). As noted, Pilot student performance in these analyses may have been influenced by both the classroom program, received by all students, and the MSL intervention, received by some students.

Intervention Analyses

To examine the effect of the MSL intervention, the study team compared the spring DIBELS scores of students in the MSL intervention with students not in the MSL intervention, controlling for kindergarten winter LNF score. As with the classroom program analyses, the intervention and nonintervention spring DIBELS means were compared using a two-level hierarchical linear model, with students nested within schools (see Appendix F for technical details).

Chapter 3. Implementation

Classroom Program Training and Implementation

The classroom program aims to enhance core instruction by providing classroom teachers with professional development aligned with the recommendations of the National Early Literacy Panel (2008), the National Reading Panel (2000), and other evidence-based approaches (Adams, 1990; Foorman et al., 2016; Kosanovich & Foorman, 2016; National Research Council, 1998; Shanahan et al., 2010). In addition to training, Pilot sites receive ongoing technical assistance from Pennsylvania Training and Technical Assistance Network (PaTTAN) staff assigned to each district. This section describes classroom program training and implementation.

Classroom Teacher Training

PDE provided both training and materials to support classroom program implementation. In spring 2015, PDE provided Pilot kindergarten classroom teachers and interventionists with 4 days of training in the first three modules of Language Essentials for Teachers of Reading and Spelling (LETRS). In the six districts that received OG training from Compass (see the following section), kindergarten classroom teachers participated in the first 20 hours of the OG training. OG trainers also provided training in Neuhaus Reading Readiness (a half day or whole day, depending on site preference).

In summer 2016, first grade teachers were trained in the classroom program. The goals of the training included the following:

- Deepen knowledge and skills of the essential components of reading instruction (phonemic awareness, phonics, vocabulary, fluency, comprehension, and oral language)
- Integrate multisensory structured language/literacy (MSL) methods into the core
- Build explicit, direct, sequential, systematic instruction practices into the classroom

First grade teachers received a 4-day training series based on the recommendations of the National Early Literacy Panel (2008), the National Reading Panel (2000), and other evidence-based approaches (Adams, 1990; Foorman et al., 2016; Kosanovich & Foorman, 2016; National Research Council, 1998; Shanahan et al., 2010). The topics were as follows:

- Day 1: Overview of the Pilot, Theoretical Frameworks, Oral Language Development, and Teaching the Five Essential Components of Reading Instruction (the “Big Five”).
- Day 2: Phonemic Awareness and Phonics
- Day 3: Oral Language and Vocabulary
- Day 4: Fluency and Comprehension

According to a teacher-level attendance summary provided by PDE, only 73% of first grade teachers attended all four trainings. However, 95% attended at least one training, and 89% attended three of the four trainings. Monet School District chose to give an additional day of training for a Fundations program refresher, which eight of 13 (62%) first grade teachers and six of 11 (55%) kindergarten teachers chose to attend.

In addition, kindergarten teachers in all districts were given the option to attend a refresher training and/or any of the trainings held for first grade teachers. Because these trainings were optional for kindergarten teachers, attendance was fairly limited among kindergarten teachers across all districts; nonetheless, many kindergarten teachers did still attend the trainings. Appendix C, Exhibit C2, provides more information about training attendance.

Implementation of the Classroom Program

Exhibit 3.1 lists the core reading programs and MSL interventions implemented in Year 2. To protect confidentiality, pseudonyms are used in place of participating site names. More information on MSL intervention programs is provided later in this chapter.

Exhibit 3.1. Core Reading Programs and MSL Interventions Implemented in 2016–17

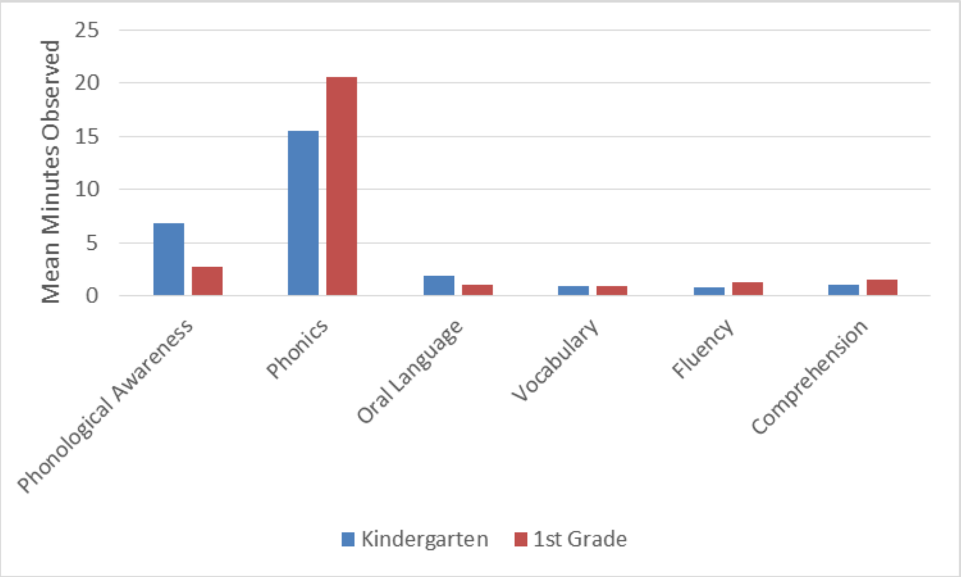
District	Core Program	Core Supplement	MSL Intervention
Dali School District	Wonders	Fundations	OG (Compass) Lindamood Phoneme Sequencing Program for Reading, Spelling, and Speech (LiPS)
Degas School District	Reading Street, Common Core Edition	Accessing the Code Neuhaus Reading Readiness	OG (Compass)
Kahlo Public Schools	Wonders	Accessing the Code	OG (Compass) Sonday
Michelangelo School District	Wonders	Neuhaus Reading Readiness in first semester of kindergarten Accessing the Code in second semester of kindergarten and first grade	Sonday
Monet School District	LEAD21 core	Fundations and Neuhaus Reading Readiness	OG (Compass)
Picasso Public Schools	Treasures	Fundations	Wilson WRS LiPS
Pollock Public Schools	Wonders	Fundations	OG (Compass)
Warhol School District	Story Town K–6	Accessing the Code	OG (Compass) Sonday

Source: Information provided by PDE.

DIBELS data confirmed that all Pilot schools conducted screening three times per year, as planned. In 2015–16, PDE reported that classroom implementation was low, due in part to a focus on intervention training and implementation. (Classroom implementation data were not analyzed due to concerns about their reliability.) For 2016–17, PDE strengthened teacher training (see above) and better aligned fidelity observations and self-reports to the lesson plans to yield more reliable data. These tools did not include criteria for adequate fidelity of implementation but do offer evidence of implementation, along with some descriptive information on instructional practices.

From October through December 2016, trained raters observed kindergarten and first grade classroom teachers. Observation data suggest that, on average across all districts, most of the 30-minute lessons were devoted to skills covered in training (*M* = 27 minutes for kindergarten and *M* = 28 minutes for first grade). As seen in Exhibit 3.2, most of this time was devoted to phonics, followed by phonological awareness, with a stronger emphasis on phonics in first grade (*M* = 21 minutes) than in kindergarten (*M* = 16 minutes). These emphases varied by district (see Exhibit C3 in Appendix C). In kindergarten, phonics instruction ranged from about 12 minutes to 20 minutes by district, and phonological awareness instruction ranged from about 2 to 9 minutes per district. In first grade, phonics instruction ranged from about 15 minutes to 24 minutes by district, and phonological awareness instruction ranged from about 1 to 5 minutes per district.

Exhibit 3.2. Reading Instruction Observed in Year 2, by Category

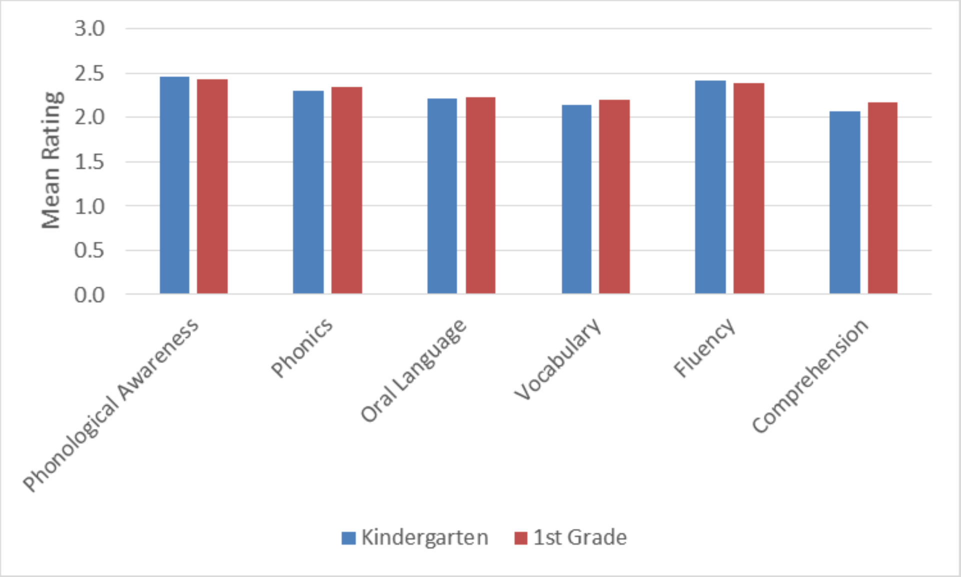


Note. Sample size = 159 Pilot school kindergarten and first grade teachers.
Source: Fidelity observation means provided by PDE.

PDE also provided self-report data collected from 143 teachers in January, February, and March 2017, reflecting 1 week of instruction. Teachers were asked how often they engaged in specific instructional practices using the following scale: 0 = *Not at all*, 1 = *A little*, 2 = *Sometimes*, and 3 = *A lot*. PDE merged these practices into the same five categories reported for the classroom revisits. As seen in Exhibit 3.3, teachers reported more balanced instruction across the categories than was seen in the classroom visits. Note that the self-report surveys reflected a week of instruction rather than a single observed lesson and occurred somewhat later in the school year. Differences between grades were small, with kindergarten teachers (Cohort 2) reporting means from 2.1 to 2.5 across categories, compared to 2.2 to 2.4 for first grade teachers (Cohort 1).

These means reflect a range from *sometimes* to midway between *sometimes* and *a lot*. As seen in Exhibit C4 in Appendix C, mean ratings varied by district, with some districts reporting greater variation across categories. The range of means across districts within a category was greater for first grade (Cohort 1). For example, the mean rating for comprehension instruction ranged from 1.4 (about midway between *a little* and *sometimes*) to 2.7 (closer to *a lot* than *sometimes*).

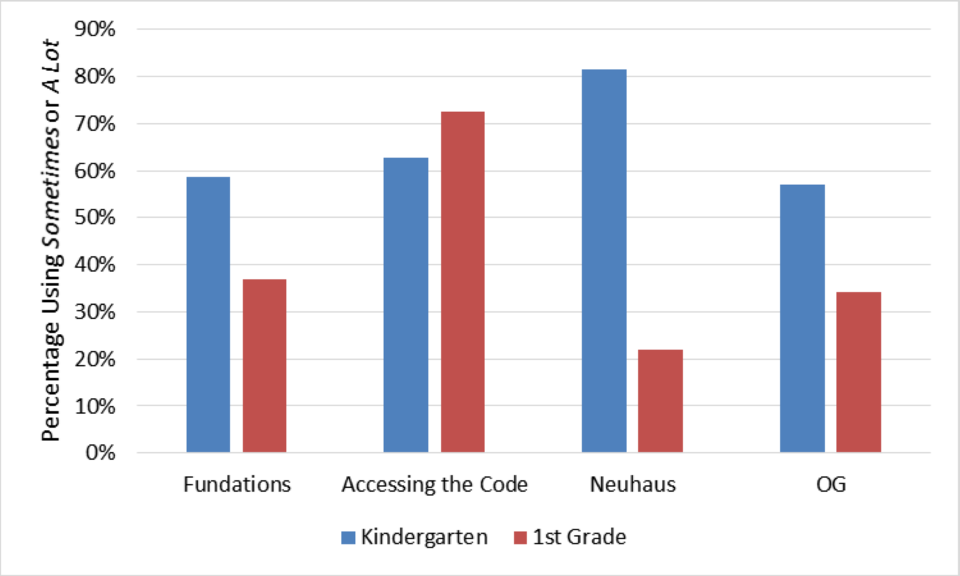
Exhibit 3.3. Self-Reported Reading Instruction by Category



Note. Sample size = 73 first grade teachers and 70 kindergarten teachers. Means reflect a 0–3 scale (described in text).
Source: Fidelity self-report means provided by PDE.

Teachers also reported how often they used each of four possible programs, using the same scale. Exhibit 3.4 shows the percentage of teachers who reported using each program *sometimes* or *a lot* of the time. At least half of all kindergarten teachers reported using each program at least sometimes; Neuhaus was most popular (81%). First grade teachers most often used Accessing the Code (73%); less than 40% of first grade teachers used the other three programs at least sometimes.

