MATHEMATICA
Policy Research

WORKING PAPER 62

Working PAPER

BY NAIHOBE GONZALEZ, SOPHIE MACINTYRE, AND PILAR BECCAR-VARELA (MILLS TEACHER SCHOLARS)

Challenges in Adolescent Reading Intervention: Evidence from a Randomized Control Trial

June 2018

ABSTRACT

This paper presents findings on the implementation and impacts of Leveled Literacy Intervention (LLI) in Oakland, California, where the school district conducted the nation's first randomized controlled trial of LLI in secondary grades. LLI is a short-term, intensive intervention designed to help teachers provide small-group instruction to struggling readers. Many school districts across the country have used LLI, which research evidence has shown to rapidly improve outcomes for students in early elementary grades. During the trial, secondary schools in Oakland faced various challenges implementing LLI, leading students to experience different levels of LLI duration, intensity, and fidelity. LLI had no impact on students' reading comprehension and a negative impact on their mastery of English language arts/literacy standards. Students who were pulled out of other classes to receive LLI were particularly negatively affected, possibly as a result of missing grade-level content. This study's findings highlight challenges in implementing effective literacy interventions for struggling adolescent readers.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305L160003 to the Oakland Unified School District. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors gratefully acknowledge the district and school staff who supported this study, including Jean Wing, Nancy Lai, Abbey Kerins, and Rinat Fried. We also thank Brian Gill (Mathematica Policy Research) and Eric Isenberg (American Institutes for Research) for their contributions to this study. Finally, we are grateful to the grant's project officer, Allen Ruby, for providing helpful suggestions throughout the project.

I. INTRODUCTION

Too many American youth leave high school without the literacy skills that colleges and employers demand. In 2015, 28 percent of 12th graders in the United States performed below the basic level on the National Assessment of Educational Progress, meaning that those students had less than partial mastery of the knowledge and skills that are fundamental for reading at their grade level (U.S. Department of Education 2018). Because literacy is associated with greater job opportunities and higher incomes, the stakes are high for finding effective ways to accelerate these students. Adults with below-basic literacy are 16.5 times more likely to receive public assistance and 5 times more likely to earn less than $300 per week relative to those with the highest level of literacy (Wood 2010).

Secondary schools often struggle with finding effective ways for quickly accelerating the progress of older, struggling readers. The challenge is apparent in literacy trends over the last decade: the average reading performance of 12th graders on the National Assessment of Educational Progress has been stagnant, as has the achievement gap between low-income students and those not eligible for free or reduced-price lunch (U.S. Department of Education 2018). As summarized by Levin et al. (2010), "Progress in strengthening young people's literacy now depends on schools a) choosing appropriate programs and b) implementing them consistently and effectively."

An ideal setting in which to study reading intervention for secondary students is the Oakland Unified School District (OUSD), where 52 percent of students in grades 6 through 12 scored multiple years below their expected reading level in 2016 on the Scholastic Reading Inventory (SRI) and almost one-third scored at least four years below grade level. As in other school districts around the country, Oakland students who are multiple years behind in reading also face other challenges. In OUSD, 33 percent of such students are English learners, 21 percent are in special education, and 88 percent are eligible for free or reduced-price lunch.

To help these adolescent readers improve their literacy skills, OUSD invested in piloting the Fountas & Pinnell Leveled Literacy Intervention (LLI), an intensive reading intervention program. After one year of piloting the program in a small number of secondary schools, the district partnered with researchers to conduct the nation's first randomized controlled trial of LLI in secondary grades. Many school districts across the country have used LLI, often as part of a Response-to-Intervention model, and the program has shown promise in rapidly improving outcomes for students in early elementary grades (What Works Clearinghouse 2017). This paper presents findings on the implementation and impacts of LLI in secondary schools.

II. OVERVIEW OF LLI

LLI is a short-term, intensive intervention system designed to help teachers provide daily, small-group instruction to students who are not achieving grade-level expectations in reading. It is intended to supplement, rather than replace, regular literacy instruction, and draws on research on reading comprehension and skill acquisition, vocabulary development, oral fluency, and student engagement and motivation (see Heinemann 2015 for a summary of the research base for LLI for grades 3 through 12). LLI was developed by Irene C. Fountas and Gay Su Pinnell and is published by Heinemann.

Although it was originally developed for students in early elementary grades, LLI has since expanded across K–12. Materials are bundled in kits for different reading levels and include a series of fiction and nonfiction texts and sequential lesson guides of progressing difficulty.[1] Odd-numbered lessons focus on discussing and revisiting the book from the previous lesson, phonics/word work, and reading a new book that is in students' instructional reading level. Even-numbered lessons focus on revisiting the book from the previous day, conducting a reading assessment for progress monitoring, writing about the book from the previous day, phonics/word study, and introducing a new book that is at students' independent reading level. Each LLI lesson guide also provides suggestions for supporting English learner students.

To implement LLI, teachers begin by assessing students with the Fountas & Pinnell Benchmark Assessment System (F&P BAS), a one-on-one assessment that matches students' instructional and independent reading abilities to the text-level gradient used by LLI. Teachers then form small groups of three to five students with similar assessment scores and deliver 45-minute daily lessons. Lessons may be adapted to 30 minutes by using guidance provided by the program. The recommended program length ranges from 12 weeks to 24 or more weeks and depends on students' starting reading level and progress. For the starting reading levels of most students in this study (corresponding to grades 3 through 5), LLI recommends that students participate for 18 to 24 weeks.

Before implementing the program, LLI teachers may participate in training delivered by Heinemann coaches. Heinemann offers various training opportunities, including three-day, on-site trainings that cost approximately $9,600 for up to 30 teachers. In addition to the training and the kits (which range in cost from $2,900 to $4,950), an important input to LLI is teacher time, particularly given the small student-to-teacher ratio required by the intervention. In OUSD, teachers were typically assigned to teach LLI half time and served two LLI groups, or about 10 students. Based on the national average teacher salary, the cost of a half-time teacher is approximately $36,800 per school year.[2] Other implementation costs include support from principals and district leadership and classroom facilities.

LLI is based on a theory of change positing that the progress made by struggling readers toward the goal of reading at grade level is affected by factors related to LLI and contextual factors (Ransford-Kaldon et al. 2013). For instance, the program requires that teachers are qualified and trained to deliver the lessons, that lessons are taught with fidelity, and that the duration of LLI meets the needs of students to produce results. Important contextual factors that influence the program's success include school-level supports for literacy instruction, the quality and continuity of regular classroom instruction, and the support that students receive at home.

[1] There are seven LLI kits, each one denoted by its own color. As of June 2018, the LLI Orange System (kindergarten) cost $2,900. The Green System (grade 1) cost $3,416, the Blue System (grade 2) cost $3,324, and the Red (grade 3), Gold (grade 4), Purple (grade 5), and Teal systems (grade 6 through 12) cost $4,950. The teachers in this study used the Blue, Red, Gold, Purple, and Teal systems.

