WestEd
WestEd.org

Evaluation of the EL Education Language Arts Curriculum in Grades K-2

Technical Report

Jennifer McMaken, Candice Bocala, and Karen Melchior

September 2019

Suggested citation: McMaken, J., Bocala, C., & Melchior, K. (2019). *Evaluation of the EL Education Language Arts Curriculum in Grades K-2: Technical report.* San Francisco, CA: WestEd.

WestEd is a nonpartisan, nonprofit research, development, and service agency that works with education and other communities throughout the United States and abroad to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has more than a dozen offices nationwide, from Massachusetts, Vermont, Georgia, and Washington, DC, to Arizona and California, with headquarters in San Francisco.

WestEd
WestEd.org

Contents

Introduction

This study examines whether using the EL Education Language Arts Curriculum for kindergarten through grade 2 (K-2) resulted in significantly higher student achievement in reading in relation to a comparison group. Prior research has found that providing teachers access to high quality curriculum in conjunction with professional development to support curriculum use leads to changes in teacher practice and increases in student achievement (Gallagher, 2016; Kirabo & Makarin, 2017; Taylor et al., 2015).[1] However, this curriculum has not undergone a summative evaluation for the early elementary grades. This study was designed to provide educators with information about whether the EL Education Language Arts Curriculum demonstrates evidence of effectiveness under the Every Student Succeeds Act (ESSA).

Curriculum Overview

The EL Education Language Arts Curriculum is a standards-based core literacy program that uses real world content and primary texts to engage students with literature. Two components of the grades K-2 curriculum were studied in this evaluation: Module Lessons, a content-based literacy curriculum that fosters student vocabulary and comprehension through read-alouds of text, and the K-2 Reading Foundations Skills Block, a structured phonics approach to literacy with a focus on systematic and predictable relationships between letters and sounds. The K-2 Reading Foundations Skills Block curriculum was designed to address the Common Core State Reading Foundations Standards by focusing on students' development of early literacy skills, including fluency, phonemic awareness, and phonics. The K-2 Reading Foundations Skills Block curriculum is based on the Phased Theory (Ehri, 2005)[2] of

[1] Gallagher, H.A. (2016). Professional Development to Support Instructional Improvement: Lessons from Research. Washington, D.C.: SRI Education.

Kirabo, J. & Makarin, A. (2017). Can Online Off-the-Shelf Lessons Improve Student Outcomes? Evidence from a Field Experiment. American Economic Association.

Taylor, J.A., Getty, S.R., Kowalski, S.M., Wilson, C.D., Carlson, J., & Van Scotter, P. (2015). Efficacy Trial of Research-Based Curriculum Materials with Curriculum-Based Professional Development. Colorado Springs, CO.

[2] Ehri, L. C. (2005). Development of Sight Word Reading: Phases and Findings. In M. J. Snowling, & C. Hulme (Eds.), The Science of Reading: A Handbook (pp. 135-154). Oxford: Blackwell Publishing.

phonics that is arranged by the types of letter-sound connections students make as they learn to read and write. K-2 Reading Foundations Skills Block is designed to be taught five days a week for one hour per day. The hour is broken into 15-20 minutes of whole-group instruction and 40-45 minutes of small-group instruction. The curriculum includes activities for instruction as well as a full set of assessments, including benchmark assessments that can identify students' skill level, cycle assessments to measure content mastery of the curriculum cycles, and daily assessments that track progress toward mastery of each day's learning target. The K-2 Reading Foundations Skills Block curriculum can be used in conjunction with EL Education's two other content-based literacy curricula: Module Lessons and K-2 Literacy Labs[3], both of which foster student vocabulary and comprehension in part through read-alouds of text.

EL Education also provided professional development and support to schools using the K-2 Reading Foundations Skills Block curriculum as part of this evaluation. The 2018-19 school year was the first year schools had utilized this curricular component in the district. For school leaders, this included direct coaching and debriefs on implementation quality. Teachers had access to an intensive introduction to the curriculum as well as direct coaching and a series of webinars provided by an EL Education Curriculum Specialist.

Research Question

Under ESSA, states and districts are expected to use evidence-based programs in instruction. This study was designed to provide information about whether the EL Education Language Arts Curriculum demonstrates evidence of effectiveness under ESSA for English language arts (ELA). The central research question was: is the ELA achievement of students using the EL Education Language Arts Curriculum significantly different than the ELA achievement of students not using the Language Arts Curriculum?