Exhibit 3.4. Self-Reported Use of Programs



Note. Sample size = 73 first grade teachers and 70 kindergarten teachers. Note that means do not reflect ratings from all teachers, some of whom selected *not applicable* for each program. Means reflect a 0–3 scale (described in text).
Source: Fidelity self-report data.

As would be expected, program use (defined as using the program at least sometimes) varied by district (see Exhibit C5 in Appendix C). All Dali School District teachers used Fundations. Degas School District teachers reported using only Neuhaus and OG. All Kahlo Public Schools teachers reported using Accessing the Code. Michelangelo School District teachers varied by grade, with all kindergarten teachers using Neuhaus and all first grade teachers using Accessing the Code. In Monet School District, all kindergarten teachers used Fundations. First grade teachers showed more variability; OG and Fundations were the most commonly used, though not by all. In Picasso Public Schools, all kindergarten teachers used both Fundations and Accessing the Code, while all first grade teachers used Fundations. Pollock Public Schools varied by grade, with all kindergarten teachers using both Neuhaus and Fundations, and all first grade teachers using Accessing the Code. All Warhol School District teachers used Accessing the Code.

PDE used findings from fidelity observations and self-reports to review Year 2 implementation and provide feedback to districts before Year 3. In spring 2017, PDE also conducted peer-to-peer conversations between classroom teachers by grade level. A notetaking guide for these conversations included questions related to the most beneficial aspects of Pilot professional development, impacts seen at the classroom and student levels, implementation challenges, beneficial Pilot materials, and supports that would be helpful moving forward. PDE included overall findings in their end-of-year summary to districts, with the intention that the information would inform programmatic changes.

Intervention Training and Implementation

The MSL intervention is aimed at providing additional support to students with lower levels of baseline literacy skills (as measured by kindergarten winter LNF). This section describes MSL intervention training and implementation.

Interventionist Training

PDE provided both training and materials to support MSL intervention implementation. Interventionists were trained before Year 1. Additional interventionists were trained during Years 1 and 2, when it was determined additional interventionists were needed. Initial interventionists participated in the spring 2015 LETRS training provided to classroom teachers and received additional training in summer and fall 2015. The additional trainings varied by the MSL intervention program(s) implemented in each district. OG interventionists received the most intensive training, some of which was not specific to the Pilot grades (kindergarten through second grade; Appendix C gives a summary of the Compass Initial Level Certification OG training requirements). However, some interventionists were trained in other MSL programs because of district preference or the need for additional interventionists. Exhibit 3.5 summarizes the number of interventionists implementing each program in Year 1 and Year 2, overall and by district.

Exhibit 3.5. Number of Interventionists Implementing Each Program

District	Orton-Gillingham	Sonday	Wilson (WRS or Fundations), LiPS	DuBard Association Method	Neuhaus	Accessing the Code
Year 1						
Dali School District	3		2			
Degas School District	4					
Kahlo Public Schools	11	8				
Michelangelo School District		2		4		
Monet School District	11					
Picasso Public Schools			3			
Pollock Public Schools	9				3	
Warhol School District	12					
Total	**50**	**10**	**5**	**4**	**3**	**0**
Year 2						
Dali School District	11		7			
Degas School District	9					
Kahlo Public Schools	11	24				
Michelangelo School District		8				
Monet School District	11					
Picasso Public Schools			4			
Pollock Public Schools	13					
Warhol School District	8				1	6
Total	**63**	**32**	**11**	0	**1**	**6**

Note. In Year 2, some interventionists used more than one program. These interventionists were counted once per program, which means that summing the program totals results in more than the actual number of interventionists.
Source: Pilot school intervention logs.

In Year 1, 48 of the 50 OG interventionists received Compass Initial Level Certification training, as evidenced by pretest, posttest, and fidelity observation scores. Of the two interventionists not participating in this training as part of the Pilot, one was a Compass trainer (and thus already trained) and the other replaced a trainee (who left the school) and received briefer on-site training. The 48 OG interventionists trained by Compass all showed growth in knowledge from pretest (*M* = 36.7%, *SD* = 11.5) to posttest (*M* = 91.3%, *SD* = 9.6). An additional 35 OG interventionists received Compass Initial Level Certification training in Year 2. They all improved from pretest (*M* = 29.8%, *SD* = 13.9) to posttest (*M* = 94.7%, *SD* = 5.3). Also in Year 2, 20 OG interventionists trained in Year 1 continued with Level 2, intermediate training (see comparison in Appendix C). This did not include tests but did require additional fidelity observations. Fidelity data are summarized in the section on *Implementation of Intervention*.

Two districts chose not to implement OG in either year. The first of these, Michelangelo School District, trained their interventionists in the Dubard Association Method. University of Southern Mississippi faculty provided 1 week of on-site training and 1 year of follow-up on-site and phone meetings. The second, Picasso Public Schools, trained their interventionists in Wilson Reading System (WRS) and Lindamood Phoneme Sequencing Program for Reading, Spelling, and Speech (LiPS), with training provided by the program developers.

After intervention assignment in Year 1, four districts needed more interventionists than had been trained. (These are the four districts in Exhibit 3.5 that implemented more than one MSL intervention program in Year 1.) As time constraints did not permit new interventionists to receive the extensive OG training, 15 additional interventionists received training in other MSL programs. PDE offered training in Sonday. Kahlo Public Schools and Michelangelo School District accepted Sonday training from Winsor Learning. Pollack Valley School District chose to extend the Neuhaus Reading Readiness program that was introduced for kindergarten teachers by the Compass trainers in the fall; the interventionists received additional support from PaTTAN staff who had taken a 6-hour course in the program. Dali School District used Wilson Fundations in their kindergarten classrooms and elected to provide a double dose of Fundations to intervention students not receiving OG. This district had a Wilson trainer on-site. The district also participated in LiPS training in summer 2015 and implemented LiPS as part of the intervention.

Some districts displayed different program usage in Year 2. Michelangelo School District continued using Sonday but stopped using the Dubard Association Method. Pollock Public Schools continued using OG but stopped using Neuhaus. Warhol School District implemented Neuhaus and Accessing the Code in addition to OG.

Implementation of Intervention

Intervention logs revealed that all Pilot schools implemented the MSL intervention. However, participation in the intervention did not always adhere to the cut score. As seen in Exhibit C1 in Appendix C, some students who qualified for the intervention did not participate (16% in Cohort 1 and 18% in Cohort 2, of students still in the sample at the end of Year 2). As previously described, this generally reflected cases where parents decided their children would not participate; rarely, students were deemed ineligible for the participation (e.g., an individualized education program [IEP] team determined it was not a good fit). Conversely, some students who did not qualify for the intervention received the intervention in Year 2 (33 Cohort 1 students and 12 Cohort 2 students). This imperfect adherence threatens the intent-to-treat RD design.

The implementation summaries in this section reflect the RD sample (requiring a kindergarten winter LNF score below the cut score and spring 2017 DIBELS scores, dropping students retained in kindergarten), with the additional requirement that qualified students must have received some intervention hours (so that opted-out students did not affect these means, providing a better sense of how the intervention was implemented). Please note that these average hours may underestimate the dose received by a typical intervention student due to cases where a student began but discontinued the intervention for some reason (e.g., due to a parental or IEP team decision). Conversely, these averages overestimate the dose received by the typical student below the cut score, due to the students who qualified for but did not receive the intervention (but were still included in the main intent-to-treat RD outcome analyses).

The number of interventionists seeing students varied by month, ranging from 75 to 127, and students would frequently switch between interventionists/groups on a daily, weekly, or monthly basis. By the end of the 2016–17 school year, there were 102 interventionists in total. As previously discussed, interventionists implemented a variety of MSL intervention programs. Exhibit 3.6 provides the number and percentage of Year 2 intervention students receiving each intervention program. OG remained the most popular method, although the percentage for both cohorts was lower than in Year 1, when 72% of Cohort 1 kindergarten students received OG.

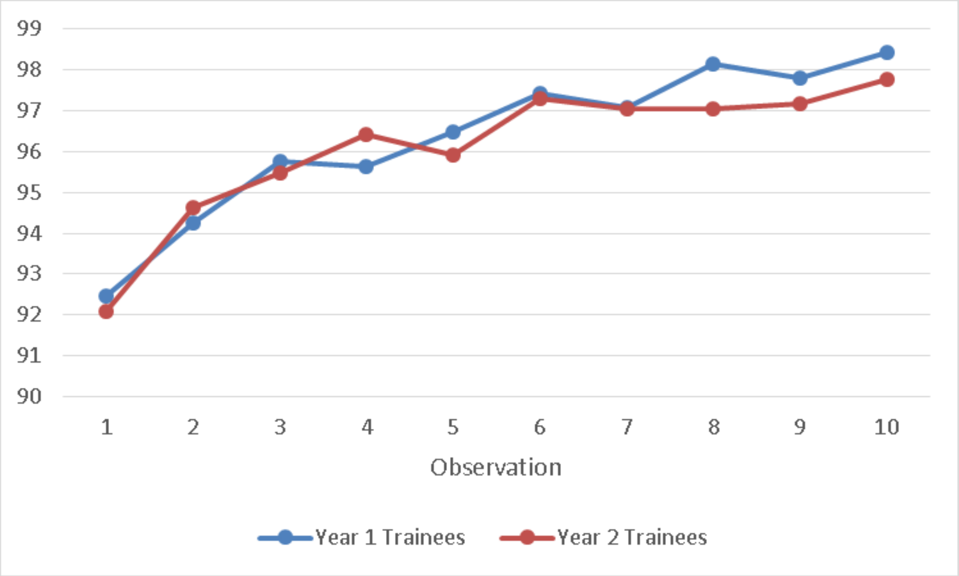
Exhibit 3.6. Number and Percentage of Intervention Students Receiving Each Intervention Program in Year 2, by Cohort

Method	Number of Students	Percentage of Students
Cohort 1 (Grade 1)		
Accessing the Code	50	12.4%
Orton-Gillingham	256	63.5%
Wilson (WRS or Fundations), LiPS	19	4.7%
Neuhaus	2	0.5%
Sonday	76	18.9%
Cohort 2 (Kindergarten)		
Accessing the Code	35	8.0%
Orton-Gillingham	230	52.5%
Wilson (WRS or Fundations), LiPS	36	8.2%
Neuhaus	1	0.2%
Sonday	136	31.1%

Note. Sample size = 841 (Cohort 1 sample size = 403, Cohort 2 sample size = 438); this does not include all intervention students because the method was not reported for all Cohort 2 students. Many students switched groups and methods during the course of the year; for the purpose of this summary, assignments were made based on which method was most frequently used during a student's intervention time throughout the 2016–17 school year.
Source: Pilot school intervention logs.

The extent to which interventionists implemented key intervention components was assessed for interventionists trained in the OG method by Compass, which requires that trainees maintain a mean score of 85 (out of a maximum score of 100).[14] All Compass-trained interventionists maintained a mean score above 85. For Year 1 trainees, the average score was 96, with a range from 88 to 99. Mean scores were similar across districts, ranging from 92 to 99. For Year 2 trainees, the average score was 96, with a range from 90 to 99. Again, scores were similar across districts, ranging from 94 to 99. Exhibit 3.7 summarizes the fidelity data across all districts by observation. These data, which show generally increasing scores across the 10 observations, suggest high overall adherence to OG components, throughout the year, and show that interventionists improved their adherence to the OG program over time. However, these observations were announced to interventionists, and scores may not reflect implementation on a typical day.

Exhibit 3.7. Mean OG Fidelity Score Across Interventionists, by Observation



Source: Compass OG fidelity observation data.