[2] According to the U.S. Department of Education (2016), the average teacher salary in the United States in the 2016–2017 school year was $58,950. We applied a 25 percent estimated fringe benefit to this figure, following Levin et al. (2010).

III. PREVIOUS LITERATURE ON LLI

Past studies indicate that LLI is effective at improving the reading skills of younger students. In three randomized controlled trials in diverse settings with students in kindergarten through grade 2, LLI produced significant gains in reading after 12 to 18 weeks, with effect sizes ranging between 0.39 and 0.81 standard deviations across grade levels (Ransford-Kaldon et al. 2010; Ransford-Kaldon et al. 2013). Other non-experimental studies have also found promising evidence of LLI's effectiveness among students in early grades (e.g., Taylor 2017; Odell 2012; Peterman et al. 2009; and Harrison et al. 2008). A review of the research on LLI by the What Works Clearinghouse (2017) determined that LLI had positive effects on general reading achievement, potentially positive effects on reading fluency, and no discernible effects on alphabetics for students in kindergarten through grade 2.

In the two studies that met the What Works Clearinghouse's review standards, LLI was implemented with a high degree of fidelity (Ransford-Kaldon et al. 2010; Ransford-Kaldon et al. 2013). Classroom observers in both studies determined that lessons were delivered as designed more than 95 percent of the time. In addition, teachers participated in eight days of training in how to implement the program and received additional professional development throughout the school year. In both studies, 91 percent of teachers felt that they had received adequate professional development for implementing LLI. Researchers also found that regular classroom instruction followed many of the same principles of LLI.

At the same time, however, the studies highlighted some implementation challenges. Students received fewer than the recommended number of instructional sessions, although they still made significant gains. Several classroom observers and teachers noted that lessons were fast-paced and could not be completed adequately as designed within the suggested time frame. Staff also observed that the reading assessment activity in the even-numbered lessons took up too much time. Finally, when asked about logistical issues, most LLI teachers mentioned time and/or scheduling, particularly coordinating with classroom teachers to pull out students for LLI.

IV. METHODS

To determine the impact of LLI on secondary students' reading achievement, this evaluation used an experimental design that randomly assigned groups of students within a school to either a treatment group, which received LLI, or a control group, which proceeded with business as usual. In this section, we describe the study's school and student recruitment, random assignment, data collection, and analysis methods.

School and student recruitment

In fall 2016, OUSD recruited 10 secondary schools that planned to implement LLI during the 2016–2017 academic year. The district offered schools a stipend of $110 per LLI student as an incentive to participate in the study. Of the more than 20 secondary schools that were contacted, 7 middle schools and 3 high schools agreed to participate in the study.

Because the schools in the study were not randomly sampled, they are not necessarily representative of all secondary schools in OUSD or other urban school districts. Schools that were interested in participating had to be prepared to implement LLI, and they had to be willing to allow students to be randomly assigned to the intervention. On average, participating schools had a high share of students who were eligible for free or reduced-price lunch (87 percent), were English learners (31 percent), and scored three or more years below grade level on the SRI in fall 2016 (50 percent).

Within the study schools, teachers selected students in grades 6 through 9 to be included in random assignment based on the same criteria they would have used to select students for LLI in the absence of the study. A student's eligibility for the study was largely based on whether the student's starting reading level on the F&P BAS test was multiple years below grade level and resembled the reading level of other students who might have been eligible for the study, so that homogeneous instructional groups could be formed as specified by LLI.

Random assignment

In fall 2016, teachers used students' F&P BAS scores (and, if relevant, any scheduling constraints) to form groups of three to seven students with similar starting reading levels. The study team then randomly assigned the groups to treatment and control groups separately for each study school. Group-level random assignment allowed teachers to maintain control over forming the instructional groups—an important component of LLI.

To improve the precision of the study's impact estimates and reduce the possibility of a chance imbalance between the treatment and control groups, we paired, when possible, each group into strata (matched pairs) before random assignment on the basis of similar median pre-intervention F&P BAS scores. In some cases, groups had to be paired if they shared a scheduling constraint. Within each stratum, we randomly assigned one group to the treatment group and the other to the control group. [3] Across all 10 study schools, 292 students in 76 groups were randomly assigned, with 145 students assigned to the treatment group and 147 assigned to the control group.

As expected, students selected for the study were low-performing students relative to the rest of the district. For example, their average z-score on the fall 2016 SRI was -0.54 standard deviations. Stated differently, they scored 3.8 years below their expected grade level on average. Students in higher grade levels tended to be further behind—on average, students in grade 6 scored 3.0 years below grade level on the SRI, while students in grade 9 scored 5.1 years below grade level. Students selected for the study also had low performance on the spring 2016 Smarter Balanced Assessment Consortium (SBAC) ELA/literacy test, with only 3.6 percent meeting standards.

In Table 1, we present baseline test scores and demographic characteristics of students by treatment status. By chance, students randomly assigned to the treatment group had slightly

[3] One school submitted 22 students for random assignment without groupings or F&P BAS scores. We paired students by using their baseline SRI score and assigned one student in each pair to treatment. The teacher used the students' F&P BAS scores to form LLI groups from the 11 students assigned to treatment. In general, teachers could move treatment students into different group configurations after random assignment. Students in groups assigned to the control condition did not remain in those group configurations, as they did not receive LLI. For these reasons, the randomly assigned groups did not necessarily reflect instructional groups.

lower baseline academic performance than that of control students, although only one difference—on the fall 2016 Scholastic Math Inventory (SMI)—is statistically significant at the 5 percent level. Overall, students in both the treatment and control groups had similar demographic characteristics. The differences in baseline measures are not jointly statistically significant.

Table 1. Baseline student characteristics

Baseline characteristic	Treatment		Control		Standardized difference
	Mean	N (of 145)	Mean	N (of 147)	(percent)
Fall 2016 F&P BAS (numeric score, 1–26)	17.3 (2.76)	134	17.9 (2.58)	136	-23.8
Fall 2016 SRI (z-score)	-0.586 (0.542)	143	-0.485 (0.537)	146	-18.7
Fall 2016 SMI (z-score)	-0.592 (0.826)	113	-0.361 (0.854)	123	-27.4*
Spring 2016 SBAC ELA/literacy (z-score)	-0.694 (0.595)	126	-0.560 (0.647)	123	-21.5
Spring 2016 SBAC math (z-score)	-0.619 (0.590)	126	-0.548 (0.743)	125	-10.6
African American (%)	40.0	145	35.4	144	9.4
Asian/Filipino/Pacific Islander (%)	6.21	145	9.72	144	-13.0
Latino (%)	48.3	145	50.0	144	-3.4
Male (%)	51.7	145	53.4	146	-3.4
Eligible for free/reduced-price lunch (%)	92.4	145	92.5	147	-0.2
English learner (%)	34.5	145	26.7	146	16.9
Special education (%)	11.7	145	10.3	147	4.6

Source: Authors' calculations using data provided by OUSD.