The rest of this report details how this question was investigated. The next section describes the research design and methodology. The final section presents findings from the study.

Study Design

The study occurred in an urban school district during the 2018-19 school year. All schools in the district were implementing Module Lessons as their content-based ELA curriculum. In addition, eight schools in the district were also implementing the K-2 Reading Foundations Skills Block curriculum during their literacy foundational skills instructional block. Other schools in the district were using a variety of skills

[3] Literacy Labs are not evaluated as part of this study as the participating district did not use this component of the Language Arts Curriculum for K-2.

materials and curricula for this instructional block. The most common alternative skills curriculum in use was Houghton Mifflin Harcourt's Journeys. Given that there are some overlapping features across EL Education's curricula, the study was designed to investigate the research question through two quasi-experimental approaches. The first approach relied on obtaining a matched sample of students and their achievement scores from outside of the study district; the second approach was a within district comparison. These approaches are detailed below.

Study 1: National Sample

As the study district was already using the Module Lessons portion of the EL Education Language Arts Curriculum, a comparison group of students with no exposure to EL Education's curriculum was obtained. The study district uses NWEA's MAP Growth assessment as an interim assessment districtwide. We were able to use NWEA's virtual comparison group reporting feature to obtain a matched sample of students to those in the study. This approach provided an opportunity to compare outcomes for students in district study schools to settings where the EL Education curriculum was not in use. The study team provided NWEA a list of all schools utilizing EL Education's curriculum so they could be excluded from the study pool. This sample also provided an opportunity to examine the generalizability of Language Arts Curriculum outcomes beyond a single district.

Measures

The outcome measure for this study was spring 2019 scores on the NWEA MAP Growth reading assessment. MAP Growth is an adaptive test that assesses both foundational skills and comprehension. MAP is scored on a RIT scale that was constructed using the Rasch model of item response theory (Lord & Novick, 1968; Lord, 1980; Rasch, 1980).[4] This approach accounts for item difficulty in scoring and results in a single, equal-interval scale that measures student achievement across grades (NWEA, 2011).[5] Because the scale is stable and uniform, it allows direct comparison of scores across students or between students over time.

The matched dataset did not include demographic data for comparison sample, only for students in the treatment group. For treatment students, information on student sex, race/ethnicity, and grade was available in addition to information on achievement.

Methods

NWEA's matching process is proprietary, but the process matches between 1 and 51 students to each treatment student based on a set of key student and school characteristics, including school poverty

[4] Lord, F. M. & Novick M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

[5] NWEA. (2011). *Technical manual: For Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG)*. Portland, OR: NWEA.

rate, grade level, subject area, starting MAP score (fall 2018), and number of instructional weeks between the test events being compared. Sensitivity analyses of this method have found that this does not introduce bias to the growth estimate between the treatment and matched sample.[6] While student-level demographic measures available for creating a matched comparison group are limited in this approach, the fact that each treatment student is matched to their own comparison group of up to 51 students allows us to compare the academic achievement of treatment students in relation to a large group of comparison students.

The matched dataset was provided to the study team by NWEA for analysis. As demographic data were unavailable for the matched comparison group to use as covariates in regression analysis, differences in spring 2019 MAP scores between the treatment and comparison groups were calculated using t-tests. Effect sizes were calculated for the overall sample as well as for treatment student subgroups. Treatment subgroup analyses included grade, sex, and race/ethnicity.

Sample

Seven of eight schools who were using the K-2 Reading Foundations Skills Block agreed to participate in the study; the study sample included MAP data on 1,098 treatment students. All but three students were successfully matched with a comparison group from outside the district by NWEA. The number of matches ranged from 1-51 comparison students per treatment student, with an average of 49.8 student matches. After matching, the final sample was comprised of 1,095 treatment students and 54,508 matched comparison students. Table 1 describes the characteristics of the treatment group. While these data are unavailable for the comparison group on an individual level, the matching process accounts for available student characteristics so the comparison group should mirror the characteristics of the study sample. This type of statistical matching should produce a balanced sample of treatment and comparison students.[7]

The treatment sample has slightly more female students than male students. While students are fairly evenly divided across grades K-2, kindergarten does have the largest enrollment in the sample, with grade 1 enrollment dropping from kindergarten, and grade 2 enrollment dropping from grade 1. The majority of students in the sample (85%) were Black, with smaller populations of White, Hispanic, Multi-Ethnic, and Asian or Pacific Islander students.

[6] Ma, L. & Cronin, J. (2009). *Evaluating the Effect of Random Selection on Virtual Comparison Group Creation.* Portland: NWEA.

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

TABLE 1.

Study 1 Treatment Characteristics

Characteristic	n	Percent
Sex		
Female	566	51.7%
Male	529	48.3%
Race/Ethnicity		
Asian or Pacific Islander	19	1.7%
Black	935	85.4%
Hispanic	32	2.9%
Multi-Ethnic	23	2.1%
White	80	7.3%
School Grade		
Kindergarten	400	36.5%
Grade 1	361	33.0%
Grade 2	334	30.5%

Study 2: District Sample

In addition to the national matched sample from outside of the study district, we also had access to student records from treatment and comparison schools within the district. This data allowed us to model student outcomes more fully as student-level data were available for all participants and we could account for clustering of students by schools. Additionally, participants from the same district likely had greater shared experiences from living in the same county and attending similar school settings. However, as the district was using the Module Lessons in all schools, the comparison between treatment and comparison students within the district was different than the comparison between treatment and comparison students in study 1. The within district comparison examines the effect of

using the K-2 Reading Foundations Skills Block curriculum in addition to the Module Lessons curriculum on student achievement.

Measures

Spring 2019 MAP reading scores were also the outcome measure for this study. In addition, indicators for school, sex, grade, race/ethnicity, English learner status, and low income status as measured by direct certification were used in analysis.

Methods

A two-stage process was used to identify a well-matched comparison sample of students. At the outset of the study, school-level data on academic performance, racial and ethnic makeup, sex makeup, proportion of low income students, proportion of English learner students, and proportion of students with disabilities were obtained. Propensity score matching (PSM) was used to identify a set of comparison schools that were well-matched to the treatment schools on academic achievement as well as study body composition (e.g., sex, race/ethnicity, economic disadvantage, English learners, and students with disabilities). Nearest neighbor matching of up to four schools was used to identify comparison schools. These selections were compiled and reviewed across all eight treatment schools. One school was paired with each treatment school based on the closest match to school characteristics, accounting for schools that were selected as matches multiple times.

In the second stage of matching, nearest neighbor matching without replacement was used to identify a subset of students at the eight comparison schools[8] who were well-matched to treatment students on fall 2018 MAP scores as well as sex, grade, and race/ethnicity. These variables were chosen for the propensity model because of theoretical import to student achievement. Model fit was examined by calculating the standardized differences between the treatment and comparison groups on the PSM model variables. Hedges' g was calculated for MAP achievement and the Cox index was calculated for the dichotomous variables. Model fit was examined in relation to the What Works Clearinghouse standards for baseline equivalence: variables with a difference of less than 0.05 standard deviations were considered balanced and did not require any statistical adjustment. Variables with a difference between 0.05 and 0.25 standard deviations were not balanced but could be controlled for in analyses to establish equivalence. Variables with a difference above 0.25 were not balanced, even when controlled for in analyses. In addition, we examined the goodness of fit of the PSM model by looking at the region of common support (i.e., the region in which treatment and comparison individuals have overlapping propensity scores). Common support is a model assumption that must be met in order to reliably estimate treatment effects as it ensures that individuals with the same characteristics could plausibly be part of either the treatment or comparison groups. The full baseline equivalence results are presented in the results section below.

Outcome analyses were conducted on the matched sample to estimate the effect on reading achievement from using the Language Arts Curriculum. To account for the clustering of students by

[8] As all eight comparison schools agreed to participate in the study, none were excluded from analysis even though only seven treatment schools were part of the study.

school, a multilevel model was specified. We were unable to control for clustering at the classroom level as well, as it was not feasible to connect classroom rosters with student administrative records for analysis. In this analysis, spring MAP score was the student-level outcome, predicted by a treatment indicator. Treatment status and a set of student-level covariates were used to estimate MAP achievement for students at level-1, clustered within schools at level-2. Covariates at level-1 included student-level demographic characteristics: sex, race/ethnicity, English learner status, low income status, grade level, and fall MAP score. At level-2, a school indicator was used to account for student clustering. The resulting model was:

Level 1 (Student):

$$y_{ij} = \beta_{0j} + \beta_{1j}(\text{Treatment})_{ij} + \beta_{2j}(\text{Fall MAP})_{ij} + \sum_{m=a}^{b} \beta_{mj}^{b}(\text{Student Char.})_{ij}$$

Level 2 (School):

$$\beta_{0j} = \gamma_{00} + \sum_{r=v}^{w} \gamma_{0r}^{w}(\text{School})_{j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{mj}^{b} = \gamma_{m0}^{b}$$

At student level-1, the outcome y_{ij} represents the estimated spring MAP score for student *i* in school *j*, given whether the student was part of the treatment group, (β_{1j}), conditioned on fall MAP achievement and a vector, *m,* of student covariates. At school level-2, the randomly varying intercept parameter for school average MAP achievement (β_{0j}) is a function of intercept γ_{00}, a vector, *r,* of school identifiers, and a between-school error term, u_{0j}. Slopes are fixed.

Sample

The sample was limited to students who did not change schools over the course of the school year and were enrolled for the full year. 2,458 students across eight comparison schools and seven treatment schools met these conditions. MAP data was missing for 88 students for at least one time period. As this represented less than four percent of the sample and was balanced between the treatment and comparison groups, listwise deletion was used to remove the missing cases from the dataset. The resulting sample of 2,370 students was used to create a well-matched comparison group for analysis. The characteristics of this sample are described in Table 2 below.

In the full, unmatched sample, treatment students scored slightly lower on the fall MAP assessment than students in comparison schools. Other differences were found between the two groups: students in comparison schools were more likely to be White on average than students in treatment schools, more likely to be female, less likely to be an English learner, and slightly less likely to be economically disadvantaged.

TABLE 2.

Study 2 Sample Characteristics—All Students

Characteristics	Comparison	Treatment
Fall 2018 MAP Score	157.3 (n=1,323)	155.9 (n=1,047)
Sex		
Female	52.5% (n=695)	51% (n=534)
Male	47.5% (n=628)	49% (n=513)
Race/Ethnicity		
Asian or Pacific Islander	0.9% (n=12)	1.6% (n=17)
Black	78% (n=1,032)	86.1% (n=901)
Hispanic	2.7% (n=36)	2.9% (n=30)
Multi-Ethnic	2.5% (n=33)	1.8% (n=19)
White	15.6% (n=206)	7.6% (n=79)
Economically Disadvantaged	61.2% (n=810)	63.9% (n=669)
English Learner	0.7% (n=9)	2.4% (n=25)
Grade		
Kindergarten	34.9% (n=461)	35.7% (n=374)
Grade 1	33.7% (n=446)	34.2% (n=358)
Grade 2	31.4% (n=416)	30.1% (n=315)

Results

Significant positive differences were found in reading achievement between the treatment and comparison groups for both studies.[9] In study 1, treatment students outperformed the comparison students on the spring MAP reading assessment by 4 points, equivalent to an effect size of 0.22. In study 2, treatment students also outperformed the comparison students on the spring MAP reading assessment by almost 4 points, equivalent to an effect size of 0.21. The overall results and subgroup analyses are presented below in greater detail.

Study 1: National Sample Results

Achievement results from the national sample analysis were significantly higher for students in the Language Arts Curriculum treatment schools than the matched virtual comparison group students. Figure 1 below shows the fall and spring MAP reading scores for both groups. The groups were equivalent for the fall administration, with an average MAP score of 155.5 (standardized difference of 0), but by spring, the treatment group was out-performing the comparison group by 4 points (173 to 168.9). This difference is equivalent to an effect size of 0.22. Subgroup analyses were also performed as part of this study to examine how widespread treatment differences were across treatment student groups since the dataset was not structured to allow for modeling of group variance.

Figure 2 presents results from the analysis of differences by sex. Significant differences were found for both male and female treatment students and the matched comparison group. The effect size is larger between treatment and comparison groups for male treatment students than the difference between groups for female treatment students, at 0.29 versus 0.17 for females.

[9] T-test differences between treatment and comparison group means on spring MAP scores for study 1 were significant at $p<0.001$. The estimated treatment effect for study 2 was significant at $p<0.05$. For study 1, subgroup differences reported on sex, gender, race/ethnicity, and grade were significant at $p<0.001$, with the exception of differences between Hispanic students and grade 2 students, neither of which were significant at $p<0.05$ between the treatment and comparison groups.

FIGURE 1.

Fall and Spring MAP Achievement by Treatment Group

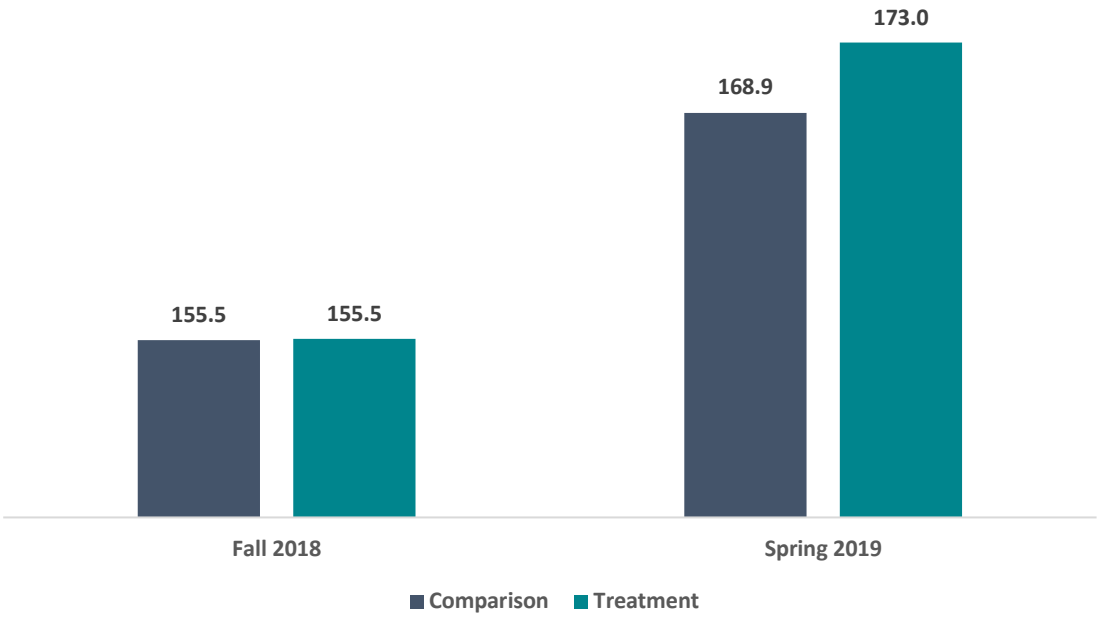
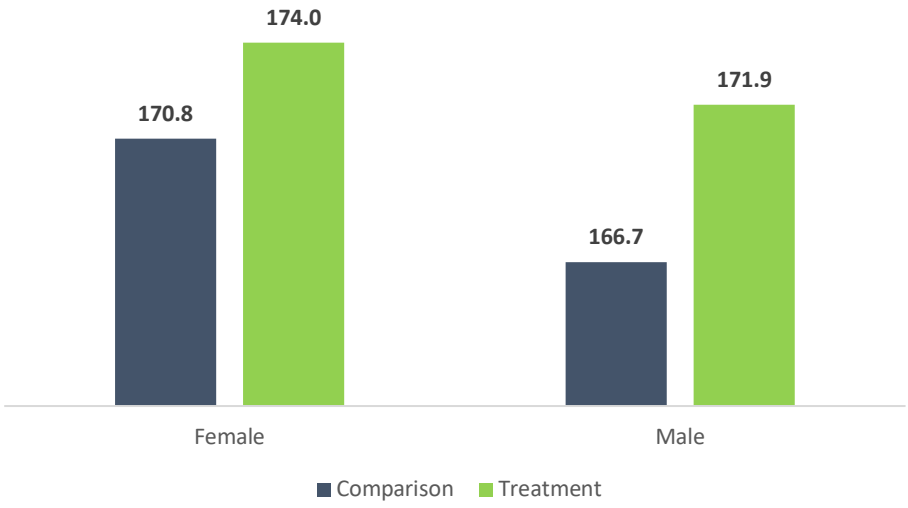


FIGURE 2.

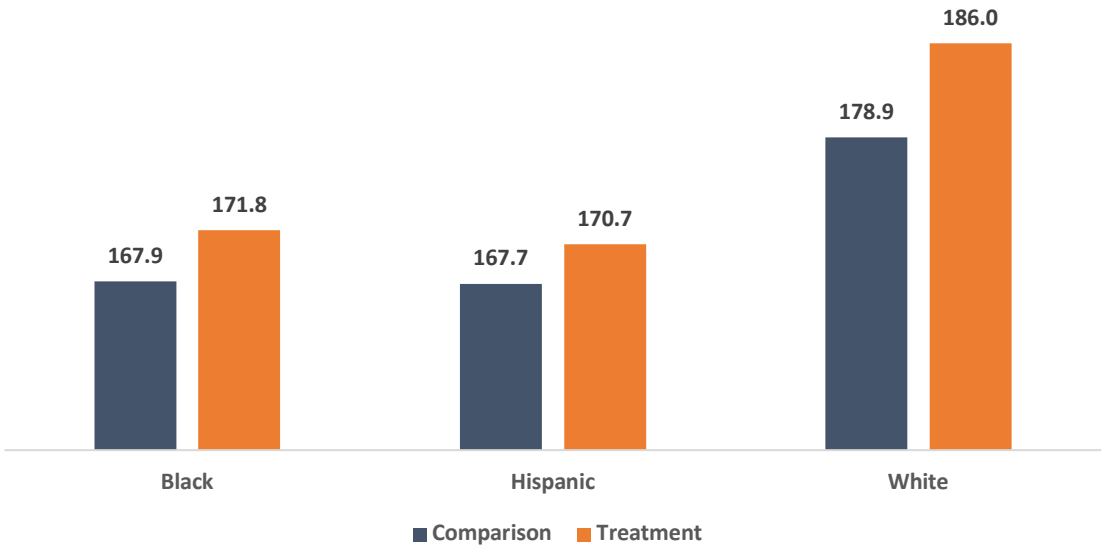
Spring MAP Achievement by Sex



Achievement differences were also found by race and ethnicity among treatment students. Treatment students outperformed comparison students across all racial and ethnic groups. Black treatment students outperformed the matched comparison students by almost 4 points (effect size of 0.22). Hispanic treatment students outperformed matched comparison students by 3 points (effect size of 0.17, though this difference was not statistically significantly at $p<0.05$ due to the small sample size of Hispanic students in the district), and White treatment students outperformed matched comparison students by 7 points (effect size of 0.38) in the national sample. Figure 3 presents these results.

FIGURE 3.

Spring MAP Achievement by Race/Ethnicity



Finally, we examined differences between treatment and comparison groups by grade level. Figure 4 presents results from this analysis. Significant differences between treatment and comparison students were found for both kindergarten and grade 1. There was not a significant difference in achievement between treatment and comparison students for grade 2. This may be partly related to the fact that student growth on the MAP declines across grades. Students in kindergarten, regardless of treatment group, had the largest ELA gains from fall to spring, followed by students in grade 1. Students in grade 2 posted the smallest ELA gains from fall to spring. This is in line with findings from NWEA's norming study which showed smaller gains as grade level increased.[10] Kindergarteners in the treatment group outperformed their counterparts in the comparison group by 7 points on the spring administration of

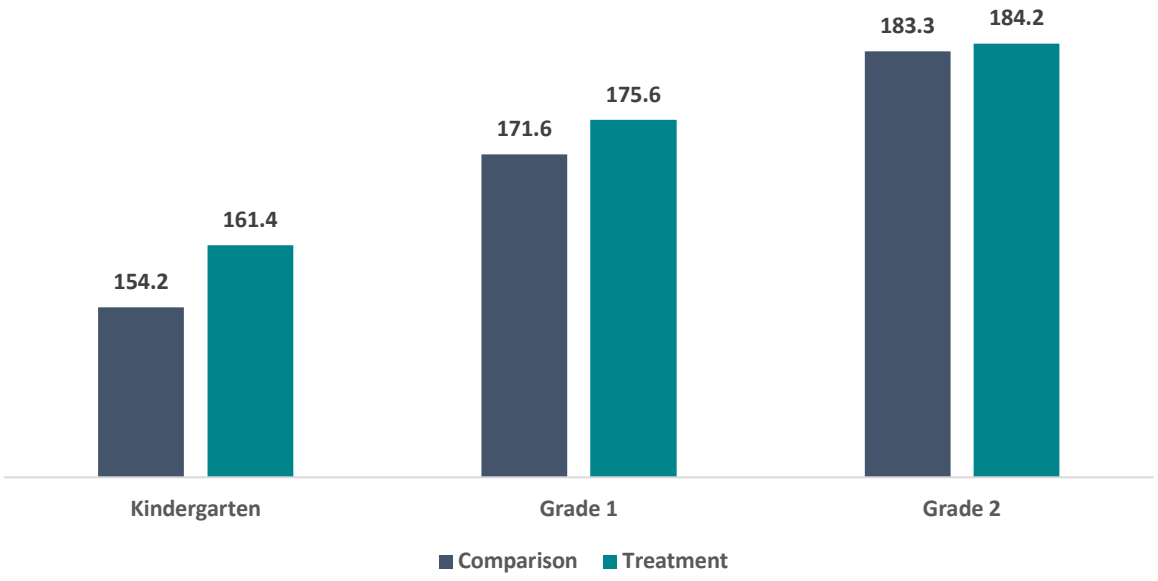
[10] Thum Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP Norms for Student and School Achievement Status and Growth*. Portland, OR: NWEA.

the MAP reading assessment; this is equivalent to an effect size of 0.54. Grade 1 treatment students outperformed the comparison group students by 4 points, equivalent to an effect size of 0.25.

While it is not possible to rule out other explanatory factors with a quasi-experimental study, the widespread extent of treatment differences as well as the magnitude of these differences relative to the comparison group provide compelling evidence that the Language Arts Curriculum may be contributing to gains in student achievement.

FIGURE 4.

Spring MAP Achievement by Grade Level



Study 2: District Sample Results

Model Balance

Results from the PSM analysis were reviewed to ensure a well-matched comparison group had been formed for analysis. Table 3 presents the baseline equivalence parameters from the model. The nearest neighbor matching procedure created a well-balanced comparison group. All variables included in the matching model other than Hispanic were below the 0.05 standard deviation threshold for demonstrating model balance. The standardized difference in rates of Hispanic students between the two groups was 0.14, which can be accounted for by controlling for the variable in outcome analyses.

TABLE 3.

PSM Model Balance

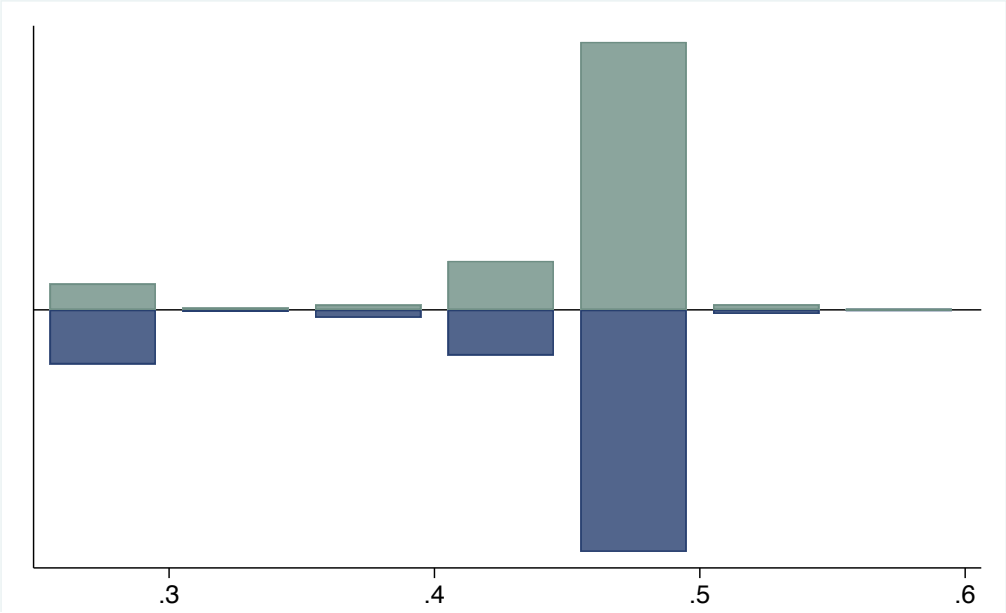
Variable	Comparison	Treatment	Standardized Difference[a]
Fall 2018 MAP Score	155.6 (n=1,047)	155.9 (n=1,047)	0.02
Sex (Female)	50.5% (n=529)	51% (n=534)	0.01
Race/Ethnicity			
Black	86.8% (n=909)	86.1% (n=901)	0.04
Hispanic	2.3% (n=24)	2.9% (n=30)	0.14
Multi-Ethnic	1.8% (n=19)	1.8% (n=19)	0.00
White	7.6% (n=79)	7.6% (n=79)	0.00
School Grade			
Kindergarten	35.6% (n=373)	35.7% (n=374)	0.00
Grade 1	35.1% (n=367)	34.2% (n=358)	0.02
Grade 2	29.3% (n=307)	30.1% (n=315)	0.02

[a] Values presented are Hedges' g for continuous variables and the Cox index for dichotomous variables.

We also examined the degree of common support for the model by looking at the overlap in propensity scores between treatment and comparison groups. Figure 5 displays this graph. Treatment students are represented by teal bars and the matched comparison group individuals are indicated by blue bars. As can be seen, there is a high degree of overlap between the two groups, both in terms of range of scores as well as in the distributional density of scores. For example, both of the bars on the chart between a propensity score of 0.4 and 0.5 are approximately balanced between the two groups. This data as well as the model fit statistics in Table 3 indicate that the two groups are highly comparable. While we cannot be sure the two groups are comparable beyond the factors balanced in the model, the model fit results provide additional evidence that the sample selected is a reasonable comparison group for the treatment group. Achieving balance between the treatment and comparison groups allows us to examine what effect using the K-2 Reading Foundations Skills Block has on achievement beyond using the Module Lessons.

FIGURE 5.

Common Support of Propensity Scores



Student-Level Outcomes

The outcome analysis for study 2 was the multilevel model predicting spring MAP scores based on a set of student characteristics, fall MAP achievement, and treatment status while accounting for the clustering of students by school. Results from this analysis mirror what was found with the national sample in study 1. Table 4 presents the results of this two-level model. Significant associations were found between spring MAP scores and treatment status, being Asian, being in kindergarten or grade 1, fall MAP scores, and not being economically disadvantaged. The two largest estimate effects were for Asian students (associated with a 5.1 point increase on the spring MAP score) and being in a school that was utilizing the K-2 Reading Foundations Skills Block curriculum. Being part of the treatment group was associated with a 3.9 point increase in the spring MAP score. This is equivalent to an effect size of 0.21.

As expected, prior achievement, as measured by the fall MAP score, was a significant predictor of spring achievement. Each point earned on the fall assessment was associated with a 0.8 point increase on the spring assessment. Smaller effects were found for grade levels and economic disadvantage. As with study 1, students in kindergarten and grade 1 demonstrated higher gains in spring achievement compared with students in grade 2: being a kindergartener was associated with a 2.3 point increase on the spring MAP assessment and being in grade 1 was associated with a 2.1 point increase on the spring MAP assessment. Therefore while treatment students across all grades had significantly higher spring MAP scores on average than students in the comparison group, kindergarten and grade 1 treatment

students had an even larger average increase in reading scores when compared to the spring scores of grade 2 treatment students.

TABLE 4.

Multilevel Model Results

Variables	Regression Coefficient	Standard Error	p-value
Treatment	3.9	1.81	0.032*
Female	-0.2	0.48	0.632
White	1.1	1.07	0.322
Hispanic	-0.2	1.62	0.919
Multi-Ethnic	3.5	1.77	0.050
Asian	5.1	2.18	0.019*
Kindergarten	2.3	0.82	0.005**
First Grade	2.1	0.63	0.001**
English Learner	-2.8	2.16	0.188
Economically Disadvantaged	-1.7	0.58	0.004**
Fall MAP Score	0.8	0.02	0.000**
Constant	47.1	3.56	0.000**

** significant at p<0.01; * significant at p<0.05

Considerations

This study uses a quasi-experimental design; therefore, we cannot rule out unobserved factors that may account for differences between groups. Additionally, we have limited information on instructional practices both for the national sample as well as the district sample. Within the district, there was wide variation in MAP performance across schools suggesting that factors such as instructional practice or school-based factors such as climate may mediate achievement outcomes.

However, the results from these two studies consistently suggest that use of the Language Arts Curriculum was associated with a significant increase in spring MAP reading achievement scores for the treatment group. While the research design cannot determine that this relationship is causal, the consistency of the findings between the two studies lend support to the curriculum playing a role in these differences.