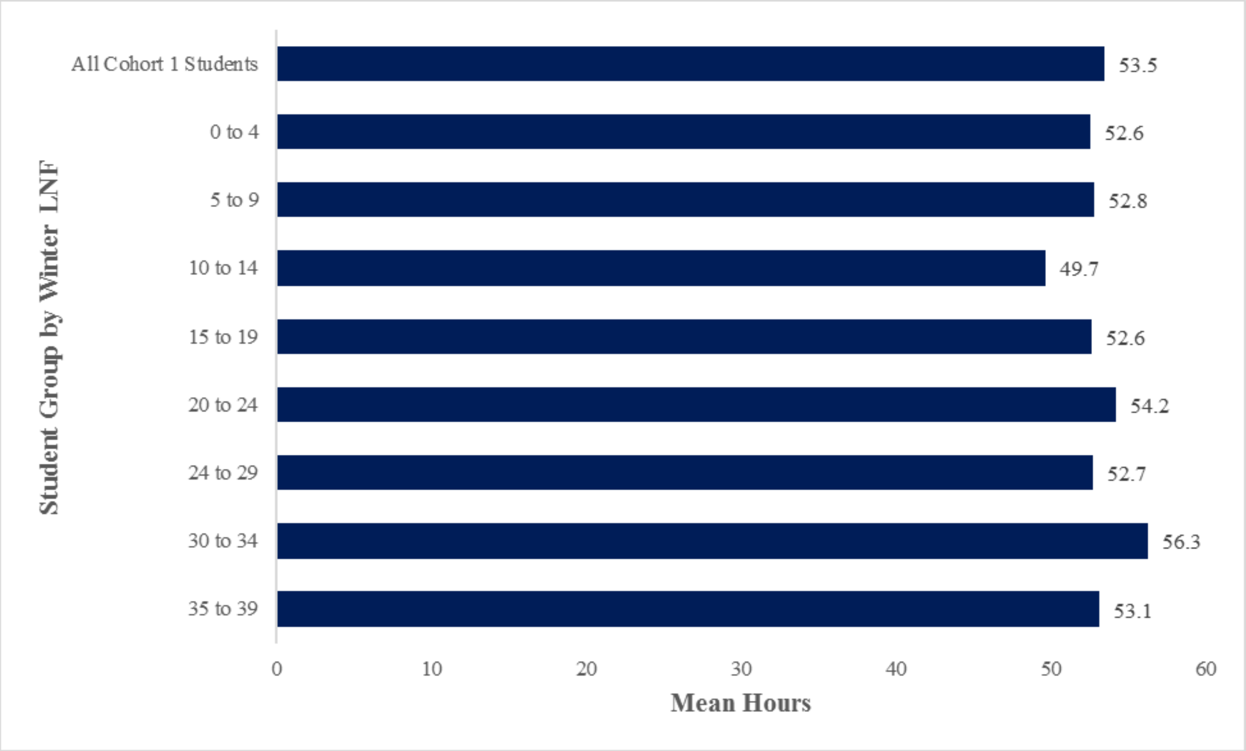
In Year 2, 20 OG interventionists continued with Level 2 training. Of these, 18 completed the required four fidelity observations, with an average score of 97 (interventionists' averages ranged from 92 to 100).

Cohort 1 students who qualified for and participated in the intervention received, on average, 54 hours of time in the intervention in Year 2 (*SD* = 14). This fell short of the target of 70 hours in first grade. Means varied by school, from approximately 37 hours at Debussy Elementary to 73 hours at Ravel Elementary (see Exhibit C6 in Appendix C). Across both kindergarten and first grade, these students averaged approximately 76 hours (*SD* = 15.8), which fell short of the target

[14] OG Compass Level 1 training requires 10 observations conducted by a Compass trainer using the Initial Trainee Observation Evaluation form. In Year 1, the expected 10 observations were completed for 45 of 48 interventionists; the other three interventionists had data for nine observations. In Year 2, 31 of 34 interventionists completed initial certification; of these, 30 had scores for 10 observations and one had scores for nine observations. The other four Year 2 trainees were delayed and were not included in this summary of fidelity.

of 100 hours across both grades. Again, means varied by school. One school, Ravel Elementary, exceeded the 100-hour goal, and three additional schools exceeded 90 hours. Across all schools, approximately 5% of Cohort 1 students participating in the intervention met the 100-hour target. As seen in Exhibit 3.8, there was no systematic variation in hours across kindergarten winter LNF ranges (e.g., students with lower baseline scores receiving more intervention time or vice versa).

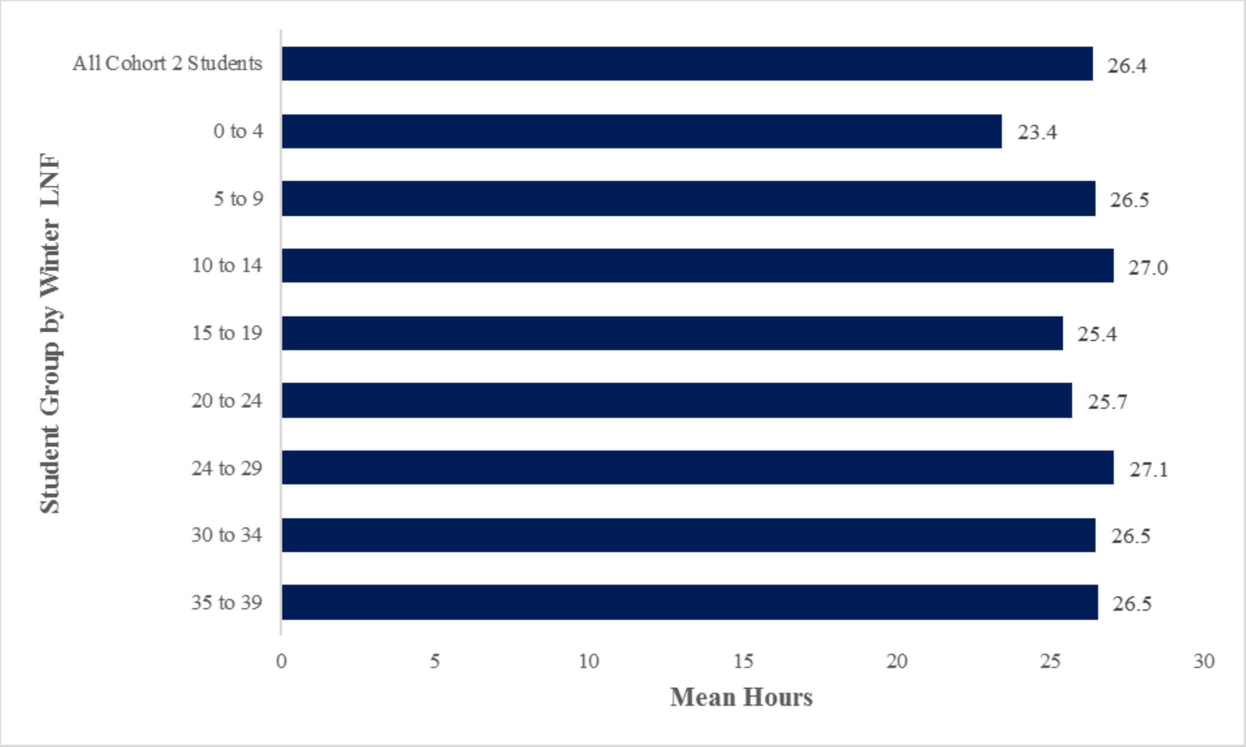
Exhibit 3.8. Cohort 1 Mean Intervention Hours per Student, Overall, and by Kindergarten Winter LNF Range



Note. Sample size = 398 (students who qualified for the intervention and participated through the end of the year).
Source: Pilot school intervention logs.

Cohort 2 students received 26 intervention hours on average (*SD* = 7.4) in kindergarten during the 2016–17 school year, falling a few hours below the 30-hour goal. This was higher than Cohort 1's kindergarten (Year 1) intervention time (*M* = 23 hours). As seen in Exhibit C7 in Appendix C, means varied greatly across schools, ranging from approximately 13 hours (Schubert Elementary) to 40 hours (Beethoven Elementary), with five schools meeting or exceeding the 30-hour target. Across all schools, approximately 41% of all Cohort 2 students participating in the intervention met the 30-hour target. This is a notable improvement over Year 1, when only 5% of participating Cohort 1 students met the target. As with Cohort 1, hours did not systematically vary across winter LNF ranges (see Exhibit 3.9).

Exhibit 3.9. Cohort 2 Mean Intervention Hours per Student, Overall, and by Winter LNF Range



Note. Sample size = 441 (students who qualified for the intervention and participated through the end of the year).
Source: Pilot school intervention logs.

In Year 2, the average group size was 2.9 for Cohort 1 and 3.0 for Cohort 2.[15] These averages met or were just below the target maximum of three per group. Averages varied by school (see Exhibit C8 in Appendix C). For Cohort 1, 13 of 21 schools had average group sizes of 3.0 or less; 19 of 21 had average group sizes of 3.4 or less (which would round to 3.0, suggesting the majority of sessions had three or fewer students). For Cohort 2, 13 of 21 schools had averages of 3.0 or less and 17 schools had averages of 3.4 or less. All school averages were below 4.0, with the exception of Vivaldi Elementary (Cohort 2 only).

[15] The Cohort 2 mean reflects 433 intervention students because group size information was not available for one interventionist.

Chapter 4. Results

Effectiveness of the Classroom Program (RQ 3), Main Analyses

As discussed above, the classroom program is designed to strengthen core instruction by providing classroom teachers with professional development aligned with the recommendations of the National Early Literacy Panel (2008), the National Reading Panel (2000), and other evidence-based approaches (Adams, 1990; Foorman et al., 2016; Kosanovich & Foorman, 2016; National Research Council, 1998; Shanahan et al., 2010). As the program is aimed at the classroom teacher and thus all students, regardless of baseline performance, the Pilot sample includes students who received both the classroom component and the MSL intervention (students with a kindergarten winter LNF of 39 or less) and those who received only the classroom component (students with a kindergarten winter LNF of 40 or higher). Analyses for this report show the effects of the classroom program on Cohort 1 students by the end of their kindergarten year (2015–16) and first grade year (2016–17), and on Cohort 2 students by the end of their kindergarten year (2016–17).

Cohort 1 Classroom Program Results

Controlling for school and baseline student characteristics, the study team did not find any statistically significant effects of the classroom program on spring DIBELS scores by the end of Cohort 1's kindergarten year. As shown in Exhibit 4.1, the effect of 2.7 points (*SE* = 1.4) on students' spring LNF approached, but did not meet, statistical significance (*ES* = 0.16, *p* = .063); the effects on NWF and PSF were statistically indistinguishable from zero.

Exhibit 4.1. End-of-Year Pilot and Matched Comparison Analyses, Cohort 1

Outcome	Pilot Group Mean	Comparison Group Mean	Estimated Difference	Standard Error	Effect Size	*p*-Value
Kindergarten (2015–16)						
LNF	57.6	54.9	2.7	1.4	0.16	.063
NWF-CLS	44.4	45.8	-1.4	2.1	-0.05	.512
PSF	54.8	55.2	-0.4	2.4	-0.03	.865
First Grade (2016–17)						
NWF-CLS	86.4	79.6	6.8*	2.8	0.18	.015
NWF-WWR	26.4	23.9	2.5*	1.1	0.17	.029
ORF-WC	64.3	67.3	-3.1	2.2	-0.09	.158
ORF-Accuracy	92.6	90.2	2.4	2.1	0.16	.262

Note. Kindergarten sample size = 2,735 students (1,591 Pilot and 1,144 Comparison); first grade sample size = 2,471–2,472 students (1,433 Pilot and 1,038–1,039 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

By the end of the first grade, however, the Cohort 1 Pilot students out-performed Comparison students on both NWF scores. Specifically, students in Pilot schools scored 6.8 points higher on the NWF correct letter sounds (NWF-CLS) score (*ES* = 0.18, *p* = .015), and 2.5 points higher on the NWF whole words read (NWF-WWR) score (*ES* = 0.18, *p* = .029).

Both kindergarten and first grade analysis results were robust to additional model specifications (e.g., adding additional school-level controls, running the analyses without covariates). Additional supplemental analyses are reported in the *Exploratory Analyses* section.

Cohort 2 Classroom Program Results

Cohort 2 Pilot students performed significantly better than the Comparison students on the LNF and NWF measures by the spring of their kindergarten year. The 4–5 point differences between the Pilot and Comparison students equates to an effect size of more than 0.2 student standard deviations. The difference in findings across the two cohorts is notable. Cohort 1 students, as shown in Exhibit 4.1, did not perform significantly better than their matched comparisons, suggesting that some factors of the classroom or intervention program changed in kindergarten between the first and second Pilot years (e.g., quality of implementation).

Exhibit 4.2. End-of-Year Pilot and Matched Comparison Analyses, Cohort 2

Outcome	Pilot Group Mean	Comparison Group Mean	Estimated Difference	Standard Error	Effect Size	*p*-Value
Kindergarten (2016–17)						
LNF	58.2	54.3	3.9*	1.5	0.23	.008
NWF-CLS	47.3	42.3	4.9*	1.5	0.21	.001
PSF	54.4	55.6	-1.3	1.8	-0.08	.487

Note. Sample size = 2,833–2,834 students (1,519 Pilot and 1,314–1,315 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

Effectiveness of Intervention (RQ 4), Main Analyses

The MSL intervention is aimed at providing additional support to students with lower levels of baseline literacy skills (as measured by kindergarten winter LNF). The primary RD analyses used a restricted sample comparing the 438 Cohort 1 Pilot students and the 320 Cohort 2 Pilot students who scored within 5 points below (winter LNF 35–39, qualifying for intervention) and 5 points above (winter LNF 40–44, not qualifying for intervention) the MSL intervention assignment cut score, controlling for the winter LNF score. As with the classroom component analyses, the analyses show the effects of the intervention on Cohort 1 students by the end of their kindergarten (2015–16) and first grade years (2016–17), and on Cohort 2 students by the end of their kindergarten year (2016–17).

Cohort 1 Intervention Results

Looking just at those students within 5 points of the cut score, the study team generally found positive but nonsignificant effects of the MSL intervention on Cohort 1's spring DIBELS scores by the end of their kindergarten and first grade school years (see Exhibit 4.3). The results were robust to most model specifications.