Notes: Standard deviations are displayed in parentheses. F&P letter scores were converted to numeric scores (i.e., 17 corresponds to the letter Q). SRI, SMI, and SBAC scores were converted to z-scores defined relative to OUSD's distribution of scores by test, administration, and grade.

* Difference is statistically significant at the 5 percent level.

Some students who were randomly assigned did not take the relevant outcome assessments in spring 2017 (the SRI and the SBAC ELA/literacy test).[4] Overall attrition was approximately 10 percent for the SRI and 6 percent for the SBAC ELA/literacy test. The differences in attrition rates between the treatment and control groups are small and not statistically significant (Table 2). After attrition, the treatment and control groups in each of the two analytic samples demonstrate similar patterns in baseline charactristics as the randomized sample and remain balanced overall (Appendix Tables A.1 and A.2).

[4] The SBAC is offered to students only in grades 6 through 8, so the analysis of ELA/literacy scores is based on those grades while the analysis of SRI scores is based on the larger sample of students in grades 6 through 9.

Table 2. Sample sizes and attrition

	Treatment	Control	Total	Overall attrition (percent)	Differential attrition (percent)
Randomly assigned	145	147	292	-	-
In SRI analysis	128	135	263	9.9	3.6
In SBAC ELA/literacy analysis	95	97	192	5.9	3.9

Source: Authors' calculations using data provided by OUSD.

Data collection

The analyze impacts, we used student achievement and demographic data collected by OUSD in the 2015–2016 and 2016–2017 school years. As mentioned above, we studied student achievement on the SRI and the SBAC ELA/literacy assessments. The district administered the SRI to students in all grades in fall and spring of each school year. Using written materials sampled from various content areas, the SRI measures how well students can read and comprehend literary and expository texts. In spring of each school year, students in grades 6 through 8 also took the SBAC, which tests mastery of grade-level standards in ELA/literacy and mathematics. Both assessments are aligned to Common Core State Standards and are computer-adaptive.[5]

The use of reading comprehension and general literacy achievement to assess the impact of LLI in this study is motivated by earlier research on adolescent literacy. Comprehension is seen as a primary challenge among struggling adolescent readers (Kamil et al. 2008). General literacy achievement, which on the SBAC includes reading comprehension, writing, listening, and research/inquiry components, has been linked to labor market outcomes later in life (Wood 2010).

To study implementation, we used LLI attendance logs and other implementation data provided by OUSD. LLI teachers tracked student attendance in every session, start and exit dates, crossover, and any literacy supports provided to students in the control group. The study's school liaison observed each LLI teacher once between December 2016 and March 2017 and rated the fidelity of each lesson component on a scale of 0 (not observed) to 3 (high fidelity) by using a version of a fidelity rubric developed by Heinemann. The school liaison also collected information about each teacher's implementation of LLI, including the setting (e.g., pullout versus in-class), the number of days per week that LLI was offered, and session length.

Data analysis

This study aimed to learn about the implementation of LLI in secondary schools and evaluate LLI's impact on student literacy outcomes. For the implementation analysis, we conducted descriptive quantitative and qualitative analyses of the implementation data collected. Below, we describe in greater detail the methods used for the impact analysis.

[5] The SBAC ELA/literacy test also contains a performance task component.

Intent-to-treat analysis

In the random assignment design used in this study, the simple difference between the outcomes of treatment and control students is an unbiased estimate of the impact of LLI. However, to improve the statistical precision of the estimates, we measured the treatment-control difference after controlling statistically for small, random differences in the baseline achievement and characteristics of students. Specifically, we controlled for baseline achievement on the SRI and SBAC ELA/literacy and math tests from spring 2016 and the SRI and SMI from fall 2016, all administered before the beginning of treatment. In addition, we controlled for the following student characteristics: gender, race/ethnicity, eligibility for free or reduced-price lunch, English learner status, special education status, and grade level.

Accordingly, we estimated student impacts using the following model:

$$(1)\ y_{ijk} = \alpha_k + \beta T_{jk} + X'_{ijk}\gamma + u_{jk} + \varepsilon_{ijk}$$

where y_{ijk} is the spring 2017 SRI or SBAC ELA/literacy score of student *i* in group *j* within stratum *k*; α_k is a vector of stratum (matched pair) fixed effects; T_{jk} is a treatment indicator that equals 1 if the group was assigned to receive LLI and 0 otherwise; X_{ijk} is a vector of baseline student test scores and demographic characteristics; u_{jk} is a group-specific random error term; and ε_{ijk} is an individual-level random error term. γ is a vector of parameters to be estimated, and β is the key parameter measuring the impact of assignment to LLI, or the intent-to-treat (ITT) impact, on student achievement. We estimated the model with ordinary least squares using standard errors that account for group-level clustering.

Treatment-on-the-treated analysis

We also conducted a treatment-on-the-treated (TOT) analysis, which estimated the impact of receiving LLI, accounting for the fact that some students randomly assigned to LLI did not actually receive it and that one student randomly assigned to the control group nonetheless received LLI. To estimate TOT impacts, we used students' group assignment status T_{jk} as an instrumental variable for the receipt of LLI in a two-stage least squares regression. We defined whether a student received LLI based on whether they attended at least four sessions (or one week's worth).[6] Random assignment status is a valid instrument because it strongly predicts receipt of LLI and affects student outcomes only through receipt of LLI.

The first stage regression (equation 2) estimates the effect of random assignment to LLI on the probability of receiving LLI, and the second stage (equation 3) estimates the impact of receiving LLI on outcomes, given by the coefficient δ.

$$(2)\ Received\ LLI_{ijk} = \eta + \lambda T_{jk} + w_{jk} + v_{ijk}$$

[6] Under this definition, 5 students in the treatment group who attended just one session are not considered to have received LLI. We also considered an alternative definition of receipt based on whether students attended the minimum number of lessons recommended by LLI. However, random assignment status may not be a valid instrument in this case, as it also predicts whether students received fewer than the recommended number of sessions, which could also affect reading outcomes.

$$(3)\ y_{ijk} = \alpha_k + \delta \widehat{Received}LLI_{ijk} + X'_{ijk}\gamma + u_{jk} + \varepsilon_{ijk}.$$

Subgroup analysis

The estimation of overall impacts may mask differences in impacts by factors such as student characteristics or fidelity of implementation. As an exploratory analysis, we tested whether the impacts of LLI varied on key dimensions, including whether (1) students were English learners; (2) students were three or more years behind grade level on the fall 2016 SRI; (3) LLI was taught in pullout groups rather than in scheduled classes; (4) teachers had any experience with LLI; and (5) LLI was taught with high fidelity.[7] These analyses are exploratory because of limited sample sizes within subgroups and the fact that types of LLI classes or teachers were not randomly assigned to students.

For each subgroup of interest, we estimated the following regression model, which adds an interaction term to the benchmark model in equation (1):

$$(4)\quad y_{ijk} = \alpha_k + \beta_1 T_{jk} + \beta_2 (T_{jk} * W_{ijk}) + X'_{ijk}\gamma + u_{jk} + \varepsilon_{ijk}.$$

W represents the relevant variable for the hypothesis being tested (for example, an indicator for whether the student was an English learner). Given that W is always a binary indicator, the coefficient on the interaction term β_2 represents how the impact for members of that subgroup differs from the impact for others, captured by β_1.

Missing baseline data

Data on one or more of the variables used as baseline controls were missing for about 27 percent of students. As described above, we included five baseline assessments as controls to improve the statistical precision of the estimates. Most commonly, students were missing scores on the spring 2016 SRI (21 percent) and fall 2016 SMI (19 percent). To address the missing data problem, we employed a multiple imputation procedure to estimate students' missing baseline values. Simulations suggest that, for randomized controlled trials, this approach results in limited bias (defined as less than 0.05 standard deviations) if missing data are not related to an unobserved variable that is correlated with the outcome (Puma et al. 2009).[8]

We imputed missing baseline test scores separately by treatment or control status by using a chained linear equations model that included all outcome variables and all student characteristic variables in the final impact regressions. We excluded students from the imputation model if they had missing data for both outcome test scores. This restriction excluded 25 students, or 8.6 percent of the randomized sample, with no outcome scores (15 in the treatment group and 10 in the control group). We then conducted the impact regression analyses described above in $M = 10$ imputed data sets.

[7] Teachers were observed once and rated from 0 to 3 on each LLI lesson component. Those component ratings were then averaged for each teacher. The average of the overall ratings was 1.4. High fidelity was defined as an average fidelity rating of 1.5 or higher.

[8] Although rates of missing baseline data were similar between the treatment and control groups, we added to the impact regressions indicators for each baseline assessment that was imputed.

After collecting coefficient and standard error estimates from each of the 10 imputed data sets, we computed multiple imputation coefficients and standard errors using Rubin's combination method (Rubin 1987), where the multiple imputation beta (β_M) coefficient is the average of the beta coefficient values in each imputed data set (β_m) and the multiple imputation standard error SE_M is the square root of the within-imputation coefficient variance plus the between-imputation coefficient variance inflated by a finite imputation correction multiplier:

$$(7)\ SE_M = \sqrt{\left(\frac{\sum_{m=1}^{M} Var_m}{M}\right) + \left(1 + \frac{1}{M}\right)\left(\frac{\sum_{m=1}^{M}(\beta_m - \beta_M)^2}{M-1}\right)}.$$

To test the sensitivity of the results to these imputation methods, we estimated a version of equation 1 with unimputed baseline data but only the fall 2016 SRI as a control for prior achievement, which no students in the analytic samples were missing.

Results

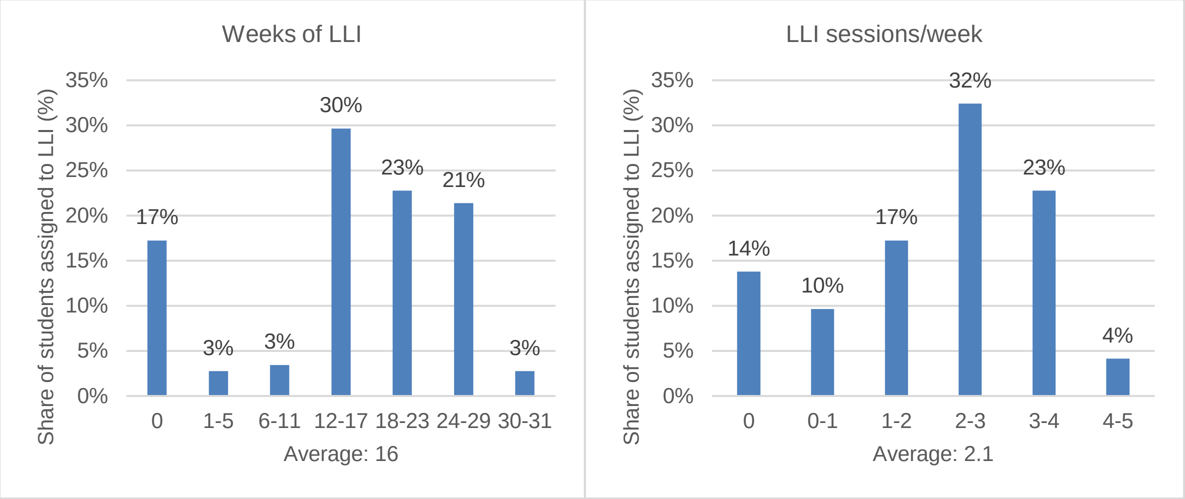
In summary, schools faced various implementation challenges, leading students to receive different levels of LLI duration, intensity, and instructional fidelity. LLI had no impact on students' reading comprehension and a negative impact on their mastery of ELA/literacy standards.

Implementation findings

Students experienced different levels of LLI duration and intensity, and most fell short of the recommended minimum number of sessions. The average student assigned to LLI participated in the program for 16 weeks, although the duration varied significantly across students (Figure 1). School-level factors, such as the program start and end date, and individual student factors, such as the decision to stop participating in LLI, affected how long students remained in the program. In addition, very few students received the recommended intensity of four or five days per week (Figure 1). Student attendance was inconsistent—on average, students assigned to LLI attended 2.1 out of 3.8 sessions offered per week—and only 2 of the 10 study schools offered daily sessions. As a result, 63 percent of students assigned to LLI received fewer than the total recommended minimum number of sessions. This includes 17 percent of students assigned to treatment who did not receive any LLI, the majority of whom were concentrated in one school.

High schools particularly struggled with student participation and engagement. Thirty-six percent of high school students assigned to LLI did not receive LLI compared to 10 percent of students in grades 6 through 8. In most cases, the disparity reflected high school students' (or their families') refusal to participate in the program. High schoolers also attended significantly fewer sessions than students in middle grades (17 compared to 49, on average). Teachers at each of the three high schools in the study reported challenges with student attendance or engagement, particularly with groups pulled out of other classes. For example, one teacher reported that students resented leaving a creative writing class they liked for LLI. Another teacher had to terminate the program in early February because of chronically low student attendance despite initial plans to offer LLI through the end of the school year.

Figure 1. Program duration and intensity for students assigned to LLI



Source: Authors' calculations using data provided by OUSD.