Exhibit 4.3. Cohort 1 Regression Discontinuity Analyses for Restricted Sample (LNF 35–44)

Outcome Variable	Estimated Effect	Standard Error	Effect Size	*p*-Value
Kindergarten (2015–16)				
LNF	1.6	2.0	0.15	.431
NWF-CLS	3.3	2.8	0.23	.229
PSF	1.5	2.2	0.12	.498
First Grade (2016–17)				
NWF-CLS	8.5	6.3	0.28	.175
NWF-WWR	2.9	2.5	0.24	.241
ORF-WC	-1.1	4.9	-0.05	.817
ORF-Accuracy	0.9	1.9	0.10	.609

Note. Kindergarten sample size = 431 students (186 intervention and 245 nonintervention); first grade sample size = 391 students (165 intervention and 226 nonintervention). The analyses were based on a two-level regression (students within schools), controlling for kindergarten winter LNF. The *p*-values for the estimated impact are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

Cohort 2 Intervention Results

Likewise, analyses of the effect of the MSL intervention on Cohort 2's spring DIBELS scores showed that students who scored within 5 points of the cut score performed similarly to each other, controlling for winter LNF (see Exhibit 4.4). These findings, too, were robust to most model specifications.

Exhibit 4.4. Cohort 2 Regression Discontinuity Analyses for Restricted Sample (LNF 35–44)

Outcome Variable	Estimated Effect	Standard Error	Effect Size	*p*-Value
Kindergarten (2016–17)				
LNF	-3.0	2.2	-0.29	.170
NWF-CLS	0.9	3.2	0.06	.781
PSF	2.5	2.4	0.22	.297

Note. Sample size = 316 students (133 intervention and 183 nonintervention). The analyses were based on a two-level regression (students within schools), controlling for kindergarten winter LNF. The *p*-values for the estimated effect are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

Exploratory Analyses

Although analyses of the classroom component in Year 1 (i.e., on Cohort 1's kindergarten scores) showed null results, both Cohorts 1 and 2 showed positive significant effects of the classroom component in Year 2. The main analyses of the MSL intervention for both cohorts in both years, however, suggested that the MSL intervention was not having a positive effect for students who were within 5 points of the cut score. Importantly, the null results say nothing about whether the MSL intervention affected students who qualified for the intervention but who were further away from the cut point at baseline (i.e., students with kindergarten winter LNF scores below 35). Likewise, the classroom component analyses did not show for whom the classroom component was effective. That is, analyses were complicated by Pilot treatment heterogeneity, because some students received both the classroom program and the intervention, and among intervention students, some received a variety of MSL programs and dosages. This section explores whether evidence exists that either the classroom program or the MSL intervention affected Cohort 1 or Cohort 2 students as intended.

Effects by MSL Intervention Qualification

As previously discussed, the analyses examining the effectiveness of the classroom program compared all Pilot students with kindergarten fall, winter, and spring DIBELS scores with all Comparison students with kindergarten fall, winter, and spring DIBELS scores.[16] The Pilot students included students who were eligible for the MSL intervention (e.g., 554 students, or 35% of the Cohort 1 Pilot sample) and those who were not (e.g., 1,039 students, or 65% of the Cohort 1 Pilot sample).[17] Because a large percentage of students (30%) participated in the MSL intervention, the classroom program analyses also picked up the partial effect of the MSL intervention. This offers an opportunity to examine the average treatment of the overall Pilot program, including both the classroom program and the MSL intervention, on all Pilot students.

To determine whether classroom program effects differed by students' eligibility to participate in the MSL intervention (i.e., whether or not there was an interaction effect for students receiving both the classroom program and the MSL intervention), the study team analyzed the effects of the classroom program on spring DIBELS for students with a kindergarten winter LNF score of 40 or above (i.e., nonintervention students, or students only exposed to the classroom component), and for students with a kindergarten winter LNF score of 39 or less (i.e., students who experienced the classroom program and *also* qualified for the MSL intervention). The study team then tested whether the estimates for the effect of the classroom program were different for these two groups (i.e., students who were and were not eligible for intervention). The following two subsections present the results separately for Cohort 1 and Cohort 2.

[16] Cohort 1 sample inclusion also requires first grade spring DIBELS scores.

[17] The sample size for the MSL intervention does not match exactly the sample size for the classroom component, as the sample for the classroom program analyses included only those students who (a) participated in DIBELS screening in the fall, winter, and spring, and (b) had data for all covariates included in the main model. The main difference between the sample for the classroom program and the sample for the MSL intervention is that the sample for the MSL intervention did not include a requirement that students participate in the DIBELS screening in the fall.

Cohort 1 Pilot Effects by MSL Intervention Qualification

Both kindergarten and first grade analysis results, shown in Exhibit 4.5 and 4.6, showed no evidence that the classroom program alone improved Cohort 1 students' preliteracy skills, as the mean differences between Pilot and Comparison nonintervention students remained nonsignificant, and in ORF in first grade, negative and significant. However, in both years, positive effects existed for students who qualified for the MSL intervention. Specifically, by the end of Cohort 1's kindergarten year, analyses showed that the students who qualified for the MSL intervention scored 4.8 points higher (*ES* = 0.29, *p* < .001) than Comparison students with winter LNF scores of 39 or less, controlling for student and school characteristics. By the end of Cohort 1's first grade year, students who qualified for the MSL intervention outperformed Comparison students with kindergarten winter LNF scores of 39 or lower on both NWF scores, scoring 11.7 points higher (*ES* = 0.32, *p* < .001) than Comparison students on NWF-CLS and 4.0 points higher (*ES* = 0.27, *p* = .002) than Comparison students on NWF-WWR.

Exhibit 4.5. Effects on DIBELS Scores in Spring of Kindergarten (2015–16) for Students Not Eligible for the MSL Intervention and Students Eligible for the MSL Intervention, Cohort 1

Outcome	Pilot Group Mean	Comparison Group Mean	Estimated Difference	Standard Error	Effect Size	*p*-Value
LNF						
Effect on nonintervention students	63.5	62.6	1.0	1.2	0.06	.416
Effect on intervention students	46.0	41.3	4.8*	1.3	0.29	.000
Difference in effects			3.8*	1.0	0.23	.000
NWF-CLS						
Effect on nonintervention students	52.0	55.0	-3.0	2.0	-0.11	.131
Effect on intervention students	30.3	29.4	0.9	2.2	0.03	.679
Difference in effects			4.0*	1.6	0.15	.017
PSF						
Effect on nonintervention students	57.8	58.5	-0.7	2.4	-0.05	.774
Effect on intervention students	49.0	49.3	-0.2	2.4	-0.02	.923
Difference in effects			0.4	1.0	0.03	.657

Note. Sample size = 2,735 students (1,591 Pilot and 1,144 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

Exhibit 4.6. Effects on DIBELS Scores in Spring of First Grade (2016–17) for Students Not Eligible for the MSL Intervention and Students Eligible for the MSL Intervention, Cohort 1

Outcome	Pilot Group Mean	Comparison Group Mean	Estimated Difference	Standard Error	Effect Size	*p*-Value
NWF-CLS						
Effect on nonintervention students	94.1	90.6	3.5	2.8	0.09	.213
Effect on intervention students	70.2	58.5	11.7*	3.2	0.32	.000
Difference in effects			8.2*	2.7	0.22	.002
NWF-WWR						
Effect on nonintervention students	29.5	28.0	1.5	1.2	0.10	.209
Effect on intervention students	19.7	15.8	4.0*	1.3	0.27	.002
Difference in effects			2.5*	1.1	0.17	.017
ORF-WC						
Effect on nonintervention students	74.9	80.4	-5.5*	2.2	-0.15	.014
Effect on intervention students	43.3	43.1	0.2	2.5	0.01	.941
Difference in effects			5.7*	2.2	0.16	.011
ORF-Accuracy						
Effect on nonintervention students	95.8	94.6	1.3	2.2	0.09	.564
Effect on intervention students	85.4	81.8	3.6	2.2	0.25	.105
Difference in effects			2.4*	1.0	0.16	.013

Note. Sample size = 2,471–2,472 students (1,433 Pilot and 1,038–1,039 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

These results for Cohort 1, along with the null results in the main kindergarten classroom component analyses and MSL intervention analyses, suggest that the MSL intervention had a positive effect on the preliteracy skills of students with the greatest need for support (i.e., students who not only qualified for the MSL intervention, but who had scores more than 5 points below the cut score). It is possible that the classroom component added to the effect as well, but this seems unlikely given that there was no indication that the classroom component, designed to support all students, affected nonintervention students' preliteracy skills. The pattern of the effects in Year 1 bolsters this finding, showing large positive effects on the NWF-CLS and

NWF-WWR measures for intervention students, and also showing significant differences in the effects between intervention and nonintervention students on all measures.

Cohort 2 Pilot Effects by MSL Intervention Qualification

The analyses for Cohort 2 showed a different pattern than those of Cohort 1. Specifically, the Pilot program had positive and significant effects on LNF and NWF-CLS for both nonintervention and intervention students. Nonintervention students scored 2.8 points higher (*ES* = 0.16, *p* = .018) on LNF than Comparison students with winter LNF scores of 40 or higher, while intervention students scored 3.5 points higher (*ES* = 0.20, *p* = .006) than Comparison students with winter LNF scores of 39 or lower. Likewise, both nonintervention and intervention students scored higher on NWF-CLS than Comparison students with similar winter LNF scores. Moreover, for both LNF and NWF-CLS scores, the effects of the Pilot program on intervention and nonintervention students were statistically indistinguishable from each other (e.g., the difference in the effects on LNF are just 0.7 points, *p* = .483).

Exhibit 4.7. Effects on DIBELS Scores in Spring of Kindergarten (2015–16) for Students Not Eligible for the MSL Intervention and Students Eligible for the MSL Intervention, Cohort 2

Outcome	Pilot Group Mean	Comparison Group Mean	Estimated Difference	Standard Error	Effect Size	*p*-Value
LNF						
Effect on nonintervention students	65.6	62.8	2.8*	1.2	0.16	.018
Effect on intervention students	45.3	41.9	3.5*	1.3	0.20	.006
Difference in effects			0.7	1.0	0.04	.483
NWF-CLS						
Effect on nonintervention students	56.2	51.6	4.6*	1.5	0.19	.002
Effect on intervention students	32.5	28.6	3.9*	1.7	0.16	.019
Difference in effects			-0.7	1.6	-0.03	.647
PSF						
Effect on nonintervention students	57.6	59.4	-1.8	1.7	-0.12	.294
Effect on intervention students	48.6	50.1	-1.5	1.8	-0.10	.401
Difference in effects			0.3	1.0	0.02	.759

Note. Sample size = 2,833–2,834 students (1,519 Pilot and 1,314–1,315 Comparison). The analyses were based on a two-level regression (students within schools), controlling for pairing blocks, free or reduced-price lunch, special education status, baseline DIBELS, and schools' Title I status. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data.

These analysis results suggest that the classroom component in Cohort 2 was primarily responsible for the effects on students' preliteracy skills. The MSL intervention may have added to the effect for students who were most in need, though these analyses did not provide any evidence of this.

Intervention Hours Analyses

Additional analyses examined the extent to which intervention hours through May 2017[18] predicted spring DIBELS scores, controlling for the kindergarten winter score (see Exhibit 4.8). Looking across all students who qualified for the MSL intervention, the analyses showed that an added hour of intervention time was associated with an increase in kindergarten LNF scores in Cohorts 1 and 2, an increase in NWF-CLS scores in Cohort 1, and an increase in PSF scores in Cohort 2. For example, an added hour of intervention was associated with an increase in the spring LNF score of 0.29 points in Cohort 1 and 0.12 points in Cohort 2. *Had students received the intended 30 hours of intervention, the RD analysis estimates likely would have been notably more positive for Cohorts 1 and 2 in their kindergarten year*. The relationship between hours in intervention and first grade DIBELS scores was notably smaller, however, with a significant but weak relationship between intervention hours and NWF-WWR.

Exhibit 4.8. Estimated Impact of an Additional Hour of Intervention Time on Spring DIBELS, Cohorts 1 and 2

Outcome Variable	Estimate	Standard Error	*p*-Value
Cohort 1			
Kindergarten (2015–16)			
LNF	0.29*	0.05	.000
NWF-CLS	0.19*	0.06	.001
PSF	0.12	0.07	.087
First Grade (2016–17)			
NWF-CLS	0.08	0.04	.067
NWF-WWR	0.04*	0.02	.044
ORF-WC	-0.03	0.03	.429
ORF-Accuracy	-0.02	0.02	.324
Cohort 2			
Kindergarten (2016–17)			
LNF	0.12*	0.05	.011
NWF-CLS	0.10	0.05	.059
PSF	0.15*	0.06	.017

Note. Kindergarten sample size = 546 students (all intervention); first grade sample size = 488 students (all intervention). The analyses were based on a two-level regression (students within schools), controlling for kindergarten winter LNF. The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*).
Source: District DIBELS data and intervention logs.

[18] June hours were excluded because they would not impact spring DIBELS scores collected in May.

Chapter 5. Summary of Key Findings and Limitations

Key Findings

This evaluation is guided by research questions regarding the implementation of training, the implementation of the classroom program, the implementation of the MSL intervention, the effectiveness of the classroom program, and the effectiveness of the MSL intervention. In Year 2, the Pilot was implemented with two cohorts: Cohort 1, now in first grade, and Cohort 2, the kindergarten class of 2016–17 (the second of three kindergarten cohorts in the study). PDE provided training to Pilot classroom teachers and interventionists. All schools conducted universal preliteracy screening in kindergarten, using DIBELS Next in the fall, winter, and spring. Winter LNF scores were used to assign students to the intervention (January 2016 scores for Cohort 1 and January 2017 scores for Cohort 2). Classroom fidelity data suggest that classroom teachers were implementing the programs and instructional practices from training to at least some extent, although there were no criteria for adequate implementation. Observational data, reflecting only a single lesson for each teacher, suggest that instruction focused primarily on phonics, followed by phonological awareness, with a stronger emphasis on phonics in first grade. Compass observation data showed that OG interventionists implemented the MSL intervention with high procedural fidelity. Intervention logs showed that all schools implemented the intervention, but that the target dosage was not met for most students. For Cohort 1, the target dosage of 100 hours across 2 years was met for only 5% of students; Cohort 1 students received approximately 76 hours of the intervention, on average. For Cohort 2, the target of 30 hours was met for 41% of students participating in the intervention; Cohort 2 students received approximately 26 hours, on average. This is an improvement over Year 1, when Cohort 1 received approximately 23 hours, on average.

The main analyses for the classroom program yielded significant findings ($p < .05$) on some measures for both cohorts. For Cohort 1, significant effects were seen for both spring first grade NWF scoring methods: NWF-CLS had an estimated difference of 6.8 points ($ES = 0.18$) and NWF-NWF had an estimated difference of 2.5 ($ES = 0.17$). For Cohort 2, significant effects were seen for two spring kindergarten measures: LNF had an estimated difference of 4.0 ($ES = 0.23$) and NWF-CLS had an estimated difference of 5.1 ($ES = 0.21$). The main analyses for the MSL intervention yielded no significant findings. However, classroom program analyses conducted separately for intervention and nonintervention students revealed that for Cohort 1, Pilot intervention students (but not nonintervention students) outperformed similar Comparison students on kindergarten spring LNF and NWF-CLS and on first grade spring NWF-CLS and NWF-WWR. These results for Cohort 1, along with the null results in the main kindergarten classroom component analyses and MSL intervention analyses, suggest that the MSL intervention had a positive effect on the preliteracy skills of students with the greatest need for support (i.e., students who not only qualified for the MSL intervention, but who had scores more than 5 points below the cut score). In contrast, for Cohort 2, both intervention and nonintervention Pilot students outperformed similar Comparison students. Taken together with the null Cohort 2 MSL analyses, these results suggest that the classroom component in Cohort 2 was primarily responsible for the effects on students' preliteracy skills. This may support the idea that classroom program implementation improved in Year 2. Additional analyses suggest that the MSL intervention effect was mediated by dosage; in kindergarten and to a lesser extent in first grade, students receiving more intervention time tended to have higher spring DIBELS scores when controlling for winter kindergarten LNF. Thus, more intervention time in Year 3 may yield stronger findings.

Limitations

The classroom program effectiveness evaluation design presents some limitations. The school-level matching did not account for unobserved variables (e.g., factors motivating the participating districts to be part of the Pilot), which may result in biased estimates of program effectiveness. Furthermore, the Comparison sample’s participation in another funded literacy initiative (which began 3 years before the Pilot) may result in an underestimation of Pilot program effects compared with typical schools (which may not use universal screening to inform core instruction and identify students to receive supplemental intervention). As previously mentioned, the classroom program analyses captured the effects of both the classroom component and any supplemental intervention. In Pilot schools, some students who qualified for the MSL intervention did not receive the intervention; in contrast, Comparison schools did not require parent consent for the intervention, so it may be assumed that a greater proportion of qualifying students received the intervention. A limitation of the RD design is that results are only generalizable near the cut point.

Data quality and availability are limitations as well. For example, classroom program fidelity data provided evidence of implementation but did not speak to whether or not that implementation was adequate. In addition, student-level attendance data were not available for Comparison schools, so that variable could not be used as a control in the classroom program effectiveness analyses.

Intervention dosage remained, on average, below targets for both cohorts, making it less likely the evaluation would detect significant effects. Treatment heterogeneity for the intervention is another limitation, as is the imperfect adherence to the cut score.

Future Reports

The final report will cover the third year of Pilot implementation. This will enable comparisons across three cohorts and expand the Pilot to second grade (Cohort 1). Other variables, such as special education status, also may be explored.

References

Adams, M. J. (1990). *Beginning to read: Thanking and learning about print.* Cambridge, MA: MIT Press.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*, 163–176. doi: 10.1177/0022219408326219

Cochran, W. G., & Rubin D. B. (1973). Controlling bias in observational studies: A review. *Sankhya*, *35*, 417–446.

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., . . ., & Wissel, S. (2016). *Foundational skills to support reading for understanding in kindergarten through 3rd grade* (NCEE 2016-4008). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from http://whatworks.ed.gov

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2013). *Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Next technical manual.* Eugene, OR: Dynamic Measurement Group.

Jacob, R., Zhu, P., Somers, M-A., & Bloom, M. (2012). *A practical guide to regression discontinuity*. New York, NY: MDRC. Retrieved from http://www.mdrc.org/sites/default/files/RDD%20Guide_Full%20rev%202016_0.pdf

Kosanovich, M., & Foorman, B. (2016). *Professional learning communities facilitator's guide for the What Works Clearinghouse practice guide: Foundational skills to support reading for understanding in kindergarten through 3rd grade* (REL 2016-227). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy. Retrieved from http://www.nifl.gov/earlychildhood/NELP/NELPreport.html

National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Reports of the subgroups*. Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health. Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/documents/report.pdf

National Research Council. (1998). *Preventing reading difficulties in young children.* Washington, DC: The National Academies Press. https://doi.org/10.17226/6023

Pennsylvania Department of Education (PDE). (2014a, September 9). *Dyslexia screening & early literacy intervention pilot program*. Retrieved from http://www.pattan.net/category/Projects/page/Dyslexia.html

Pennsylvania School Code of 1949, Dyslexia Screening and Early Literacy Intervention Pilot Program, Pub. L. 773, No. 69 (2014b). Retrieved from http://www.pattan.net/category/Resources/Misc.%20Materials/Browse/Single/?id=54060c460c1c44cb178b4574

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group by multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33–38.

Rubin, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics, 36,* 293–298.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://files.eric.ed.gov/fulltext/ED512029.pdf

Appendix A. Study Design

In recent years, randomized experiments have become the “gold standard” for evaluating educational interventions. When implemented properly, randomization guarantees that the treatment and control groups produced are equivalent in expectation at baseline, so any difference between the two groups after the start of treatment can be attributed to the effect of the treatment. For this reason, randomized experiments provide unbiased estimates of program impacts that are easy to understand and interpret. For a variety of reasons, however, it is not always practical or feasible to implement a randomized experiment, in which case a nonexperimental design must be used instead.[19]

When using a nonexperimental design, researchers estimate the impact of a program by selecting a comparison group that looks similar to the treatment group on observed characteristics, typically through matching methods. An important threat to the causal validity of such designs is selection bias: Differences in outcomes between the treatment and comparison groups may be because of preexisting or unobserved differences between the two groups, rather than the effect of the program being evaluated. An important challenge in the use of nonexperimental designs is to identify a comparison group that is equivalent to the treatment group in all ways except program participation.[20]

To test the effect of classroom teachers receiving the training and implementing the classroom program (RQ 3), Pilot schools were matched to similar schools in nonparticipating school districts that were not implementing the program but were collecting DIBELS data. The school-level matched design allowed researchers to test whether schools in which teachers received the training and implemented the program had different student outcomes than the matched Comparison schools.[21] The outcomes for this design can include DIBELS spring scores and referral information (e.g., percentage of students referred or identified to special education).

Multiple different methods can be used to match schools to create a comparison group, including propensity score matching and Mahalanobis distances approaches. Both matching approaches rely on observable data and, as a result, cannot remove biases caused by unobserved factors (such as factors related to districts participating in the Pilot). AIR selected the Mahalanobis distance approach because it does not involve statistical modeling; thus, it requires fewer assumptions and is more flexible. The Mahalanobis distance approach is particularly well suited to situations in which the number of units to be matched is small, as in this evaluation of the Pilot. The basic process for pair matching with unpaired data is to calculate Mahalanobis distances based on observed covariates. Mahalanobis metric matching is based on a measure of overall similarity (i.e., “Mahalanobis distance”) between two units with respect to a set of covariates. Mahalanobis distance is calculated on the basis of covariate differences between the

[19] In addition to providing unbiased evidence of effectiveness, randomized experiments are more efficient (requiring smaller sample sizes) and simple to analyze, reducing the cost of analytical work.

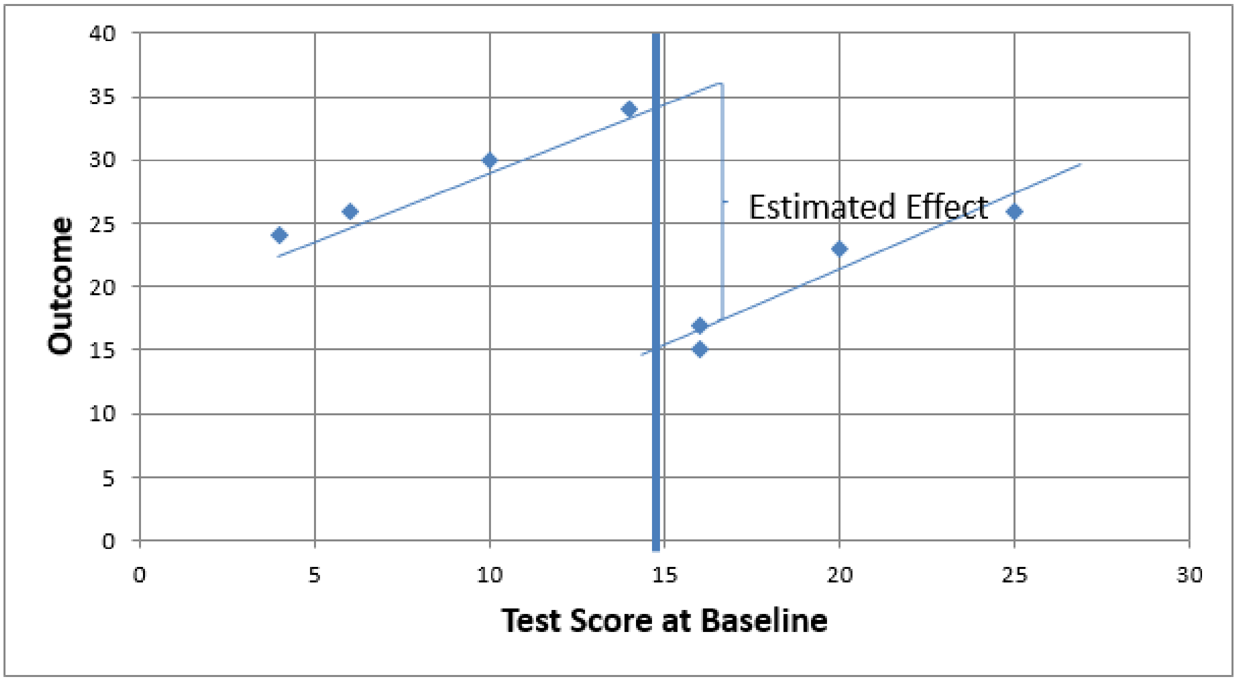
[20] Nonexperimental designs that do not incorporate a comparison group are significantly less rigorous and provide weaker evidence.

[21] The matched-school design had to use schools from other districts because all elementary schools from the Pilot districts were participating.

units (e.g., schools) and the sample variance-covariance matrix (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1985; Rubin, 1980).[22]

The effectiveness of the MSL intervention (RQ 4) was evaluated using an RD design. RD "analysis is a rigorous nonexperimental approach that can be used to estimate program impacts in situations in which candidates are selected for treatment based on whether their value for a numeric rating exceeds a designated threshold or cut point" (Jacob et al., 2012, p. iii). The rating variable may be any continuous variable measured before treatment, such as a pretest on the outcome variable (e.g., DIBELS score). An illustration of the RD approach is shown in Exhibit A1. The graph in the figure portrays a relationship that might exist between an outcome for candidates being considered for a prospective treatment and a rating (e.g., DIBELS cut score) used to prioritize candidates for that treatment. The vertical line in the center of the graph designates a cut point, below which candidates are assigned to the treatment and above which they are not assigned to the treatment. As can be seen, the relationship between outcomes and ratings is upward, sloping to the right, which indicates that the mean student test score increases as the pretest score increases. There is a sharp downward jump at the cut point in the relationship between outcomes and the pretest score, indicating that students who had low pretest scores and received the intervention benefited from the intervention.