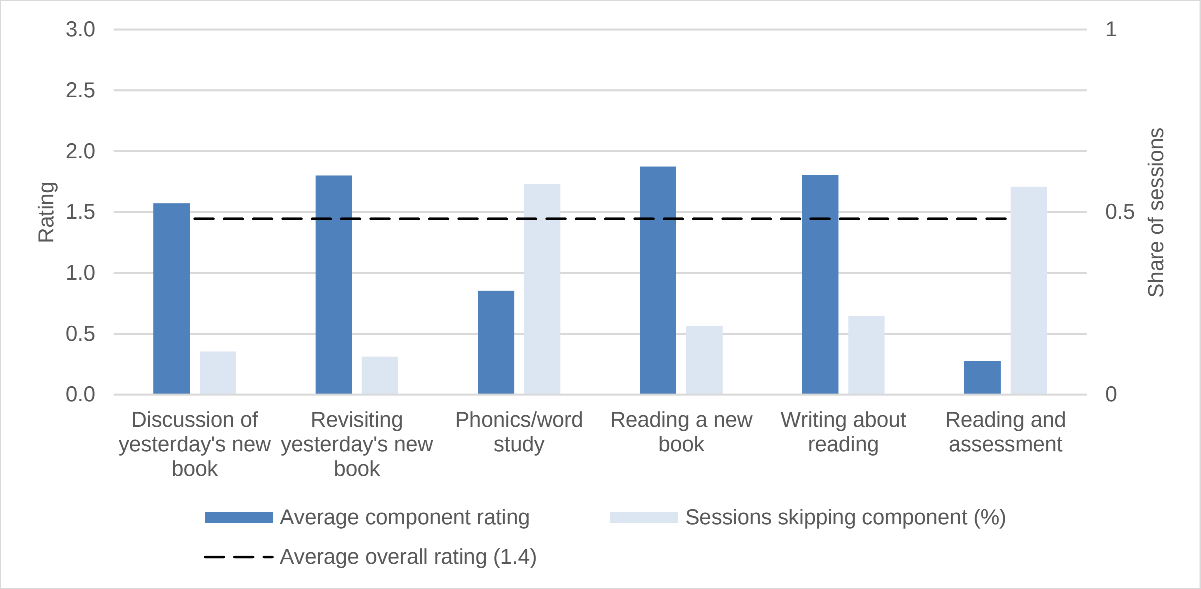
Program start and end dates varied widely across schools. Although most schools planned to start offering LLI in the fall, actual start dates ranged from October to February. Schools reported that securing the appropriate LLI materials, receiving teacher training, and completing the initial one-on-one student assessments sometimes took longer than anticipated. Schools' end dates also varied from February to June. In seven schools, the duration of LLI was fixed (for example, all of second semester), while in the other three schools in the study, LLI was offered through the end of the school year and the duration depended on individual student progress, though exit criteria varied or were not always specified. On average, schools began the program in early December and ended in mid-April.

Scheduling LLI classes proved difficult for some schools, and pulling students out of other classes presented challenges. Of the 10 study schools, 5 offered LLI in a scheduled class, 4 pulled students out of other classes, and one used both approaches. Some schools noted that it was difficult to create a regularly scheduled LLI class because LLI serves a small number of students and their participation is not determined until after the start of the school year. Although offering LLI in pullout groups was logistically easier, students had to miss other classes and make up the work, which concerned some students, parents, and classroom teachers. Some LLI teachers tried to minimize this burden by holding sessions at a different time each day. Students in pullout groups were more likely to refuse LLI and attend fewer sessions than those in regularly scheduled classes, particularly at the high school level. On average, students in pullout groups attended 28 sessions, while students in regularly scheduled classes attended 47 sessions.

Instructional fidelity was relatively low, primarily because of skipped or modified lesson components. The average teacher received a rating of 1.4 on a scale of 0 to 3 (a rating of 1 indicates low fidelity); ratings ranged between 0.6 and 2.5 across the 20 teachers in the study. The observed lessons varied in length from 30 to 55 minutes and on average lasted 43 minutes. A typical LLI lesson is fast-paced and involves several components, with most components being 5 to 10 minutes long. Many teachers struggled to complete all of the components in a day's lesson within the allotted time and either skipped components or completed some components in the following session. Across the 20 observations, 11 teachers were observed continuing a lesson or finishing a book from the previous day and 9 were observed skipping components altogether. For the components that were taught, the overall fidelity rating was 2.3 (a rating of 2 indicates medium fidelity).

The most commonly skipped components were phonics/word study and reading and assessment (Figure 2). Although teachers sometimes skipped phonics for timing reasons, some felt that the content was not appropriate for the needs of their older students. The assessment component presented different challenges. In addition to time constraints, some teachers reported that they struggled to keep the rest of the group on task during one-on-one assessments. Conducting these assessments also required experience or training, familiarity with LLI's online resources, and time to download and print the forms in advance. These requirements presented barriers to teachers new to LLI and even to experienced teachers with limited planning time.

Figure 2. LLI lesson fidelity



Source: Authors' calculations using data provided by OUSD.

Finally, it was common for teachers to modify the lesson guides in other ways, such as by reorganizing the components or deviating from the suggested language. For example, two teachers combined the reading and assessment component with the writing about reading component because they felt that the writing activity was a better way to keep other students occupied while they conducted one-on-one assessments.

Although most teachers were new to LLI, not all of them received sufficient training in the program. Of the 20 teachers in the study, 14 received training, typically in a two-day session led by a Heinemann representative, a half-day training led by OUSD central staff, or both. Only one of the 6 teachers who did not receive training had experience with LLI. Overall, 20 percent of teachers had experience with the program. The fidelity data suggest that these training opportunities might not have sufficiently prepared all teachers to conduct LLI as designed. During classroom observations, some teachers were found to need reminders about various aspects of implementation, including lesson timing, suggested modifications for shorter sessions, how to find and use the online resources to prepare lessons in advance, how to assess students regularly, and how to conduct the lesson components as written.

Impact findings

The results of the regression analysis indicate that LLI had no discernible impact on students' reading comprehension and a statistically significant, negative impact on their mastery of ELA/literacy standards (Table 3). The estimated effect size of being assigned to LLI on SRI scores is -0.02 standard deviations (*p*-value = 0.67), and the estimated effect size on SBAC ELA/literacy scores is -0.15 standard deviations (*p*-value = 0.01). These results are not sensitive to using unimputed baseline data and fewer covariates (Appendix Table A.4).[9]

Table 3. ITT and TOT impact estimates in effect sizes

	SRI	SBAC ELA/literacy
Impact of assignment to LLI	-0.018 (0.044)	-0.154* (0.062)
Impact of receipt of any LLI	-0.021 (0.045)	-0.163** (0.059)
Number of students analyzed	263	192

Source: Authors' calculations using data provided by OUSD.