Exhibit A1. Illustration of the RD Design With Intervention Effect



In this design, students within Pilot schools who were in need of additional supports were assigned to receive the MSL intervention if their DIBELS scores were below a preset cut point at the kindergarten winter screening. Possible outcomes of interest could include spring DIBELS

[22] Practically, the distance score is calculated by first transforming the data into standardized uncorrelated data and then computing the Euclidean distance for the transformed data. Therefore, the Mahalanobis distance is like a univariate *z* score.

scores, patterns or paths of students being screened out from the targeted services or screened in for more intensive supports, and reading test scores. To test whether the MSL intervention component of the program was effective, data for students who received the additional supports were compared with data from those students who did not, while acknowledging the clustered structure of the data (e.g., students nested within schools). For the RD design to work, it is imperative that students were assigned to receive additional support based only on the DIBELS assessment.

Appendix B. Matching to Establish Comparison Sample for RQ 3

Exhibit B1 shows the average school-level characteristics of the 21 Pilot and 21 matched Comparison schools originally selected based on summer 2015 matching analyses, conducted using historical data provided by PDE. In March 2016, two Comparison schools were dropped because they began implementing OG after the Pilot started; they were replaced with backup matches. In Year 2, another Comparison school closed; this was replaced with the school now attended by the majority of the students from the closed school. School-by-school matches are not shown to protect school confidentiality.

Exhibit B1. Pilot and Recommended Comparison Sample Characteristics (Historical Data)

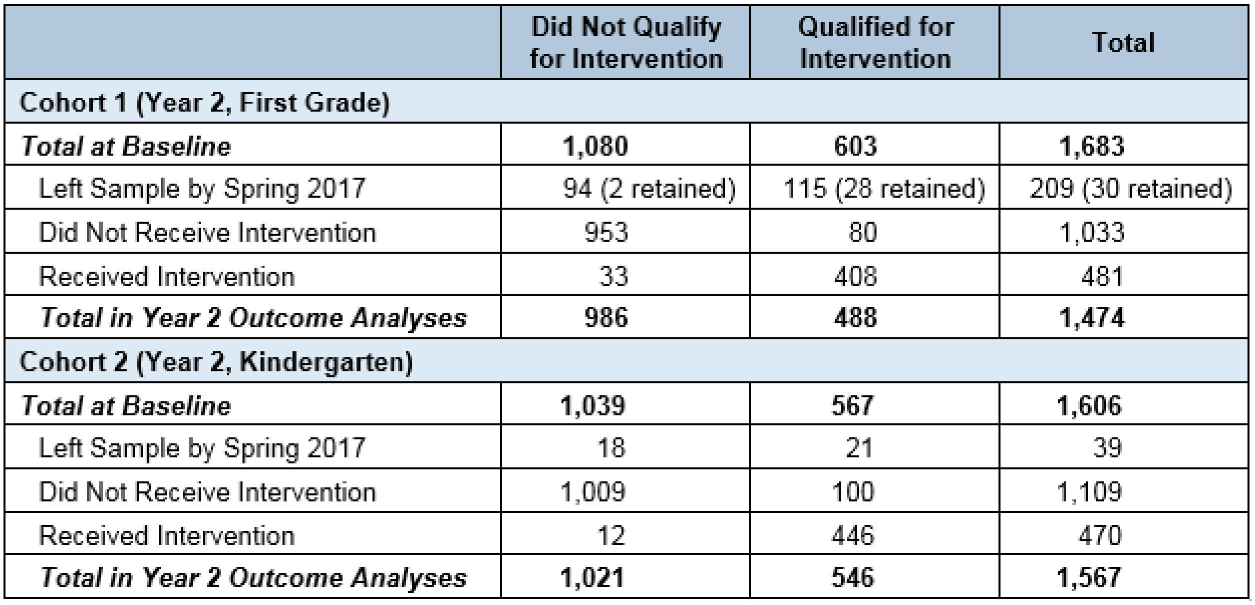
	Pilot Schools	Matched Comparison Schools	Difference	*p*-Value
Title I School (percentage)	81.0	90.5	-9.5	.390
Mean school total enrollment[a]	503.8	409.1	94.7	.073
Urbanicity (percentage)				
Urban	0.0	4.8	-4.8	.323
Suburban	42.9	42.9	0.0	1.000
Town	28.6	19.0	9.5	.481
Rural	28.6	33.3	-4.8	.746
DIBELS				
Kindergarten DIBELS score[a]	40.1	37.1	3.0	.215
First grade DIBELS score[a]	126.6	126.0	0.6	.873
Student Achievement				
Third grade scale score[a]	1339.3	1333.3	6.0	.570
Student Characteristics (percentage)				
Free or reduced-price lunch[a]	41.2	46.0	-4.7	.257
Limited English proficiency	1.1	1.6	-0.5	.439
Race/Ethnicity (percentage)				
African American[a]	3.4	4.9	-1.5	.354
Asian	1.8	1.4	0.5	.419
Hispanic[a]	5.5	5.7	-0.3	.896
Pacific Islander	0.1	0.1	0.0	.974
White	85.3	84.4	0.9	.818
Native American	0.3	0.1	0.1	.191
Multiracial	3.7	3.4	0.3	.765

[a] The variable was used in the Mahalanobis distance matching.

Appendix C. Supplementary Implementation Information

Exhibit C1 summarizes sample retention, attrition, and intervention participation according to whether or not students qualified for the intervention. Exhibit C2 summarizes Year 2 training attendance. Exhibit C3 summarizes observed instruction by category and district. Exhibit C4 summarizes self-reported frequency of instruction by category and district. Exhibit C5 summarizes self-reported program use. Exhibits C6 and C7 present the average intervention time per student, by Pilot school, for Cohorts 1 and 2, respectively. Exhibit C8 summarizes mean group size in Year 2 by cohort and school. The last section summarizes the Compass training requirements for OG Initial Level Certification and provides an overview of the three training levels.

Exhibit C1. Sample Attrition and Retention With Intervention Participation by Qualification Status



	Did Not Qualify for Intervention	Qualified for Intervention	Total
Cohort 1 (Year 2, First Grade)			
Total at Baseline	**1,080**	**603**	**1,683**
Left Sample by Spring 2017	94 (2 retained)	115 (28 retained)	209 (30 retained)
Did Not Receive Intervention	953	80	1,033
Received Intervention	33	408	481
Total in Year 2 Outcome Analyses	**986**	**488**	**1,474**
Cohort 2 (Year 2, Kindergarten)			
Total at Baseline	**1,039**	**567**	**1,606**
Left Sample by Spring 2017	18	21	39
Did Not Receive Intervention	1,009	100	1,109
Received Intervention	12	446	470
Total in Year 2 Outcome Analyses	**1,021**	**546**	**1,567**

Sources: Pilot DIBELS scores and intervention logs.

Exhibit C2. Year 2 Classroom Program Training Series Attendance by Session and Grade

District	N Teachers	"Big Five" Overview	Phonemic Awareness and Phonics	Oral Language and Vocabulary	Fluency and Comprehension	K Refresher
First Grade Teachers						
Dali School District	7	6	0	6	7	
Degas School District	4	4	4	5	4	
Kahlo Public Schools	16	17	17	18	16	
Michelangelo School District	6	6	6	6	6	
Monet School District	12	12	12	12	12	
Picasso Public Schools	2	2	1	2	2	
Pollock Public Schools	12	12	13	13	12	
Warhol School District	13	11	12	11	13	
All Districts	**81**	**72 (88.9%)**	**70 (86.4%)**	**65 (80.2%)**	**73 (90.1%)**	
Kindergarten Teachers (attendance optional)						
Dali School District	8	0	0	4	0	4
Degas School District	5	2	2	0	0	0
Kahlo Public Schools	20	2	2	2	13	0
Michelangelo School District	6	0	0	0	0	6
Monet School District	11	0	0	0	6	5
Picasso Public Schools	4	4	4	4	4	0
Pollock Public Schools	13	0	0	0	0	0
Warhol School District	12	0	1	0	0	9
All Districts	**79**	**8 (10.1%)**	**9 (11.4%)**	**10 (12.7%)**	**23 (29.1%)**	**24 (30.4%)**

Note. Sample size = 81 first grade teachers and 79 kindergarten teachers.
Source: Teacher-level attendance summary provided by PDE.

Exhibit C3. Mean Minutes of Observed Reading Instruction, by Category and District

District	Phonological Awareness	Phonics	Oral Language	Vocabulary	Fluency	Comprehension
Kindergarten Teachers (Cohort 1 Students)						
Dali School District	6.1	17.0	1.6	1.6	1.1	1.1
Degas School District	7.2	11.6	2.6	1.4	2.2	2.6
Kahlo Public Schools	7.4	11.8	2.8	0.7	0.2	1.2
Michelangelo School District	5.8	14.2	2.5	1.2	0.3	0.0
Monet School District	6.5	17.5	1.6	0.2	0.3	0.5
Picasso Public Schools	1.5	19.0	3.0	1.5	2.8	1.8
Pollock Public Schools	8.5	19.6	0.8	0.6	0.0	1.7
Warhol School District	6.8	16.0	1.1	1.3	1.5	0.5
First Grade Teachers (Cohort 2 Students)						
Dali School District	4.0	14.8	0.9	2.3	3.6	4.3
Degas School District	3.4	19.4	2.4	0.0	3.2	0.8
Kahlo Public Schools	1.2	23.1	0.6	0.7	0.5	0.1
Michelangelo School District	1.2	19.8	0.0	2.0	0.7	0.0
Monet School District	5.3	16.1	2.8	2.2	1.7	0.2
Picasso Public Schools	1.0	21.5	2.5	0.0	3.5	1.5
Pollock Public Schools	2.2	22.2	0.6	0.2	0.4	5.4
Warhol School District	3.1	23.5	0.3	0.2	0.9	0.1

Note. Sample size = 159 Pilot school kindergarten and first grade teachers.
Source: Fidelity observation means provided by PDE.

Exhibit C4. Mean Self-Reported Frequency of Reading Instruction by Category and District

District	N	Phonological Awareness	Phonics	Oral Language	Vocabulary	Fluency	Comprehension
Kindergarten Teachers (Cohort 2 Students)							
Dali School District	8	2.3	2.2	2.1	2.2	2.2	1.8
Degas School District	5	2.4	2.4	2.4	2.5	2.8	2.3
Kahlo Public Schools	19	2.5	2.5	2.4	2.1	2.5	2.3
Michelangelo School District	6	2.8	2.3	2.2	2.4	2.3	1.9
Monet School District	9	2.5	2.2	2.4	2.1	2.5	2.1
Picasso Public Schools	4	2.4	2.2	2.0	2.0	2.5	2.0
Pollock Public Schools	11	2.4	2.3	2.1	2.1	2.4	2.1
Warhol School District	8	2.4	2.2	2.0	1.9	2.3	1.9
First Grade Teachers (Cohort 1 Students)							
Dali School District	8	2.4	1.9	2.1	1.9	1.7	1.9
Degas School District	5	2.5	2.9	2.4	2.8	2.9	2.4
Kahlo Public Schools	18	2.5	2.5	2.5	2.5	2.6	2.5
Michelangelo School District	6	2.5	2.3	2.2	2.2	2.1	2.0
Monet School District	12	2.1	1.9	1.6	1.6	2.1	1.4
Picasso Public Schools	2	2.7	2.6	2.5	2.8	2.8	2.7
Pollock Public Schools	12	2.5	2.5	2.3	2.1	2.5	2.3
Warhol School District	10	2.6	2.6	2.5	2.4	2.7	2.4

Note. Sample size = 73 first grade teachers and 70 kindergarten teachers. Means reflect a 0–3 scale (described in text).
Source: Fidelity self-report data.

Exhibit C5. Self-Reported Frequency of Program Use, by District (Percentage of Teachers Using the Program "Some" or "A lot")

District	N	Fundations	Accessing the Code	Neuhaus	OG
Kindergarten Teachers (Cohort 2 Students)					
Dali School District	8	100%	25%	88%	75%
Degas School District	5	0%	0%	100%	100%
Kahlo Public Schools	19	32%	100%	84%	53%
Michelangelo School District	6	0%	50%	100%	50%
Monet School District	9	100%	11%	78%	33%
Picasso Public Schools	4	100%	100%	0%	0%
Pollock Public Schools	11	100%	64%	100%	64%
Warhol School District	8	38%	100%	63%	75%
First Grade Teachers (Cohort 1 Students)					
Dali School District	8	100%	63%	0%	0%
Degas School District	5	0%	0%	100%	100%
Kahlo Public Schools	18	0%	100%	6%	6%
Michelangelo School District	6	33%	100%	17%	0%
Monet School District	12	58%	17%	33%	67%
Picasso Public Schools	2	100%	0%	0%	0%
Pollock Public Schools	12	50%	100%	8%	33%
Warhol School District	10	20%	100%	40%	70%

Note. Sample size = 73 first grade teachers and 70 kindergarten teachers.
Source: Fidelity self-report data.

Exhibit C6. Cohort 1 Mean Total 2016–17 Intervention Hours and Mean Grand Total (Across 2 Years) Intervention Hours per Student, by School

District	School	2016–17 Mean Hours	2016–17 *SD*	Grand Total Mean Hours	Grand Total *SD*
Dali School District	Bach Elementary	58.3	3.4	85.4	3.8
Dali School District	Mozart Elementary	67.0	4.5	95.3	5.7
Degas School District	Verdi Elementary	44.1	11.1	72.7	11.3
Kahlo Public Schools	Copland Elementary	55.1	5.9	80.1	6.4
Kahlo Public Schools	Gershwin Elementary	56.5	15.7	84.5	15.6
Kahlo Public Schools	Ravel Elementary	73.4	2.4	101.2	3.7
Kahlo Public Schools	Schumann Elementary	57.6	18.5	76.4	18.5
Kahlo Public Schools	Vivaldi Elementary	62.9	21.1	78.5	22.5
Michelangelo School District	Debussy Elementary	36.6	4.4	57.8	8.9
Monet School District	Liszt Elementary	65.2	3.7	90.2	3.3
Monet School District	Mendelssohn Elementary	61.6	5.4	84.5	6.1
Monet School District	Strauss Elementary	49.9	10.7	68.9	11.7
Picasso Public Schools	Beethoven Elementary	64.6	23.1	92.6	32.1
Pollock Public Schools	Chopin Elementary	42.2	2.3	64.3	3.7
Pollock Public Schools	Haydn Elementary	50.9	12.4	70.3	11.3
Pollock Public Schools	Stravinsky Elementary	44.9	8.9	62.7	11.8
Warhol School District	Brahms Elementary	49.0	2.4	70.3	3.5
Warhol School District	Handel Elementary	63.5	18.8	82.4	20.4
Warhol School District	Schubert Elementary	50.9	7.2	71.5	8.0
Warhol School District	Tchaikovsky Elementary	56.2	5.7	72.5	8.1
Warhol School District	Wagner Elementary	55.0	4.4	80.1	8.0
All Districts	**All Schools**	**53.5**	**14.0**	**76.4**	**15.8**

Note. Sample size = 408 (students who qualified for and participated in the intervention, and remained in the school through the end of 2016-17).
Source: Pilot school intervention logs.

Exhibit C7. Cohort 2 Mean Total Intervention Hours per Student, by School

District	School	2016–17 Mean Hours	2016–17 *SD*
Dali School District	Bach Elementary	29.4	6.1
Dali School District	Mozart Elementary	27.8	2.3
Degas School District	Verdi Elementary	31.1	2.0
Kahlo Public Schools	Copland Elementary	29.5	2.5
Kahlo Public Schools	Gershwin Elementary	31.5	1.4
Kahlo Public Schools	Ravel Elementary	26.7	0.9
Kahlo Public Schools	Schumann Elementary	27.8	7.0
Kahlo Public Schools	Vivaldi Elementary	20.4	5.0
Michelangelo School District	Debussy Elementary	34.2	2.1
Monet School District	Liszt Elementary	28.4	1.1
Monet School District	Mendelssohn Elementary	24.4	2.1
Monet School District	Strauss Elementary	22.3	1.4
Picasso Public Schools	Beethoven Elementary	40.3	1.0
Pollock Public Schools	Chopin Elementary	24.3	2.6
Pollock Public Schools	Haydn Elementary	17.1	4.0
Pollock Public Schools	Stravinsky Elementary	21.7	4.5
Warhol School District	Brahms Elementary	19.2	2.7
Warhol School District	Handel Elementary	31.1	3.3
Warhol School District	Schubert Elementary	13.2	6.8
Warhol School District	Tchaikovsky Elementary	18.9	7.3
Warhol School District	Wagner Elementary	26.1	8.8
All Districts	**All Schools**	**26.4**	**7.4**

Note. Sample Size = 446 (students who qualified for and participated in the intervention, and remained in the school through the end of 2016-17).
Source: Pilot school intervention logs.

Exhibit C8. Mean Group Size in Year 2, by School

District	School	Cohort 1 Mean	Cohort 2 Mean
Dali School District	Bach Elementary	2.9	2.9
Dali School District	Mozart Elementary	2.7	2.9
Degas School District	Verdi Elementary	3.0	2.7
Kahlo Public Schools	Copland Elementary	3.2	3.4
Kahlo Public Schools	Gershwin Elementary	2.7	2.8
Kahlo Public Schools	Ravel Elementary	2.4	1.4
Kahlo Public Schools	Schumann Elementary	3.5	3.7
Kahlo Public Schools	Vivaldi Elementary	3.8	4.0
Michelangelo School District	Debussy Elementary	2.8	3.1
Monet School District	Liszt Elementary	3.4	2.9
Monet School District	Mendelssohn Elementary	3.4	3.0
Monet School District	Strauss Elementary	2.6	2.7
Picasso Public Schools	Beethoven Elementary	2.7	2.3
Pollock Public Schools	Chopin Elementary	3.2	3.3
Pollock Public Schools	Haydn Elementary	3.4	3.8
Pollock Public Schools	Stravinsky Elementary	2.6	3.1
Warhol School District	Brahms Elementary	2.4	3.2
Warhol School District	Handel Elementary	2.6	2.9
Warhol School District	Schubert Elementary	2.7	2.9
Warhol School District	Tchaikovsky Elementary	2.8	2.8
Warhol School District	Wagner Elementary	2.8	2.9
All Districts	**All Schools**	**2.9**	**3.0**

Note. Cohort 1 Sample Size = 451; Cohort 2 Sample Size = 450 (students who qualified for and participated in the intervention); group size not available for all sessions.
Source: Pilot school intervention logs.

Compass Reading Center's Initial Level Certification Orton-Gillingham Training Requirements

The following summarizes the OG training:

- Trainee receives 50 hours of lectures:
  - History of OG, phonological awareness, multisensory instruction, guided discovery teaching, decoding, encoding, morphology, lesson planning, and scope and sequence
- Trainee observes five sessions (5 hours) of an experienced OG tutor in session with students:
  - Live or video
- Trainee receives 10 observations with written feedback from trainer or supervisor. Must maintain an 85% average.
- Trainee completes three one-page book summaries on the following texts:
  - *Overcoming Dyslexia*, by Sally Shaywitz
  - *Unlocking Literacy: Effective Decoding and Spelling Instruction*, by Marcia Henry
  - *About Dyslexia: Unraveling the Myth*, by Priscilla Vail
- Trainee completes 12 one-page chapter outlines from *Multisensory Teaching of Basic Language Skills, Third Edition*, by Judith Birsh.
- Trainee takes four quizzes, maintaining an 85% average.
- Trainee completes one take-home final exam.
- Trainee completes progress narrative report at the end of 50 sessions for each student.
- Documentation of all training requirements must be maintained and submitted to Compass Reading Center's Trainer for certification.

Comparison of Compass Reading Center's Three Levels of Training

	Year 1 **Level 1**	**Year 2** **Level 2**	**Year 3** **Level 3**
Requirements	• 50 hours of lectures • 100 hours of supervised practicum • 10 observations/ feedback • Quizzes • Reading assignments/ reports	• 15 hours of lectures • 100 hours of supervised practicum • 4 observations/ feedback	• 36 hours of training including shadowing trainer, observations, reviewing lesson plans, grading reading assignments/quizzes, and administering assessments • 200 hours of practicum • (No observations)
Compass Levels	Initial Level Certificate of Completion	Intermediate Level Compass Certificate	Supervisor Level Compass Certificate
Compass Accreditation Levels	• Completion of IMSLEC[23] Teacher Level accredited course • Eligibility to sit for the Alliance Exam to become *Certified Academic Language Practitioner (CALP)* http://www.allianceaccreditation.org/standards.asp		
IMSLEC Level	Teacher name on website http://www.altaread.org/members-directory.asp?action=search1		

[23] The International Multisensory Structured Language Education Council. Accreditation standards found at https://www.imslec.org/standards.asp?_sm_au_=iMVF6qLsZVFk6n4B

Appendix D. Comparisons of Analysis Samples

Exhibit D1 summarizes the sample characteristics of the full intervention analysis sample, comparing students who did and did not qualify for intervention. Exhibit D2 summarizes the sample characteristics of the restricted sample used for the main RD analyses.

Exhibit D1. MSL Intervention Baseline (Winter of Kindergarten) Sample Characteristics, Full Sample, by Cohort

	Intervention Mean	Nonintervention Mean	Estimated Difference	p-Value
Cohort 1 Student Characteristics (percentage)				
Female	45.3	52.0	-6.8*	.008
White	78.3	84.6	-6.3*	.001
Free or reduced-price lunch	42.9	27.2	15.7*	.000
Special education	15.2	5.6	9.6*	.000
Cohort 2 Student Characteristics (percentage)				
Female	41.9	51.5	-9.6*	.000
White	79.4	85.9	-6.5*	.001
Free or reduced-price lunch	49.8	31.9	17.9*	.000
Special education	14.7	5.3	9.4*	.000

Note. Cohort 1 sample size = 1,682–1,683 students (602–603 intervention and 1,080 nonintervention students); Cohort 2 sample size = 1,567–1,568 students (549 intervention and 1,018–1,019 nonintervention students). The analyses were based on a two-level regression (students within schools). The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the *p* < .05 level is indicated by an asterisk (*). Intervention students had a winter LNF score of 39 or less. Nonintervention students had a winter LNF score of 40 or higher.
Source: District demographic data.

Exhibit D2. MSL Intervention Baseline (Winter of Kindergarten) Sample Characteristics, Main RD Analysis Sample (Winter LNF 35–44), by Cohort

Variable	Intervention Mean	Nonintervention Mean	Estimated Difference	*p*-Value
Cohort 1				
Student Characteristics (percentage)				
Female	47.1	52.0	-5.0	.305
White	79.4	84.6	-5.3	.159
Free or reduced-price lunch	27.9	27.2	0.6	.882
Special education	5.3	5.6	-0.3	.904
Diagnostic Assessments				
PPVT	107.2	109.7	-2.5	.060
CTOPP-2 Blending	9.8	10.0	-0.2	.426
CTOPP-2 Elision	9.4	9.5	-0.1	.767
CTOPP-2 Rapid Color Naming	9.8	10.3	-0.6*	.029
CTOPP-2 Rapid Object Naming	9.5	10.6	-1.1*	.000
Cohort 2				
Student Characteristics (percentage)				
Female	44.6	48.6	-4.0	.481
White	80.3	91.4	-11.1*	.003
Free or reduced-price lunch	45.4	33.0	12.4*	.022
Special education	4.0	9.2	-5.2	.068
Diagnostic Assessments				
PPVT	108.3	107.9	0.4	.832
CTOPP-2 Blending	10.1	10.1	0.0	.893
CTOPP-2 Elision	8.9	9.5	-0.6	.056
CTOPP-2 Rapid Color Naming	10.2	10.5	-0.3	.631
CTOPP-2 Rapid Object Naming	9.7	10.1	-0.4	.309

Note. Cohort 1 sample size for student characteristics = 437–438 students (189–190 intervention and 248 nonintervention students); Cohort 1 sample size for student diagnostic assessments = 343–362 students (152–164 intervention and 191–198 nonintervention students). Cohort 2 sample size for student characteristics = 315 students (130 intervention and 185 nonintervention students); Cohort 2 sample size for student diagnostic assessments = 263–264 students (109–110 intervention and 154 nonintervention students). The analyses were based on a two-level regression (students within schools). The *p*-values for the estimated difference are based on *t* tests. Two-tailed statistical significance at the $p < .05$ level is indicated by an asterisk (*). Intervention students had a winter LNF score of 39 or less. Nonintervention students had a winter LNF score of 40 or higher.
Source: District demographic and school diagnostic data.

Appendix E. Sample Parent Notification and Opt-Out Template Provided by PDE to Pilot Districts

SCHOOL DISTRICT LOGO

Dyslexia Screening and Early Literacy Pilot Program
Parent Notification/Opt Out of Diagnostic Assessments and Intervention

Date

Dear Parent or Guardian,

As part of **School District's Name** participation in the Dyslexia Screening and Early Literacy Pilot Program, your child is eligible to receive reading intervention services. Through early literacy screenings, we've learned that your child is likely to benefit from additional reading instruction. Reading intervention beginning in Kindergarten is optimal for achieving the best reading outcomes in the future.

We are including information and evidence-based resources for parents/guardians including:

- Dyslexia Screening and Early Literacy Pilot Program Notification/Opt Out
- International Dyslexia Association (IDA) Handbook for Parents
- Brochure-Dyslexia and Early Literacy Intervention Parent Fact Sheet
- Learning to Read – A Fact Sheet for Parents
- Basic Early Reading Skills
- Other resources found at http://tinyurl.com/PADyslexia4Parents

Many school districts already provide intervention to students who need additional reading support as part of their daily instructional practice. Because of our district's participation in the Pilot Program, a parent/guardian can decide to opt their child out of the program. Participation includes taking 2 brief diagnostic assessments to pinpoint your child's reading strengths and needs. It also qualifies your child for additional Multi-Sensory Structured Language intervention. If you **DO NOT** want to have your child participate in the program, please indicate that below and return this signed document to your child's teacher.

Please know that participation in this pilot program does not prevent parents/guardians from requesting an evaluation for special education at any time. If you are interested in requesting an evaluation, please contact ____ at your school district.

Within your district, parent liaisons are available to answer any of your questions or concerns. Your parent liaison's contact information is below:

Parent Liaison Name:
Email Address:
Phone Number

We are privileged to provide this opportunity for additional reading instruction to your child. Please let us know if you have any questions or concerns.

Sincerely,

Name, Position
School District Information

SCHOOL DISTRICT LOGO

Dyslexia Screening and Early Literacy Pilot Program

Parent Notification/Opt Out of Diagnostic Assessments and Intervention

If you would like your child to participate in the Dyslexia Screening and Early Literacy Pilot Program Intervention**, no further action is needed on your part.**

Check here and return this form **ONLY** if you **DO NOT** want your child to receive diagnostic assessments and participate in the Dyslexia Screening and Early Literacy Pilot Program Intervention _________

Child's Name ____

Parent/Guardian Signature___________________________ Date__________________

Appendix F. Technical Information

Statistical Analyses

Descriptive Analyses

Researchers conducted descriptive analyses regarding the level (e.g., training dosage) and fidelity of the implementation (e.g., in classrooms; by interventionists) in order to assess the level and variation in implementation quality, on the basis of implementation data provided by PDE. In addition, researchers conducted descriptive analyses regarding the students, teachers, and schools participating in the study, as well as the referral patterns of students over time (e.g., how many students were screened to receive additional supports, how many students screened out of the targeted services because of not needing the services anymore, how many students exited targeted services because of being referred to more intensive supports).

Impact Analyses

RQ 3. To address RQ 3 concerning the impact of the classroom program, the analysis team drew on the data collected by the districts or PDE (e.g., DIBELS, information about identification for special education) using a two-level hierarchical linear model (students nested within schools). The two-level model acknowledged the clustering of children within schools. The analysis team modeled student outcomes as a function of students' baseline DIBELS scores and intervention status. The model predicted child outcomes as a function of individual characteristics, such as pretest and child characteristics (e.g., gender) and school-level factors, including the treatment status, averaged baseline DIBELS scores, and blocking variables at the school level. The model for testing the impact of the Pilot classroom program was as follows:

$$Y_{ij} = \beta_{00} + \beta_{10}*(\text{Baseline\_LNF})_{ij} + + \beta_{20}*(\text{Baseline\_FSF})_{ij} + \beta_{30}*(\text{Student \_Characteristics})_{ij} + \beta_{01}*(\text{CLASS\_TRT})_{j} + \beta_{02}(\text{Title\_I})_{\text{j}} + \beta_{03}(\text{Block})_{\text{j}} + e_{ij} + r_{0j}$$

where

- Y_{ij} is the student outcome for student *i* in school *j*
- CLASS_TRT is an indicator variable that takes a value of 1 for a school implementing the Pilot program and 0 for a business as usual school (i.e., a Comparison school)
- Baseline_LNF$_{ij}$ is the measure of the fall LNF DIBELS for student *i* in school *j*, grand-mean centered
- Baseline_FSF$_{ij}$ is the measure of the fall FSF DIBELS for student *i* in school *j*, grand-mean centered
- Student _Characteristics$_{ij}$ is a vector of characteristics for student *i* in school *j*, grand-mean centered
- Title_I$_{\text{j}}$ is a school-level indicator variable identifying Title I schools taking a value of 1 if the school is a Title I school, and 0 if the school is not a Title I school

- Block_j is a school-level indicator variable identifying matched pairs taking a value of 1 if the school is part of the match and 0 if the school is not part of the match
- β_{00} is the average student outcome across all schools (i.e., grand mean)
- β_{10} is the average effect of the baseline LNF DIBELS score
- β_{20} is the average effect of the baseline FSF DIBELS score
- β_{30} is the average effect of baseline measures on the student outcome across all schools
- β_{01} is the average treatment effect across all schools
- β_{02} is the average effect of Title I Status across all schools
- β_{03} is the average effect of block indicator variable across all schools
- e_{ij} is a random error associated with student i in school j, $e_{ij} \sim N\ (0, \sigma 2)$
- r_{0j} is a random error associated with school average student outcome, $r_{0j} \sim N\ (0, \tau 00)$

Of primary interest is β_{01}, which represents the Pilot program's main effect on the outcome across all schools. A statistically significant positive value of β_{01} supports the hypothesis that students who received the Pilot classroom program would demonstrate better achievement outcomes than their counterparts who received the business-as-usual supports.

RQ 4. To answer RQ 4 using RD design, researchers conducted the following sequence of analyses to estimate the effect of the Pilot MSL intervention on those students who were below the agreed DIBELS cut point.

1. Graphical analyses: Graph the outcome versus the rating variable. Visually inspect the graph to assess whether there is a discontinuity at the cut point.
2. Conduct a parametric estimation of the program's effect and conduct nonparametric analyses as sensitivity tests.[24]
3. Unless evidence strongly suggests otherwise, use the simplest model possible to conduct analyses. Use more complex models as sensitivity checks only.

While conducting the parametric estimation of the program's effect, researchers:

1. Selected the appropriate functional form for the regression estimation, starting from a simple linear regression and adding higher order polynomials and interaction terms to it.
2. Used the *F* test approach to eliminate overly restrictive model specifications; in general, used the simplest functional form possible, unless the test results clearly indicated otherwise.

[24] The parametric approach uses all data and a regression-based technique to estimate the program's effect. The nonparametric approach uses a *localized regression* approach and includes only data that are close to the predetermined cut-off point to estimate the effect of the program. Which approach is more appropriate depends on the distribution of the data; researchers adjust the analytical strategy accordingly.

3. Added baseline characteristics (determined prior to the treatment) to the regression to improve precision.
4. Checked the robustness of the findings by "trimming" data points at the tails of the rating distribution.

The simplest linear estimation model for the RD modeled student outcomes as a function of students' baseline values (the rating variable) of outcomes and intervention status. Researchers used a two-level hierarchical model, with children nested within schools. The difference compared with RQ 3 was the level of the treatment indicator. In RQ 3, schools were either implementing the Pilot program or were Comparison schools, whereas in RQ 4, only students in schools that implemented the Pilot program were assigned to either receive the additional component of the Pilot program (more personalized supports delivered by an interventionist) or receive supports typically provided by the school or district based on their scores on the DIBELS test. The model for testing the impact of the targeted component of the Pilot program was as follows:

$$Y_{ij} = \beta_{00} + \beta_{10}*(\text{MSL\_TRT})_{ij} + \beta_{20}*(\text{Baseline\_LNF})_{ij} + e_{ij} + r_{0j} + r_{1j}*(\text{Pilot\_TRT})_{ij}$$

where

- Y_{ij} is the student outcome for student *i* in school *j*
- MSL_TRT is the indicator of whether student *i* in school *j* is assigned to receive targeted Pilot program supports delivered by an interventionist, group-mean centered
- Baseline_LNF$_{ij}$ is the measure of the kindergarten winter LNF DIBELS for student *i* in school *j*, grand-mean centered
- e_{ij} is a random error associated with student *i* in school *j*, $e_{ij} \sim N\ (0, \sigma 2)$
- β_{00} is the average student outcome across all schools (i.e., grand mean)
- r_{0j} is a random error associated with school average student outcome, $r_{0j} \sim N\ (0, \tau 00)$
- β_{10} is the average treatment effect across all schools
- r_{1j} is a random error associated with school *j* on the treatment effect, $r_{1j} \sim N\ (0, \tau 10)$
- β_{20} is the average effect of baseline measures on the student outcome across all schools

Of primary interest is β_{10}, which represents the Pilot program's main effect on the outcome across all schools. A statistically significant positive value of β_{10} supports the hypothesis that students who received the targeted component of the Pilot program would demonstrate higher achievement outcomes than their counterparts who received the business-as-usual supports.

Effects of Pilot Program by MSL Qualification. In order to estimate the effects of the pilot program by MSL qualification, researchers used a model similar to that used to answer RQ 3. The analysis team modeled student outcomes as a function of students' baseline DIBELS, school characteristics, and pilot school status, and also included an indicator for MSL qualification, and an interaction term on pilot school status and MSL qualification. The model was a two-level hierarchical model, with children nested within schools. The two-level model acknowledged the clustering of children within schools. The model predicted child outcomes as a function of individual characteristics, such as pretest and child characteristics (e.g., gender) and school-level factors, including the treatment status, averaged baseline DIBELS scores, and blocking variables at the school level. The model for testing the impact of the overall Pilot program by MSL qualification was as follows:

$$Y_{ij} = \beta_{00} + \beta_{10}\text{*(MSL\_TRT)} + \beta_{20}\text{*(MSL\_PILOT\_TRT)} + \beta_{30}\text{*(Baseline\_LNF)}_{ij} + \beta_{40}\text{*(Baseline\_FSF)}_{ij} + \beta_{50}\text{*(Student \_Characteristics)}_{ij} + \beta_{01}\text{*(Pilot\_TRT)}_{j} + \beta_{02}\text{(Title\_I)}_{\text{j}} + \beta_{03}\text{(Block)}_{\text{j}} + e_{ij} + r_{0j}$$

where

- Y_{ij} is the student outcome for student *i* in school *j*
- MSL_TRT is the indicator of whether student *i* in school *j* is eligible to receive targeted MSL intervention program delivered by an interventionist, group-mean centered
- MSL_PILOT_TRT is the indicator of whether student *i* in school *j* is eligible to receive targeted MSL intervention program delivered by an interventionist and is in a school implementing the Pilot program, group-mean centered
- Pilot_TRT is an indicator variable that takes a value of 1 for a school implementing the Pilot program and 0 for a business as usual school (i.e., a Comparison school)
- Baseline_LNF$_{ij}$ is the measure of the fall LNF DIBELS for student *i* in school *j*, grand-mean centered
- Baseline_FSF$_{ij}$ is the measure of the fall FSF DIBELS for student *i* in school *j*, grand-mean centered
- Student _Characteristics$_{ij}$ is a vector of characteristics for student *i* in school *j*, grand-mean centered
- Title_I$_{\text{j}}$ is a school-level indicator variable identifying Title I schools taking a value of 1 if the school is a Title I school, and 0 if the school is not a Title I school
- Block$_j$ is a school-level indicator variable identifying matched pairs taking a value of 1 if the school is part of the match and 0 if the school is not part of the match
- β_{00} is the average student outcome across all schools (i.e., grand mean)
- β_{10} is the difference between students who would have qualified for the MSL and those who would not have, in the Comparison schools
- β_{20} is the difference in the treatment effect for students who qualified for the MSL intervention and those who did not
- β_{30} is the average effect of the baseline LNF DIBELS score
- β_{40} is the average effect of the baseline FSF DIBELS score
- β_{50} is the average effect of baseline measures on the student outcome across all schools

- β_{01} is the average treatment for students who did not qualify for the MSL intervention effect across all schools
- β_{02} is the average effect of Title I Status across all schools
- β_{03} is the average effect of block indicator variable across all schools
- e_{ij} is a random error associated with student *i* in school *j*, $e_{ij} \sim N\ (0, \sigma2)$
- r_{0j} is a random error associated with school average student outcome, $r_{0j} \sim N\ (0, \tau00)$

Of primary interest is β_{01}, which represents the Pilot program's main effect on the outcome across all schools, for students who did not qualify for the intervention, and β_{20}, which represents the difference in the treatment effects between students who qualified for the intervention and those who did not. A statistically significant value of β_{01} supports the hypothesis that students who did not qualify for the intervention were affected by the classroom program, while a statistically significant value of β_{01+} β_{20} supports the hypothesis that students who qualified for the MSL intervention were effected by a combination of the classroom program and the MSL intervention. The estimate of β_{20} alone is the difference between the effects of the pilot program by MSL qualification.

Power Calculations

We computed the minimum detectable effect size (MDES) based on the actual analysis sample and impact results for each spring DIBELS outcome for the primary classroom program analysis model, and for the primary RD model. The MDESs are shown in Exhibit F1.

Exhibit F1. Realized Minimum Detectable Effect Sizes for Primary Classroom Component Regression Discontinuity Analyses

Outcome	MDES
Classroom Program Model	
Spring LNF	0.222
Spring NWF	0.195
Spring PSF	0.461
Regression Discontinuity Model	
Spring LNF	0.311
Spring NWF	0.427
Spring PSF	0.364

Source: District DIBELS data.

ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

Making Research Relevant

LOCATIONS

Domestic

Washington, D.C.

Atlanta, GA

Austin, TX

Cayce, SC

Chapel Hill, NC

Chicago, IL

Columbus, OH

Frederick, MD

Honolulu, HI

Indianapolis, IN

Metairie, LA

Monterey, CA

Naperville, IL

New York, NY

Reston, VA

Rockville, MD

Sacramento, CA

San Mateo, CA

Waltham, M

International

El Salvador

Ethiopia

Haiti

Honduras

Kyrgyzstan

Tajikistan

Zambia