Notes: This table displays impact estimates in z-scores (standard deviations). Standard errors are displayed in parentheses below each impact estimate.

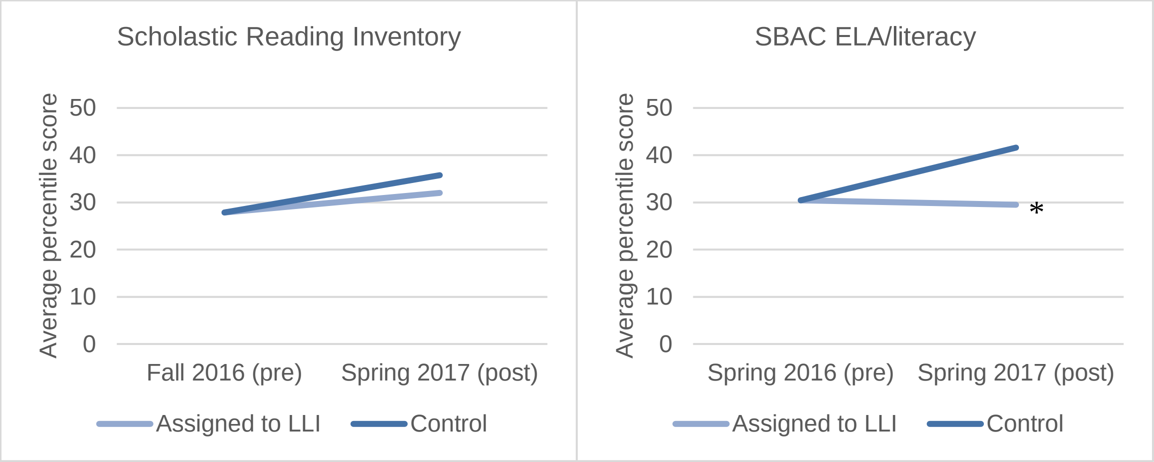
* Impact is statistically significant at the 5 percent level.

** Impact is statistically significant at the 1 percent level.

Based on the typical annual growth in reading of students in grades 6 through 8, the negative impact of being assigned to LLI on ELA/literacy scores is roughly equivalent to losing 5.5 months of learning. To convert impacts into months of learning, we divided the impact estimate by the average of the typical annual growth in reading for students in grades 6 through 8 and assumed a nine-month school year. The accuracy of this conversion depends on the extent to which learning growth on the SBAC ELA/literacy test is similar to the exams analyzed in Bloom et al. (2008). Another way to interpret these results is in terms of regression-adjusted percentile scores (Figure 3). In spring 2016, treatment and control students had an average percentile score of 30 on the SBAC ELA/literacy test. One year later, students in the control group improved their scores by 11 percentile points, on average, while the performance of students assigned to LLI decreased by 1 percentile point.

[9] Mean outcome test scores by treatment status are also reported in Table A.3 in the Appendix.

Figure 3. ITT impacts in percentile scores



Source: Authors’ calculations using data provided by OUSD.

Notes: Percentile scores range from 1 to 100 and are relative to all secondary students in OUSD. The control group averages were regression-adjusted using the estimated ITT effect sizes in Table 3.

*Impact is statistically significant at the 5 percent level.

As discussed in the implementation findings, several students in the treatment group did not participate in LLI as assigned, which could affect the program’s effectiveness. Of the 145 students in the treatment group, 83 percent received at least four sessions. There was limited crossover from the control group, with only one student assigned to control receiving LLI. TOT analyses estimating the impact of receiving LLI showed that LLI had no discernible impact on SRI scores and a statistically significant, negative impact on SBAC ELA/literacy scores (Table 3). Specifically, receiving at least four LLI sessions had an impact of -0.16 standard deviations on SBAC ELA/literacy scores (*p*-value = 0.01).

When we test for differences in impacts along student characteristics (English learner status and baseline reading level) and facets of implementation (teacher experience, instructional fidelity, and instructional setting), we find statistically significant differences only between students who received LLI in regularly scheduled classes and those who received LLI in a pullout group (Table 4). The estimated differences between these two groups are -0.21 standard deviations for the SRI (*p*-value = 0.01) and -0.27 standard deviations for the SBAC ELA/literacy test (*p*-value = 0.05), with students in pullout groups performing worse than those in regularly scheduled classes. The estimated impacts for students in regularly scheduled classes were positive for the SRI (0.06 standard deviations) and negative for SBAC ELA/literacy (-0.10 standard deviations) but were not statistically significant. The other factors we examined were not associated with different impacts, although small sample sizes limited our ability to obtain precise estimates.

Table 4. Differences in impacts by subgroup

	SRI	SBAC ELA/literacy
Impact for non-English learners (reference group)	0.026 (0.061)	-0.044 (0.084)
Difference between impacts for English learners	-0.129 (0.107)	-0.296 (0.173)
Impact for students fewer than three years behind grade level at baseline (reference group)	0.096 (0.079)	-0.176 (0.117)
Difference between impacts for students three or more years behind grade level at baseline	-0.161 (0.101)	0.036 (0.146)
Impact for students taught by an experienced LLI teacher (reference group)	0.137 (0.097)	-0.245** (0.093)
Difference between impacts for students taught by a first-year LLI teacher	-0.211 (0.112)	0.142 (0.126)
Impact for students taught LLI with lower fidelity (reference group)	-0.046 (0.072)	-0.163 (0.108)
Difference between impacts for students taught LLI with higher fidelity	0.056 (0.091)	0.017 (0.136)
Impact for students who did not receive LLI in a pullout group (reference group)	0.064 (0.057)	-0.096 (0.070)
Difference between impacts for students who received LLI in a pullout group	-0.214** (0.082)	-0.255* (0.131)
Number of students analyzed	263	192

Source: Authors’ calculations using data provided by OUSD.

Notes: This table displays impact estimates in z-scores (standard deviations). Standard errors are displayed in parentheses below each impact estimate.

* Impact is statistically significant at the 5 percent level.

** Impact is statistically significant at the 1 percent level.

Discussion

Null and mixed findings are not uncommon in evaluations of adolescent literacy interventions. As of this writing, the What Works Clearinghouse has identified research meeting standards for 21 adolescent literacy interventions. Seven interventions were determined to have no effect on reading comprehension or general literacy achievement, 13 had mixed or potentially positive effects, and only one had positive effects (READ 180). Finding effective reading interventions for secondary students who are multiple years behind grade level may be especially difficult. An experimental study of four reading interventions found that two programs (READ 180 and RISE) were effective with moderate-risk grade 9 students, but none had an impact on the achievement of grade 9 students reading below a grade 4 level (Lang et al. 2009).

The finding of a negative effect on SBAC ELA/literacy performance in this study is more unusual. A possible explanation for this result is that, by participating in LLI, some students missed grade-level content covered in the SBAC ELA/literacy assessment. This theory is supported by the fact that the negative effects were greater for students who received LLI in a pullout group. It does not appear that the results can be explained by students in the control group receiving other, more effective literacy supports that students assigned to LLI did not receive. Teachers recorded additional literacy supports outside of regular classroom instruction for only 18 percent of students in the control group.[10]

Implementation challenges may have affected the results. Despite the prevalence of null and mixed findings in the adolescent literacy literature, one reason OUSD elected to implement LLI in secondary schools is that—consistent with the experience of many of its elementary schools, which have implemented the program for a number of years—research evidence suggests that the program is effective with students in early grades. However, aside from focusing on a markedly different grade span, these earlier studies featured other important distinctions related to implementation: students received a greater number of sessions through daily lessons, instructional fidelity ratings were high, and all teachers received eight days of professional development, along with continuing support throughout the period of implementation. In contrast, fidelity to LLI's model was uneven and varied across the schools in the study.[11]

Previous research suggests that experience can be an important factor in implementation. Most LLI teachers in this study were new to the program, yet research has found that it takes teachers one year to feel confident with a new instructional approach (Fullan 2001; Hall and Hord 2001). Lack of experience may also delay improvements in student outcomes. For example, a study of four reading interventions for students in grade 5 found that the only positive impact was observed in the second year of the study, after teachers had had one year of experience in using the program (James-Burdumy et al. 2012). On the other hand, the other three intervention programs in that study were still ineffective in the second year, highlighting the prevalence of null findings in this literature.

There is also reason to think that additional professional development might improve implementation, potentially leading to better results. A systematic review of 33 studies of adolescent literacy programs determined that the approaches found to be effective provided extensive professional development and significantly changed teaching practices; further, programs designed to change daily teaching practices had greater research support than those focused on curriculum alone (Slavin et al. 2008). Like other intensive programs, LLI is fast-paced and requires nuanced judgment calls that may require more extensive teacher training. In addition, secondary literacy teachers are not typically trained in the foundational reading strategies that are part of LLI and might thus require more professional development than teachers in early grades.

There may be other factors that affect the implementation of reading interventions at the secondary level. For example, despite the greater downsides to pulling secondary students out of grade-level instruction to receive LLI, scheduling a regular LLI class can present logistical challenges for middle and high schools. In addition, because of students' low starting reading levels, most schools used LLI materials designed for students in elementary grades—which

[10] In the three most common examples, 10 control students participated in a reading pullout program using nonfiction texts, 6 control students were enrolled in an English enrichment class, and 6 control students worked with Newsela, a nonfiction personalized learning technology. It is unclear whether some students in the treatment group also received additional supports.

[11] Other studies of literacy interventions have similarly found variable implementation across schools, even despite relatively rigid implementation guidelines (e.g., Levin et al. 2010 and King 1994).

might help explain why teachers skipped some lesson components (particularly phonics) and why high school students were less engaged. Finally, according to LLI's theory of change, continuity of instruction from the regular classroom to intervention may be important to student progress, yet is more common in elementary schools. For students in this study, LLI instruction tended to be dramatically different from any other literacy instruction they received.

More broadly, the study's findings highlight the importance of assessing whether the success of intervention programs in one context can be replicated with different populations and under different conditions. Although LLI was effective with early grades in multiple studies, it did not lead to positive results for secondary students in Oakland. In addition, given that implementation varied across teachers and schools, future studies of adolescent reading interventions should consider assessing treatment effect heterogeneity in their designs. Finally, the results of this study illustrate the challenges that schools and teachers face in implementing effective reading intervention programs at the secondary level.

REFERENCES

Bloom, Howard S., Carolyn J. Hill, Alison Rebeck Black, and Mark W. Lipsey. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." *Journal of Research on Educational Effectiveness,* vol. 1, no. 4, 2008, pp. 289–328.

Fullan, Michael. *The New Meaning of Education Change* (3rd ed.). New York, NY: Teachers College Press, 2001.

Hall, Gene, and Shirley M. Hord. *Implementing Change: Patterns, Principles, and Potholes*. Boston, MA: Allyn & Bacon, 2001.

Harrison, Lynn, Rachel Peterman, Anna Grehan, Steven Ross, Emily Dexter, and Fethi Inan. "Evaluation of the Leveled Literacy Intervention: Year 1." Paper presented at the annual meeting of the American Educational Research Association, New York, NY, 2008.

Heinemann. "Fountas & Pinnell Leveled Literacy Intervention Grades 3–12, Levels L–Z Summary of Research Base." Portsmouth, NH: Heinemann, 2015a.

James-Burdumy, Susanne, John Deke, Russell Gersten, Julieta Lugo-Gil, Rebeca Newman-Gonchar, Joseph Dimino, Kelly Haymond, and Albert Liu. "Effectiveness of Four Supplemental Reading Comprehension Interventions." *Journal of Research on Educational Effectiveness*, vol. 5, no. 4, 2012, pp. 345–383.

Kamil, Michael L., Geoffrey D. Borman, Janice Dole, Cathleen C. Kral, Terry Salinger, andJoseph Torgesen. *Improving Adolescent Literacy: Effective Classroom and Intervention Practices* (NCEE Publication No. 2008-4027). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008.

King, Jennifer A. "Meeting the Educational Needs of At-Risk Students: A Cost Analysis of Three Models." *Educational Evaluation and Policy Analysis*, vol. 16, no. 1, 1994, pp. 1–20.

Lang, Laura, Joseph Torgesen, William Vogel, Carol Chanter, Evan Lefsky, and Yaacov Petscher. "Exploring the Relative Effectiveness of Reading Interventions for High School Students." *Journal of Research on Educational Effectiveness*, vol. 2, no. 2, 2009, pp. 147–175.

Levin, Henry M., Doran Catlin, and Alex Elson. "Adolescent Literacy Programs: Costs of Implementation." New York, NY: Carnegie Corporation of New York, 2010.

Odell, Kristi. "The Effect of Fountas & Pinnell's Leveled Literacy Intervention on Kindergarten Students Reading Below Grade Level." 2012 (doctoral dissertation). Available at http://www.nwmissouri.edu/library/researchpapers/2012/Odell,%20Kristi.pdf. Accessed May 29, 2018.

Peterman, Rachel, Anna Grehan, Steven Ross, Brenda Gallagher, and Emily Dexter. "An Evaluation of the Leveled Literacy Intervention Program: A Small-Group Intervention for Students in K–2." Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, 2009.

Puma, Michael, Robert Olsen, Stephen Bell, and Cristofer Price. "What to Do When Data Are Missing in Group Randomized Controlled Trials." NCEE 2009-0049. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2009.

Ransford-Kaldon, Carolyn, E. Sutton Flynt, Cristin Ross, Louis Franceschini, Todd Zoblotsky, Ying Huang, and Brenda Gallagher. "Implementation of Effective Intervention: An Empirical Study to Evaluate the Efficacy of Fountas & Pinnell's Leveled Literacy Intervention System (LLI)." Memphis, TN: Center for Research in Education Policy, 2010.

Ransford-Kaldon, Carolyn, Cristin Ross, Christine Lee, E. Sutton Flynt, Louis Franceschini, and Todd Zoblotsky. "Efficacy of the Leveled Literacy Intervention System for K–2 Urban Students: An Empirical Evaluation of LLI in Denver Public Schools." Memphis, TN: Center for Research in Education Policy, 2013.

Rubin, Donald B. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley, 1987.

Slavin, Robert E., Alan Cheung, Cynthia Groff, and Cynthia Lake. "Effective Reading Programs for Middle and High Schools: A Best-Evidence Synthesis." *Reading Research Quarterly*, vol. 43, no. 3, 2008, pp. 290–322.

Taylor, Lisa. "The Effects of Leveled Literacy Intervention for Students in the RtI Process." 2017 (doctoral dissertation). Available at https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=5063&context=dissertations. Accessed May 29, 2018.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Digest of Education Statistics. "Estimated Average Annual Salary of Teachers in Public Elementary and Secondary Schools, by State: Selected Years, 1969–70 Through 2016–17." 2016. Available at https://nces.ed.gov/programs/digest/d17/tables/dt17_211.60.asp. Accessed May 29, 2018.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 2015 Reading Assessment. Available at https://www.nationsreportcard.gov/ndecore/xplore/NDE. 2018. Accessed May 29, 2018.

What Works Clearinghouse. "Leveled Literacy Intervention Intervention Report." Washington, DC: U.S. Department of Education, Institute of Education Sciences, 2017.

Wood, William. "Literacy and the Entry-Level Workforce: The Role of Literacy and Policy in Labor Market Success." Washington, DC: Employment Policies Institute, 2010.

APPENDIX A. ADDITIONAL TABLES

Table. A.1. Baseline student characteristics, SRI analytic sample

Baseline characteristic	Treatment		Control		Standardized difference
	Mean	N (of 128)	Mean	N (of 135)	Percent
Fall 2016 F&P BAS (numeric score, 1–26)	17.3 (2.76)	119	17.9 (2.56)	126	24.8
Fall 2016 SRI (z-score)	-0.573 (0.542)	128	-0.460 (0.522)	135	21.4
Fall 2016 SMI (z-score)	-0.553 (0.836)	100	-0.337 (0.854)	115	25.6
Spring 2016 SBAC ELA/literacy (z-score)	-0.683 (0.616)	113	-0.549 (0.629)	117	21.6
Spring 2016 SBAC math (z-score)	-0.623 (0.607)	112	-0.530 (0.735)	119	13.8
African American (%)	35.9	128	34.1	135	3.9
Asian/Filipino/Pacific Islander (%)	6.25	128	9.63	135	12.5
Latino (%)	52.3	128	51.9	135	1.0
Male (%)	52.3	128	52.6	135	0.4
Eligible for free/reduced-price lunch (%)	95.3	128	92.7	135	11.0
English learner (%)	37.5	128	28.5	135	19.2
Special education (%)	13.3	128	10.2	135	9.5

Source: Authors’ calculations using data provided by OUSD.

Notes: Standard deviations are displayed in parentheses. F&P letter scores were converted to numeric scores (i.e., 17 corresponds to the letter Q). SRI, SMI, and SBAC scores were converted to z-scores defined relative to OUSD’s distribution of scores by test, grade, and administration.

None of the differences is statistically significant.

Table. A.2. Baseline student characteristics, SBAC analytic sample

Baseline characteristic	Treatment		Control		Standardized difference
	Mean	N (of 95)	Mean	N (of 97)	Percent
Fall 2016 F&P BAS (numeric score, 1–26)	16.5 (2.53)	86	17.3 (2.44)	87	28.9
Fall 2016 SRI (z-score)	-0.479 (0.513)	95	-0.398 (0.495)	97	16.1
Fall 2016 SMI (z-score)	-0.536 (0.828)	76	-0.391 (0.845)	82	17.3
Spring 2016 SBAC ELA/literacy (z-score)	-0.665 (0.626)	86	-0.532 (0.632)	87	21.1
Spring 2016 SBAC math (z-score)	-0.616 (0.607)	86	-0.559 (0.772)	89	8.2
African American (%)	0.326	95	0.299	97	5.9
Asian/Filipino/Pacific Islander (%)	0.063	95	0.072	97	3.6
Latino (%)	0.568	95	0.577	97	1.8
Male (%)	0.516	95	0.480	97	7.2
Eligible for free/reduced-price lunch (%)	0.968	95	0.949	97	9.7
English learner (%)	0.400	95	0.327	97	15.2
Special education (%)	0.137	95	0.102	97	10.7

Source: Authors' calculations using data provided by OUSD.

Notes: Standard deviations are displayed in parentheses. F&P letter scores were converted to numeric scores (i.e., 17 corresponds to the letter Q). SRI, SMI, and SBAC scores were converted to z-scores defined relative to OUSD's distribution of scores by test, grade, and administration.

None of the differences is statistically significant.

Table. A.3. Outcome test scores by treatment status

	Spring 2017 SRI			Spring 2017 SBAC ELA/literacy		
	Mean	Standard deviation	N	Mean	Standard deviation	N
Control group	-0.485	0.538	128	-0.455	0.680	97
Assigned to LLI	-0.586	0.542	135	-0.700	0.612	95
Difference	-0.101			-0.245**		

Source: Authors' calculations using data provided by OUSD.

Notes: This table displays test scores in z-scores. Each z-score represents the number of standard deviations above or below the districtwide mean score in that test, grade, and administration.

**Difference is statistically significant at the 1 percent level.

Table. A.4. ITT impact estimates in effect sizes, sensitivity analysis

	SRI	SBAC ELA/literacy
Simplified model with unimputed baseline data	-0.038 (0.044)	-0.182** (0.068)
Number of students analyzed	263	192

Source: Authors' calculations using data provided by OUSD.

Notes: This table displays impact estimates in z-scores (standard deviations). Standard errors are displayed in parentheses below each impact estimate.

**Impact is statistically significant at the 1 percent level.

ABOUT THE SERIES

Policymakers and researchers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers access to our most current work.

For more information about this paper, contact Naihobe Gonzalez, Researcher, at ngonzalez@mathematica-mpr.com.

Suggested citation: Gonzalez, Naihobe, Sophie MacIntyre, and Pilar Beccar-Varela. "Challenges in Adolescent Reading Intervention: Evidence from a Randomized Control Trial." Working Paper 62. Oakland, CA: Mathematica Policy Research, June 2018.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD