The Beaverton School District Arts for Learning (A4L) Lessons Project: An Investing in Innovation (i3) Development Grant

Student Impact Findings from Years 1, 2, and 3

Jonathan Nakamoto
Sandy Sobolew-Shubin
Martin Orland

October 22, 2015

WestEd
WestEd.org

WestEd — a national nonpartisan, nonprofit research, development, and service agency — works with education and other communities to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has 15 offices nationwide, from Washington and Boston to Arizona and California, with its headquarters in San Francisco. For more information about WestEd, visit WestEd.org; call 415.565.3000 or, toll-free, (877) 4-WestEd; or write: WestEd / 730 Harrison Street / San Francisco, CA 94107-1242.

© 2015 WestEd, All rights reserved.

Table of Contents

Executive Summary

The Beaverton School District, the third largest school district in Oregon, in partnership with Young Audiences (YA), Inc., Young Audiences Oregon and Southwest Washington, and the University of Washington, developed the Arts for Learning (A4L) Lessons Project, an intervention designed to improve students' reading and writing achievement through the integration of arts into the language arts curriculum. The A4L Lessons Project was implemented in the Beaverton School District in 2011-12 (study year 1), 2012-13 (study year 2), and 2013-14 (study year 3) with students in grades 3, 4, and 5. The A4L Lessons Project is an Investing in Innovation (i3) Development Grant that was funded by the U.S. Department of Education's Office of Innovation and Improvement (OII) in 2010.

The A4L Lessons Project involves the integration of reading, writing, and the arts, with exposure to a variety of art forms and literary genres. Students in the treatment group receive two A4L Lessons Units and one teaching artist Residency each school year. The two main elements of the program are: (1) Units of instruction, which are delivered by a classroom teacher trained by program staff, focus on a particular art form (i.e., theater, visual arts, music, or dance), and are built around one or more central texts; and (2) Residencies aligned with each A4L Unit, in which a trained teaching artist works in collaboration with the classroom teacher during five one-hour sessions. Each A4L Unit is comprised of 10 to 19 Lessons, with the suggested instructional time for the Units varying from 13 to 20 hours. In addition, the Residencies provide more concentrated focus on the study and direct experience of an art form, while also extending and reinforcing the literacy learning of the aligned Unit. Students work together in groups and practice public presentations. The A4L Lessons place an emphasis on students practicing what have been called "21st century skills," which include critical thinking, creative problem solving, and life skills, such as planning and working as a team (Seidel, Tishman, Winner, Hetland & Palmer, 2009; Silva, 2008).

To evaluate the impact of the A4L Lessons Project on students in grades 3 through 5, WestEd designed and implemented a cluster-randomized trial, randomly assigning 32 elementary schools in the Beaverton School District to receive the A4L intervention or the status-quo control condition. WestEd developed and tested three primary or confirmatory research questions and four exploratory research questions to guide the study.

Findings from the Confirmatory Research Questions

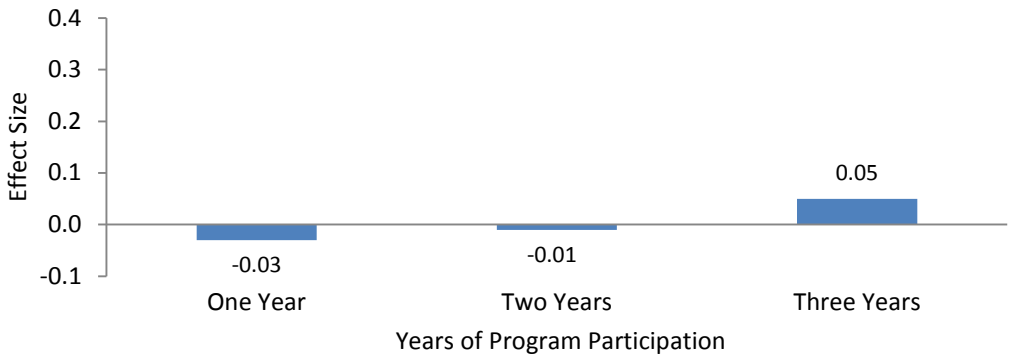
WestEd designed the evaluation of the A4L Lessons Project so that the findings from the confirmatory research questions would receive the highest rating from the What Works Clearinghouse (WWC; i.e., *Meets WWC Group Design Standards without Reservations*) and allow strong conclusions to be drawn about the program's impact. The WWC, which is a U.S. Department of Education initiative, aims to be a "trusted source of scientific evidence for what works in education" and assesses the quality of studies and their findings. Studies that receive the highest rating from the WWC provide the strongest evidence for the causal link between the intervention under study and the outcomes of interest (U.S. Department of Education, 2014). The three confirmatory research

questions relying on the Oregon Assessment of Knowledge and Skills (OAKS) Reading/Literature test are each stated below and followed by a summary of the findings.

> Three Confirmatory Research Questions: What is the impact on students' reading achievement on the OAKS after (a) one year of participation in the A4L Lessons Project, (b) two years of participation in the A4L Lessons Project, and (c) three years of participation in the A4L Lessons Project?

The results from the confirmatory analyses are shown in Exhibit E-1 and revealed no impact of the A4L Lessons Project on students' achievement on the OAKS Reading/Literature test. The differences between the treatment and control students on the OAKS after one, two, and three years of program participation were not statistically significant. The magnitude of the differences between treatment and control students, as indexed by the effect sizes, ranged from -0.03 to 0.05 and were very small. Impacts in this range, which are less than one-tenth of a standard deviation, are not considered meaningful effects by educational research standards (Lipsey et al., 2012).

Exhibit E-1. Findings from the OAKS Impact Analyses



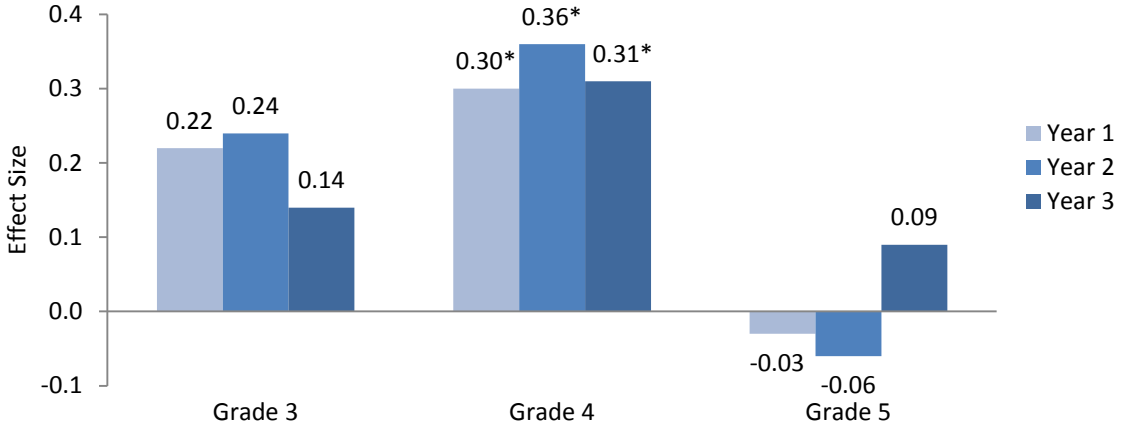
Findings from the Exploratory Research Questions

The first three exploratory research questions dealt with the impact of the A4L Lessons Project on students' life and literacy skills and relied on the Comprehensive Cross Unit (CCU) Assessments. The CCUs were developed by literacy and learning experts at the University of Washington specifically to measure the impact of the A4L Lessons supplemental literacy curriculum. Due to the fact that the CCU Assessments were administered to students in only 12 schools; six treatment schools and six control schools that participated in this portion of the study, we could only examine the impact of each year of program participation with the CCU Assessments rather than the cumulative impact of multiple years of program participation. In comparison to the confirmatory analyses, the strength of the conclusions concerning the impact of the intervention based on the exploratory analyses was much lower due to the large number of students who did not complete the CCU Assessments and the relatively small number of schools that administered these assessments. The first three exploratory questions are stated below, followed by a summary of the findings.

> Three Exploratory Research Questions: What is the impact of participating in (a) year 1, (b) year 2, and (c) year 3 of implementation of the A4L Lessons Project on the literacy and life skills of students in grades 3 through 5, as measured by the CCU Assessments?

The results from the exploratory analyses of the CCU Assessments are presented in Exhibit E-2 separated by grade and study year. There was a consistent pattern of findings with the students in grade 4. In each study year, treatment group students in grade 4 had significantly higher scores than control group students in grade 4, indicating a positive impact of the A4L Lessons Project on students' literacy and life skills as assessed by the CCUs. The effect sizes indexing the differences ranged from 0.30 to 0.36, which are considered small positive program effects by educational research standards. The students in grade 3 had higher scores than the control students in years 1 through 3 and, based on the effect sizes, the impacts would be considered small positive program impacts. However, none of the differences concerning grade 3 students were statistically significant. Finally, the differences between the treatment and control students in grade 5 were very small and not statistically significant, indicating the A4L Lessons Project did not have a reliable positive impact for students at this grade level.

Exhibit E-2. Findings from the CCU Impact Analyses, by Grade and Study Year



Note. *$p < .01$.

The final exploratory research question concerned the impacts of the A4L Lessons Project for various subgroups of students, including English language learners (ELLs) and low-income students. Given the recommendations that significant impacts for subgroups of students are more likely to be reliable when there is a significant impact for the full sample (Schochet, Puma, & Deke, 2014), WestEd conducted the exploratory subgroup analyses only when the full sample analyses with the OAKS Reading/Literature test and CCU Assessments (i.e., the previously reviewed confirmatory and exploratory analyses) were statistically significant. The specific exploratory research question is stated below, followed by the findings.

> Exploratory Research Question 4: Do the impacts on the OAKS Reading/Literature test and CCU Assessments vary by the students' ELL status or eligibility for free/reduced-price lunch?

We conducted the ELL and free/reduced-price lunch subgroup analyses with grade 4 CCU Assessments only. The results showed that the impact of the program differed significantly across ELL and non-ELL students in years 1 and 2, but not in year 3. In years 1 and 2, the impact of the A4L Lessons Project was substantially higher for ELL students, suggesting that the program had a greater impact on the literacy and life skills of ELL students. The effect sizes for ELL students were

0.87 and 0.69 compared to 0.23 and 0.31 for the non-ELL students in years 1 and 2, respectively. The effect sizes for the ELL students are considered large program impacts by educational research standards, but should be viewed extremely cautiously because the findings are based on a very small number of ELL students. In addition, the subgroup analyses demonstrated that the impact of the program on ELL and non-ELL students was nearly identical in year 3, which weakens confidence in the findings from years 1 and 2. In addition, the free/reduced-price lunch subgroup analyses based on the CCU Assessments with the students in grade 4 did not reveal any statistically significant differences between the free/reduced-price lunch and non-free/reduced-price lunch students. In other words, the results indicated that the A4L Lessons Project had an equally positive impact on the free/reduced-price lunch and non-free/reduced-price lunch students' performance on the CCU Assessments.

The findings presented in this report concern student achievement outcomes only. A fuller understanding of the implementation and outcomes of the A4L Lessons Project in the Beaverton School District can be found in the final report generated by all i3 project partners.

Introduction

The Beaverton School District, the third largest school district in Oregon, in partnership with Young Audiences (YA), Inc., Young Audiences Oregon and Southwest Washington, and the University of Washington, developed the Arts for Learning (A4L) Lessons Project, an intervention designed to improve students' reading and writing achievement through the integration of arts into the language arts curriculum. The A4L Lessons Project was implemented in the Beaverton School District in 2011-12 (study year 1), 2012-13 (study year 2), and 2013-14 (study year 3) with students in grades 3, 4, and 5. The A4L Lessons Project is an Investing in Innovation (i3) Development Grant that was funded by the U.S. Department of Education's Office of Innovation and Improvement (OII) in 2010.

Designed by YA, Inc., in partnership with researchers at the University of Washington, led by cognitive scientist Dr. John Bransford, A4L Lessons is a supplemental literacy curriculum that blends the creativity and discipline of the arts with learning science to raise student achievement in reading and writing, as well as to develop learning and life skills. The "How People Learn" framework (Bransford, Brown, & Cocking, 1999) serves as the foundation for the program's pedagogy and strategies for student engagement. It emphasizes teacher-guided, student-initiated activities, encourages students to think and learn independently, and provides tools and strategies to help students approach challenging schoolwork. The arts-integrated curricula provide students opportunities to excel in the classroom through activities that tap into a wide variety of skill sets and learning styles.

The A4L Lessons Project involves the integration of reading, writing, and the arts, with exposure to a variety of art forms and literary genres. Students in the treatment group receive two A4L Lessons Units and one teaching artist Residency each school year. The two main elements of the program are: (1) Units of instruction, which are delivered by a classroom teacher trained by program staff, focus on a particular art form (i.e., theater, visual arts, music, or dance), and are built around one or more central texts; and (2) Residencies aligned with each A4L Unit, in which a trained teaching artist works in collaboration with the classroom teacher during five one-hour sessions. Each A4L Unit is comprised of 10 to 19 Lessons, with the suggested instructional time for the Units varying from 13 to 20 hours. In addition, the Residencies provide more concentrated focus on the study and direct experience of an art form, while also extending and reinforcing the literacy learning of the aligned Unit. Students work together in groups and practice public presentations. The A4L Lessons place an emphasis on students practicing what have been called "21st century skills," which include critical thinking, creative problem solving, and life skills, such as planning and working as a team (Seidel, Tishman, Winner, Hetland & Palmer, 2009; Silva, 2008).

As part of the i3 Development Grant, WestEd evaluated the impact of the A4L Lessons supplementary literacy curriculum on students' reading and writing achievement using a cluster random-assignment design, in which 32 elementary schools were randomly assigned to the A4L Lessons treatment condition or a status-quo control condition. The impacts of the A4L Lessons were assessed by comparing the outcomes for students in the 16 intervention schools to outcomes for students in the 16 control schools, using multi-level modeling to adjust for the nesting of

students within schools. The study also examined whether high needs students (i.e., English language learners (ELLs) and economically disadvantaged students) benefited more from receiving the A4L Lessons than other students.

Research Questions

Consistent with one of the goals of the i3 program, WestEd designed the evaluation of the A4L Lessons Project so that the findings from the study's confirmatory research questions would receive the highest rating from the What Works Clearinghouse (WWC; i.e., *Meets WWC Group Design Standards without Reservations*). The WWC, which is a U.S. Department of Education initiative, aims to be a "trusted source of scientific evidence for what works in education" and assesses the quality of studies and their findings. Studies that receive the highest rating from the WWC provide the strongest evidence for the causal link between the intervention under study and the outcomes of interest (U.S. Department of Education, 2014). The evaluation design for the study's exploratory research questions were not designed to meet the WWC Standards with or without reservations. As a result, the findings from the exploratory research questions should be viewed cautiously because they cannot provide strong evidence regarding the causal link between the intervention and improved student achievement.

Confirmatory Research Questions

The current analyses, which were conducted after the third year of A4L Lessons implementation, allowed us to answer the following three confirmatory research questions concerning program impacts on students' reading achievement:

- After one year of participation in the A4L Lessons Project, what is the impact on students' reading achievement, as measured by the Oregon Assessment of Knowledge and Skills (OAKS) Reading/Literature test?[1]
- After two years of participation in the A4L Lessons Project, what is the impact on students' reading achievement, as measured by the OAKS Reading/Literature test?[2]
- After three years of participation in the A4L Lessons Project, what is the impact on students' reading achievement, as measured by the OAKS Reading/Literature test?[3]

Exploratory Research Questions

We examined the following exploratory research questions concerning impacts on the students' life and literacy skills:

---

[1] This analysis included students who were in grades 3, 4, and 5 during year 1 of the study and students who were in grade 3 during years 2 and 3 of the study.

[2] This analysis included students who started in grades 3 and 4 in year 1 of the study and were followed into year 2 and students who started in grade 3 in year 2 of the study and were followed into year 3.

[3] This analysis included students who started in grade 3 in year 1 and were followed into year 3.

- What is the impact of participating in year 1 of implementation of the A4L Lessons Project on the literacy and life skills of students in grades 3 through 5, as measured by the Comprehensive Cross Unit (CCU) Assessments?

- What is the impact of participating in year 2 of implementation of the A4L Lessons Project on the literacy and life skills of students in grades 3 through 5, as measured by the CCU Assessments?

- What is the impact of participating in year 3 of implementation of the A4L Lessons Project on the literacy and life skills of students in grades 3 through 5, as measured by the CCU Assessments?

In addition, when the impact analyses with the OAKS Reading/Literature test and CCU Assessments showed significant impacts on the students' literacy and life skills for the full sample, we asked the following exploratory research question about student subgroups:

- Do the impacts on the OAKS Reading/Literature test and CCU Assessments vary by the students' ELL status and eligibility for free/reduced-price lunch?

Methodology

Overview

WestEd's cluster-randomized experimental trial relied on the random assignment of schools to treatment or control conditions, permitting causal inferences to be drawn about the impact of the A4L Lessons on students' reading and writing achievement. At the beginning of the study, 32 of the 33 elementary schools in the Beaverton School District in Oregon were randomly assigned to either treatment or control conditions. Given that the number of participating schools was relatively small, we employed a matched random assignment procedure to ensure that the process resulted in equivalent treatment and control groups. Since only students in regular third-, fourth-, and fifth-grade classrooms participated in the A4L Lessons Project during 2011-12, 2012-13, and 2013-14, we excluded students in grades 3 through 5 in self-contained special education classrooms from the analyses.

We obtained the participating students' scores on the OAKS Reading/Literature test, the Developmental Reading Assessment®–2nd edition (DRA2), and the CCU Assessments from the district. We additionally obtained student- and school-level demographic data from the district to include as covariates in our analyses to improve the precision of impact estimates. In order to appropriately account for the multi-level structure of the data (i.e., students nested within schools), we utilized hierarchical linear modeling (HLM) to estimate the program impacts on the OAKS and CCU Assessments. We also used HLM to conduct the subgroup analyses to calculate impact estimates for ELLs and students eligible for free/reduced-price lunch, as well as to conduct baseline equivalence testing.

Sample Selection and Assignment

Selection and Random Assignment of Schools

Given the small number of schools participating in the study, we employed a matched random assignment procedure to increase the likelihood of group equivalence on five key predictor (nuisance) variables commonly associated with student achievement. These variables included: (1) school enrollment; (2) percentage of ELL students; (3) percentage of students who participate in the free/reduced-price lunch program; (4) student race/ethnicity; and (5) student achievement as measured by the OAKS Reading/Literature test and the OAKS Grade 4 Writing test, which has since been discontinued.

The matched random assignment procedure was accomplished in two stages. First, the 32 schools were matched into 16 pairs according to their similarity as defined by the variables delineated above. Next, one of the schools comprising each matched pair was randomly assigned to the treatment group, while the other school was assigned to the control group.

To measure the similarity of the 32 schools, we used a mathematical algorithm embedded in the cluster analysis module of the Statistical Package for the Social Sciences (SPSS) 18 to calculate each

school's Euclidean distance from the minimal point in the ordinal space composed of the predictor variables described above. Generally speaking, Euclidean distance is a geometric distance between two data points (or cases) and it is defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

where d_{ij} is the distance between cases i and j, x_{ik} is the value of kth variable for the ith case, x_{jk} is the value of kth variable for the jth case, and p is the total number of variables used. The equation illustrates that the Euclidean distance is a measure in the form of square root, with the smaller the Euclidean distance between two points indicating greater similarity between cases. Before taking the square root, we summed the squared differences for all the variables for the two cases measured. By using the squared differences, we avoided having to work with negative distances.

Three steps were taken to implement this algorithm and group the 32 schools into 16 pairs. First, all the variables described above were standardized or converted into a common metric/scale. Second, we specified a common reference point to which the 32 schools could be compared. In this study, the minimum values of all the nuisance variables were listed and fed into the algorithm as the reference point. The Euclidean distances between this common reference and each of the 32 schools were calculated and rank ordered from the lowest to the highest value. Finally, based on these ordered 32 distances, we paired every two adjacent schools starting from the two lowest values, ultimately resulting in 16 pairs. Within each pair, the school with the shorter distance from the reference point was given a code of "1" and the other school was given a code of "2." This within-pair coding system was created for the next stage of the procedure, the random assignment of one school in each pair to the treatment group.

In the second stage of the procedure, we separated the schools with different within-pair codes (either 1 or 2) into two columns. We then generated a random uniform variable for the 16 pairs with the values of the variable ranging from a minimum value of 0 to a maximum value of 1. For a particular pair, if the value of the random uniform variable was equal to or less than 0.5, we assigned a pointer of "<" to the pair; if the value of the random variable was greater than 0.5, we assigned a pointer of ">" to the pair. Next, the school within each pair facing the narrow end of the pointer (as determined by a coin toss) was assigned to the treatment group. The final results are provided in Exhibit 1 with schools in the treatment group listed on the left and schools in the control group listed on the right.

Finally, statistical tests were conducted to assess the equivalence of the treatment and control groups resulting from the matched random assignment procedure. A series of *t*-tests were conducted with group membership (treatment versus control) as the independent variable and each of the nuisance variables as the dependent measure. No statistically significant mean differences emerged between treatment and control groups (see Exhibit A-1 in the Appendix). Similarly, between group variances were not different for any of the dependent measures.

Exhibit 1. Schools in the Treatment and Control Groups

Treatment		Control	
School ID	School Name	School ID	School Name
1155	Beaver Acres Elementary School	1153	Aloha-Huber Park School
1156	Bethany Elementary School	1154	Barnes Elementary School
1159	Chehalem Elementary School	4671	Bonny Slope Elementary School
1160	Cooper Mountain Elementary School	1162	Elmonica Elementary School
1370	Findley Elementary School	1161	Errol Hassell Elementary School
1163	Fir Grove Elementary School	1157	Greenway Elementary School
1166	Kinnaman Elementary School	1164	Hazeldale Elementary School
1168	McKay Elementary School	1165	Hiteon Elementary School
1169	McKinley Elementary School	3437	Jacob Wismer Elementary School
1303	Nancy Ryles Elementary School	1170	Montclair Elementary School
1173	Raleigh Park Elementary School	1171	Oak Hills Elementary School
1174	Ridgewood Elementary School	1172	Raleigh Hills Elementary School
1175	Rock Creek Elementary School	2781	Scholls Heights Elementary School
4712	Springville K-8 School	1270	Sexton Mountain Elementary School
1176	Terra Linda Elementary School	1177	Vose Elementary School
1179	William Walker Elementary School	1178	West Tualatin View Elementary School

Identification of Eligible Students

In order to draw causal inferences about the impact of the A4L Lessons Project on individual students, the students included in the confirmatory analyses had to be enrolled in the participating schools prior to random assignment, which took place in January 2011 (Price, 2014). Specifically, the students in grades 3, 4, and 5 in 2013-14 had to be enrolled in the participating schools in 2010-11 (i.e., the year prior to the start of the intervention) when they were in kindergarten, grade 1, and grade 2, respectively. The district provided the enrollment rosters from October 2010 so that WestEd could identify these students. The students who enrolled in the participating schools after 2010-11 were classified as "joiners" and are included in the exploratory analyses with the CCU Assessments. The CCU analyses included all students in the six treatment schools who received the CCU pre-test and the post-test, regardless of whether they were enrolled in the schools prior to random assignment. Because the intervention was not delivered in self-contained special education classrooms, we excluded students from the analyses who received less than 40 percent of their instruction in regular classrooms during 2011-12, 2012-13, and 2013-14 prior to conducting the analyses.

Measures

Student Demographic Data

Data on student demographic characteristics were obtained from the district in the summer 2012, 2013, and 2014 for the 2011-12, 2012-13, and 2013-14 school years, respectively. Student characteristics included eligibility for free/reduced-price lunch, ELL status, and race/ethnicity (e.g., Asian, Hispanic/Latino, and Multi-Racial) among other variables. We included the student

characteristics in our models as covariates to improve the precision of impact estimates. In addition, we used the ELL status and free/reduced-price lunch status as subgroup indicators in the exploratory analyses concerning the differential impacts of the A4L Lessons intervention on student subpopulations.

Oregon Assessment of Knowledge and Skills (OAKS) Reading/Literature Test

The OAKS is the Oregon state accountability test, a criterion-reference test without normed scores. During 2010-11 and 2011-12, it was administered to students up to three times per year, beginning in January and ending in May. Once students reached the highest level of proficiency (i.e., "Exceeds"), they were locked out of further testing. Consistent with the state's methodology used for reporting purposes, the students' best score was used in the analyses. The administration of the OAKS changed slightly in subsequent years. During 2012-13 and 2013-14, the Reading/Literature test was administered to students once or twice between January and May. If a student received a score of "Meets" or "Exceeds" on the first administration of the OAKS in these two years, they could not be retested without the consent of their parent. If a student did not receive a score of "Meets" or "Exceeds," the district could retest them once later in the school year after the student had been provided with additional instruction (Oregon Department of Education, 2012). Consistent with the state's methodology used for reporting purposes, the best score was used in the analyses of student impacts. Also, students were classified as below grade-level readers if their scale score on the OAKS Reading/Literature test did not place them into the "Meets" or "Exceeds" categories. For students in grades 3 and 4 in 2010-11, we utilized their OAKS scores from 2010-11 as covariates in the models. We used the OAKS scores from 2011-12, 2012-13, and 2013-14 as the outcome measures in the confirmatory analyses for all students.

The goal of the A4L Lessons supplementary literacy curriculum is to improve students' reading and writing achievement, making the OAKS Reading/Literature test an appropriate outcome measure not over-aligned with the intervention. The OAKS Reading/Literature test contained up to eight Score Reporting Categories (SRCs). Students in grades 3 through 5 completed the following six SRCs: (1) SRC 1 – Vocabulary; (2) SRC 2 – Read to perform a task; (3) SRC 3 – Demonstrate general understanding; (4) SRC 4 – Develop an interpretation; (5) SRC 8 – Reading informational text; and (6) SRC 9 – Reading literary text. Students in grades 4 and 5 additionally complete SRC 5 – Examine content and structure: Informational text. Furthermore, students in grade 5 complete SRC 6 – Examine content and structure: Literary text. Descriptions of each SRC are included in Exhibit A-2 in the Appendix. The scores from the different SRCs were very highly correlated with one another and the total score on the OAKS. Additionally, our analyses from years 1, 2, and 3 of the study showed no differences in the pattern of findings across the total score and the individual SRCs. As a result of these findings and the need to limit the number of outcome variables for the confirmatory analyses (Schochet, 2008), we utilized only the OAKS total score in the current analyses.

Developmental Reading Assessment – 2nd Edition (DRA2)

The Beaverton School District administers the DRA2 to all second graders. The DRA2, published by Pearson, is a valid measure of reading accuracy, fluency, and comprehension as evidenced by its

criterion, construct, and content validity. With students in grades 1 through 3, the DRA2 exhibited correlations between $r = .60$ and $r = .74$ with other reading tests, such as the Gray Oral Reading Test. Additionally, the test-retest reliability of the DRA2 across a two-week period for students in grades 1 through 3 ranged from $r = .97$ to $r = .99$ (Pearson Education Inc., 2009). For the confirmatory analyses, we utilized the DRA2 scores as covariates for students who were in grade 2 during 2010-11, 2011-12, and 2012-13. Consistent with the district's methodology for categorizing students, we classified students as below grade-level readers if they scored below 28 on the DRA2 at the end of grade 2. The district tested the vast majority of students on the DRA2 at their independent reading level and these scores are available in the district's data warehouse. The very small number of students who were tested at their instructional level, but did not have scores in the data warehouse were not included in the analyses.

Comprehensive Cross-Unit (CCU) Assessments

Students' literacy and life skills were measured by the Comprehensive Cross Unit (CCU) Assessments, which were developed specifically for the A4L Lessons supplemental literacy curriculum by Dr. Diana Sharp and informed by learning and literacy experts at the University of Washington.[4] In all three years, the Joy Test was administered to students in grade 3 and the Ruth Test was administered to students in grade 4. In 2011-12 and 2012-13, the Ruth Test was administered to students in grade 5, while the Jackie Test was administered to students in grade 5 during 2013-14.

These assessments were constituted by a set of items that focus exclusively on the development of Cross-Unit skills, rather than items that tap Unit-Specific skills. Cross-Unit items assess DEEP skills (Decision Enhanced by Empathy and Perspective) and include describing a character's or the author's traits, emotions, thoughts, or internal motivations based on a text, as well as analyzing how a character's or author's perspective impacts other key genre-specific elements (e.g., the problem, events, and resolution in a story; visual representations in a graphic novel; or the mood or feelings in a poem). By contrast, Unit-Specific items assess students' skills specific to an A4L Lessons Unit such as the ability to analyze the structural elements of a story (e.g., protagonist, overarching problems, events, resolution, and setting), make inferences to create meaning, identify the theme in a novel, or identify and describe images from a poem.

Pre-test CCU data were collected before A4L Lessons implementation in the fall of each school year, while post-test CCU data collection occurred after A4L Lessons implementation in the spring of each year and after the teaching artist Residency. Six treatment and control school pairs were randomly selected to administer the CCUs in grades 3, 4, and 5 for the three-year study.

The CCUs ask students to respond to open-ended questions about literature selections and assess literacy achievement, as well as 21st century learning and life skills. The three CCUs are scored with

[4] Literacy expert, Diana Sharp, and learning experts from the University of Washington, John Bransford, Nancy Vye, and Allison Moore developed the assessments. These individuals were also members of the University of Washington team that developed the curriculum units.

similar rubrics. The rubric for the Joy Test has nine criteria, while the Ruth and Jackie Tests have 12 criteria each The criteria for the tests are rated using 0 (e.g., *does not make sense*) to 2 (e.g., *mentions what others were thinking*), 0 (e.g., *0 traits*) to 4 (e.g., *4 traits*), 0 to 6, and 0 to 9 scales. Dr. Leslie Murrill, Professor in the Roanoke College (VA) School of Education and an expert in elementary literacy instruction, oversaw the scoring of CCU Assessments.

We assessed the internal reliability of the Joy, Ruth, and Jackie Tests using Cronbach's alpha at the pre-test and post-test in each year of the study. The Joy Test showed acceptable reliability (John & Benet-Martinez, 2000) and the Cronbach's alphas ranged from $\alpha = .70$ to $\alpha = .75$. In addition, the Ruth Test showed good reliability and the Cronbach's alphas ranged from $\alpha = .77$ to $\alpha = .82$. Similarly, in year 3, the Jackie Test showed good reliability at pre-test ($\alpha = .80$) and post-test ($\alpha = .79$).

Validity studies have not been conducted on the CCU Assessments and no parallel forms of the test are in use. Using the current study's sample, however, we obtained estimates of convergent validity by correlating the CCU Assessments with the students' scores on the 2014 OAKS Reading/Literature test. These correlations showed the CCU Assessments and OAKS Reading/Literature test measured related but not exactly the same constructs. In grade 3, the correlations between the Joy pre-/post-tests and the OAKS Reading/Literature total score were $r = .63$ and $r = .62$, respectively. For fourth graders, the correlations between the Ruth pre-/post-tests and the OAKS Reading/Literature total score were $r = .66$ and $r = .65$. For fifth graders, the correlations between the Jackie pre-/post-tests and the OAKS Reading/Literature total score were $r = .68$ and $r = .67$. Correlations in these ranges are indicative of convergent validity.

School-Level Achievement and Demographic Data

Data on school characteristics prior to the start of the program were obtained from the district in September 2012. School characteristics included: (1) school enrollment in 2011-12; (2) percentage of ELL students in 2010-11; (3) percentage of students who participated in the free/reduced-price lunch program in 2010-11; (4) percentage of students who were racial/ethnic minorities in 2010-11; and (5) percentage of students who were proficient (i.e., "Meets" or "Exceeds") on the OAKS Reading/Literature test in 2010-11. With the exception of school enrollment, all of these school-level variables were included in the confirmatory analysis models as covariates to improve the precision of estimates.

Data Analyses

The use of a cluster-randomized design in which students are nested within schools necessitates the use of an analytic technique that can account for multi-level data structures, making statistical corrections to account for the nesting (Raudenbush & Bryk, 2002). The primary analyses for this study involved fitting conditional multi-level regression models (i.e., HLM models; Murray, 1998; Raudenbush & Bryk, 2002). Random effects of school site were included in the models to account for the nesting of students within schools. Fixed effects included treatment status, baseline (pre-test) measures of reading achievement (e.g., DRA2), dummy codes representing the randomization strata, and other student-level (e.g., ELL status) and aggregate school-level (e.g., the percentage of each

school that qualified for free/reduced-price lunch) covariates. All of the predictor variables included as fixed effects in the models for the confirmatory analyses are outlined in Exhibit 2. Although random assignment reduces the likelihood of non-equivalence between treatment and control groups, there is always the possibility that non-equivalence will occur by chance alone. The purpose of including the covariates was to minimize random error and to increase the precision of the impact estimates.

Analysis of Program Impacts on OAKS Scores

Exhibit 2. Predictor Variables Used in the OAKS Confirmatory Analyses

Variable	Description
School-level	
Treatment status	1 = assigned to treatment group; 0 = assigned to control group
Dummy codes for the strata	15 dummy coded variables representing the 16 pairs of schools used in the random assignment process
Prior OAKS Reading/Literature test	The percentage of students scoring meets and above on the OAKS Reading/Literature test from 2010-11
ELL percentage	The percentage of each school that was ELL in 2010-11
Free/reduced-price lunch percentage	The percentage of each school that qualified for free/reduced-price lunch in 2010-11
Minority percentage	The percentage of each school that was a racial/ethnic minority in 2010-11
Student-level	
Baseline reading achievement	Scores on the DRA2 or OAKS from the year prior to the students' first year of program participation
ELL status	1 = ELL; 0 = not ELL or former/exited ELL
Free/reduced-price lunch status	1 = qualified for free/reduced-price lunch; 0 = did not qualify for free/reduced-price lunch
Below grade-level reading status	1 = grade-level reader or above; 0 = below grade-level reader
Race/Ethnicity	Four dummy coded variables contrasting African American, Asian/Pacific Islander, Hispanic/Latino, and Other (i.e., Multi-racial and Native American/Alaskan Native) with White

The confirmatory analyses with the OAKS Reading/Literature test included all students in the treatment and control schools, regardless of the extent to which they participated in the program. The confirmatory analyses are termed Intent-to-Treat (ITT) analyses because the students are analyzed in the treatment condition to which they were originally assigned even if they changed conditions later in the study (Shadish, Cook, & Campbell, 2002). The confirmatory analyses included all students who were in the participating schools in 2010-11 and remained in the study until the post-test measure for the analyses was administered in 2011-12, 2012-13, or 2013-14, regardless of whether they moved between treatment and control schools or left the district at some point during the three years of the study.

Exhibit 3 shows the cohorts of students included in the three confirmatory analyses across the pre-intervention year and the three years of program implementation. The one-year participants (i.e., students who participated in the program for one year) included students who were in grades 3, 4,

and 5 during year 1 of the study and students who were in grade 3 during years 2 and 3 of the study. The two-year participants included students who started in grades 3 and 4 in year 1 of the study and were followed into year 2 and students who started in grade 3 in year 2 of the study and were followed into year 3. Finally, the three-year participants included students who started in grade 3 in year 1 and were followed into year 3. As shown in the exhibit, some students were included in multiple analyses. For example, students in grade 2 in 2010-11 were included in all three confirmatory analyses because they could be used to evaluate the impact of the program after one, two, and three years of participating in the intervention.

Exhibit 3. Tracking the Cohorts Included in the OAKS Confirmatory Analyses across Years

	Pre- Intervention 2010-11	Implementation Year 1 2011-12	Implementation Year 2 2012-13	Implementation Year 3 2013-14
One- year participants	grade 2	**grade 3**	-	-
One- year participants	grade 3	**grade 4**	-	-
One- year participants	grade 4	**grade 5**	-	-
One-year participants	grade 1	grade 2	**grade 3**	-
One-year participants	kindergarten	grade 1	grade 2	**grade 3**
Two-year participants	grade 2	**grade 3**	**grade 4**	-
Two-year participants	grade 3	**grade 4**	**grade 5**	-
Two-year participants	grade 1	grade 2	**grade 3**	**grade 4**
Three-year participants	grade 2	**grade 3**	**grade 4**	**grade 5**

Note. The bolded cells indicate the years the cohorts received the intervention.

The following equation illustrates the two-level HLM model we used to assess the overall impact of the A4L curriculum on the OAKS Reading/Literature test:

$$Outcome_{ijk} = \alpha_0 + \beta_1 Treatment_{jk} + \sum \beta_I I_{ijk} + \sum \beta_T S_{jk} + \sum \beta_{ST} Strata_k + \tau_{jk} + \varepsilon_{ijk}$$

where subscripts *i*, *j*, and *k* denote student, school, and randomization strata, respectively. *Outcome* represents the OAKS scores for students nested in schools nested in strata. The OAKS scores used for the outcomes were standardized within grade and year using the state-level means and standard deviations. Using this technique, a score of 0.25 represents an OAKS score that is 0.25 standard deviations above the state mean, which is approximately the 60th percentile. Treatment is a dichotomous variable that indicates school assignment to treatment and control groups.

I represents a vector of student-level control variables. For all three confirmatory analyses, the models included dummy coded variables contrasting ELLs with non-ELLs, free/reduced-price lunch students with non-free/reduced-price lunch students, and whites with Asian/Pacific Islanders, African Americans, Hispanics, and multiple race students/Native Americans.

For the one-year impact analyses, the vector of student-level control variables had a pre-test measure that included the DRA2 (z-scored for each cohort and grade level) or the OAKS (z-scored for each cohort and grade level) in a single variable. The model included dummy coded variables contrasting students who had a DRA2 score used as the pre-test (i.e., students in grade 3 during the intervention

year) with students who had a grade 3 OAKS score used as the pre-test (i.e., students in grade 4 during the intervention year) and with students who had a grade 4 OAKS score used as the pre-test (i.e., students in grade 5 during the intervention year). In addition, the model included interaction terms between the dummy coded variables representing the pre-test measures and the score for the pre-test variables. This technique was equivalent to a linear spline model (Marsh & Cormier, 2002) and accounted for the differences in how the tests were scored. Finally, the model included a dummy coded variable contrasting above and below grade-level readers based on the cut-off used by the district for the DRA2 (i.e., students with scores above 24 were classified as grade-level readers or above) and for the OAKS (i.e., the scale score cut-offs used by the state to classify students as proficient).

For the two-year impact analyses, the vector of student-level control variables had a pre-test measure that included the DRA2 (z-scored for each cohort and grade level) or the OAKS (z-scored for the one cohort) in a single variable. The model included a dummy coded variable contrasting students who had a DRA2 score used as the pre-test (i.e., students in grades 3 and 4 during the intervention) with students who had a grade 3 OAKS score used as the pre-test (i.e., students in grades 4 and 5 during the intervention). In addition, the model included an interaction term between the dummy coded variable representing the pre-test measures and the score for the pre-test variables. Finally, the model included a dummy coded variable contrasting above and below grade-level readers based on the cut-off used by the district for the DRA2 (i.e., students with scores above 24 were classified as grade level readers or above) and for the OAKS (i.e., the scale score cut-offs used by the state to classify students as proficient).

For the three-year impact analyses, the vector of student-level control variables had the DRA2 from grade 2 and a dummy coded variable contrasting above and below grade-level readers based on the cut-off used by the district (i.e., students with scores above 24 were classified as grade-level readers or above).

S is a vector of school-level control variables. The vector included the school-level OAKS reading proficiency rate from spring 2011. It also included the percentage of each school's student population in 2010-11 that was ELL, was comprised of racial/ethnic minorities, and qualified for free/reduced-price lunch.

Strata is a set of 15 dichotomous variables representing fixed effects for strata. Lastly, τ_{jk} represents a random effect for schools (clustering groups), and ε_{ijk} is an error term for individual sample members. In this model, the intervention effect is represented by β_1.

Analysis of Program Impacts on CCU Scores

The exploratory CCU analyses with the students in grades 3, 4, and 5 from years 1 through 3 assessed the impact of one year of program participation by using the pre-test CCU from the same year as a control variable. For example, even though the students in grade 5 in year 3 participated in the intervention in two previous years, the use of the pre-test CCU from year 3 meant that the analyses controlled for any prior program effects and assessed the impact of only the last year of program participation.

Our sample inclusion criteria for the CCUs analyses differed from the criteria used for the confirmatory analyses. The confirmatory analyses included only students enrolled in the participating schools in the pre-intervention year (i.e., 2010-11), which allowed us to draw the strongest possible conclusions about the impact of the intervention on students rather than schools (Price, 2014). In contrast, we included all students with pre- and post-test CCU data in the analyses regardless of whether they were enrolled in one of the CCU schools in the pre-intervention year. Our inclusion criteria for the CCU analyses allowed us to include more students in the CCU analyses, which increased our statistical power. Additionally, the CCUs are resource intensive to administer and score so it was appropriate not to exclude students with complete CCU data from the analyses because they were not enrolled in the schools during the pre-intervention year. However, this inclusion criterion would not permit the findings to meet *Meets WWC Group Design Standards without Reservations* and allow for drawing causal conclusions concerning the impact of the intervention on individual students.

Given that the CCU Assessments were administered to students in only 12 schools, HLM models relying on these data would have an increased likelihood of not converging or producing inadmissible solutions because of the small number of level-2 units (Maas & Hox, 2005). Given these potential issues, we opted to remove all of the school-level variables that were included in the OAKS analyses with the exception of the treatment status variable in order to create more parsimonious models that were more likely to converge and less likely to produce inadmissible solutions. Instead of the DRA2 or OAKS as the pre-test reading measure, we used the fall CCUs as the pre-test measure. In addition, we excluded the below grade-level reading status as a control variable because there is no grade-level cut-off on the CCUs. The remaining student-level control variables from the OAKS analyses (i.e., ELL status, free/reduced-price lunch status, and race/ethnicity) were included in the CCU analyses.

The following equation illustrates the two-level HLM model we used to assess the overall impact of the A4L curriculum on the CCU Assessments:

$$Outcome_{ij} = \alpha_0 + \beta_1 Treatment_{jk} + \sum \beta_I I_{ijk} + \tau_{jk} + \varepsilon_{ijk}$$

where subscripts *i* and *j* denote the student and school levels in the models. *Outcome* represents the post-test CCU scores for students nested in schools. Treatment is a dichotomous variable that indicates school assignment to treatment and control groups.

I represents a vector of student-level control variables. Consistent with the OAKS analyses, the models included dummy coded variables contrasting ELLs with non-ELLs, free/reduced-price lunch students with non-free/reduced-price lunch students, and whites with Asian/Pacific Islanders, African Americans, Hispanics, and multiple race students/Native Americans. The models also included the pre-test CCU score from the fall of each school year. Finally, τ_{jk} represents a random effect for schools (clustering groups), and ε_{ijk} is an error term for individual sample members. In this model, the intervention effect is represented by β_1.

We used $p = .05$ as the threshold for statistical significance. We applied the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) to protect against Type I errors (i.e., false-positive

findings). For example, using the Benjamini-Hochberg correction, the resulting *p* values from a domain with nine contrasts (i.e., three grade levels and three years) are ordered from lowest to highest, with the thresholds for statistical significance starting at .006 and moving downward to .05.

Subgroup Analyses

We conducted exploratory subgroup analyses with the CCU Assessments to calculate impact estimates for ELL students and those students eligible for free/reduced-price lunch. These subgroup analyses also allowed us to determine whether the impact of the intervention differed across various subgroups. For example, we assessed whether the impact of the intervention was stronger for ELL students in comparison to non-ELL students. To conduct the subgroup analyses, we created subgroup × treatment status interaction terms and included these interaction terms as predictors in the models. It should be noted that the results of these subgroup analyses should be viewed extremely cautiously because of the small number of students in the subgroups of interest and the unequal distribution of the subgroups across the participating schools.

Baseline Balance Testing for the OAKS Analyses

For each confirmatory analysis, we tested the baseline equivalence of the pre-test measures for the sample. In accordance with the What Works Clearinghouse (WWC; 2014) Procedures and Standards Handbook, the analysis sample for a particular contrast in this study was defined as the set of treatment and control students who had non-missing values on the outcome variables.

The analysis models for the baseline balance testing for the confirmatory analyses used the same structural components as the statistical models used to estimate intervention impacts on the outcome variables. In other words, the models used to test for baseline equivalence had the same two-level structure with students nested within schools and strata. Given that blocks of school pairs were formed before random assignment and the impact models used dummy variables for the blocks, we included the same dummy variables for blocks in the models for baseline balance testing. The following example of a two-level HLM model demonstrates the type of analyses we used for determining the baseline equivalence of the pre-test achievement measures:

$$Pretest_{ijk} = \alpha_0 + \beta_1 Treatment_{jk} + \sum \beta_{ST}\, Strata_k + \tau_{jk} + \varepsilon_{ijk}$$

where subscripts *i*, *j*, and *k* denote student, school, and randomization strata, respectively. Pre-test represents the baseline assessment scores for students nested in schools nested in strata. For the one- and two-year samples, the DRA-2 or OAKS were z-scored within cohort and grade level. The raw DRA2 scores were used for the three-year sample. *Treatment* is a dichotomous variable that indicates school assignment to treatment and control groups. *Strata* is a set of 15 dichotomous variables representing fixed effects for strata; and τ_{jk} represents a random effect for schools (clustering groups), and ε_{ijk} is an error term for individual sample members. For the one-year analysis, the model additionally included two dummy coded variables contrasting students who had the DRA2 score as the pre-test measure with students who had the grade 3 and grade 4 OAKS as the pre-test. For the two-year analysis, the model also included a dummy coded variable contrasting students who had the DRA2 score as the pre-test measure with students who had the grade 3

OAKS as the pre-test. Fitting the above model to the data produced an estimate for β_1, which represents the estimated treatment-control difference in the pre-test measure.

When reporting the results of baseline equivalence testing for the pre-test achievement measures, we followed the criteria described in the WWC (2014) Procedures and Standards Handbook. For each baseline equivalence test, we calculated the ratio between the estimated treatment-control difference in the pre-test measure and the pooled standard deviation of the treatment and control groups (i.e., the effect size for the difference between the groups). In addition, for the baseline balance testing with the demographic characteristics (i.e., categorical variables), we utilized multi-level logistic regression models that included only the treatment status variable as a predictor variable because several models that included the dummy codes representing the strata failed to converge.

Baseline Balance Testing for the CCU Analyses

For each sample included in the exploratory CCU analyses, we conducted baseline equivalence analyses with the pre-test measures. The analysis models for the baseline balance testing for the exploratory CCU analyses used the same structural components as the statistical models used to estimate intervention impacts on the outcome variables. The following example of a two-level HLM model shows the type of analyses we used for determining the baseline equivalence of the pre-test achievement measures:

$$Pretest_{ij} = \alpha_0 + \beta_1 Treatment_{jk} + \tau_{jk} + \varepsilon_{ij}$$

where subscripts *i* and *j* denote student and school, respectively. Pre-test represents the baseline assessment scores for students nested in schools. *Treatment* is a dichotomous variable that indicates school assignment to treatment and control groups. τ_{jk} represents a random effect for schools (clustering groups) and ε_{ijk} is an error term for individual sample members. Fitting the above model to the data produced an estimate β_1, which represents the estimated treatment-control difference in the pre-test measure. Finally, for the baseline balance testing with the demographic characteristics, we utilized multi-level logistic regression models that included only the treatment status variable as a predictor variable.

Treatment of Missing Data

WestEd removed all students from the analyses with missing pre-test and post-test data on the reading and writing measures. The numbers of students who were removed from the confirmatory analyses are outlined below in the section on attrition. A small number of students had to be excluded from the analyses due to missing demographic data. Our missing data strategy is consistent with the WWC (2014) recommendations for conducting impact analyses and assessing baseline equivalence.

Attrition for the OAKS Reading/Literature Test Analyses

At the school-level, all 16 treatment and 16 control schools that were randomly assigned in January 2011 remained in the study through the end of 2013-14. Exhibit 4 displays the number of students

randomized (i.e., enrolled in the school in 2010-11) by treatment condition for the one-, two-, and three-year analyses. The number of students excluded from the analyses that did not have a pre-test score or a post-test score because they left the district or were not tested is shown in Exhibit 4. Additionally, the exhibit displays the number of students who were excluded from the analysis samples because they were missing demographic data or received less than 40 percent of their instruction in regular classrooms in years 2 or 3 of the study.[5] The attrition rates for the treatment and control students ranged from 19.2 percent to 26.9 percent across the analysis samples. According to the WWC guidelines (U.S. Department of Education, 2014), these attrition rates are likely to produce an acceptable level of bias and would not prevent the study's findings from meeting the *WWC Group Design Standards without Reservations*.

Exhibit 4. Number of Students Included in and Excluded from the OAKS Confirmatory Analyses and the Attrition Rates

	Years of Program Participation					
	One Year		Two Years		Three Years	
	Treatment	Control	Treatment	Control	Treatment	Control
Students randomized	7,299	7,559	4,383	4,627	1,474	1,630
Students missing post-test scores	1,364	1,359	980	963	355	390
Students missing pre-test scores	210	80	74	66	38	27
Students missing demographic data	6	10	0	0	0	0
Special education students	0	0	4	8	4	5
Students include in impact analyses	5,719	6,110	3,325	3,590	1,077	1,208
Attrition rate	21.6%	19.2%	24.1%	22.4%	26.9%	25.9%

Note. The special education students received less than 40 percent of their instruction in regular classrooms in their second or third year of participation.

Attrition for the CCU Analyses

Overall, the attrition rates for the CCU analyses were higher than the attrition rates for the OAKS analyses. The aim was to administer the CCU Assessments to all students in the 12 CCU schools. An examination of the data revealed that the vast majority of the missing scores resulted from entire classrooms of students or grade levels in individual schools not participating in the testing. Anecdotal reports indicated that certain teachers did not want to participate in the testing and, as a result, did not administer the CCUs to their students.

For the grade 3 analyses, all 12 CCU schools participated in the testing in years 1 through 3. Additionally, all 12 CCU schools participated in the testing for grades 4 and 5 in year 3. For these samples, the cluster-level attrition rates for the treatment and control groups were zero percent. However, in year 1, only four treatment and five control schools participated in the testing for grades 4 and 5. As a result, the attrition rates were 33 percent and 17 percent for the treatment and

[5] The students who received less than 40 percent of their instruction in year 1 of the study were excluded from the attrition calculations.

control groups, respectively. Furthermore, in year 2, one treatment school did not participate in the testing for grades 4 and 5, which produced an attrition rate of 17 percent for the treatment schools.

We calculated the student-level attrition rates for the CCU analyses in two ways and the results are presented in Exhibits A-3 to A-5 in the Appendix. We first calculated the attrition rates the same way we calculated them for the OAKS confirmatory analyses. These calculations are labeled "WWC Attrition Calculations" in the exhibits and are based on the students enrolled in the CCU schools prior to the randomization in 2010-11 only. Second, we calculated the attrition rates based on the students enrolled in the CCU schools in the fall of the year the CCUs were administered. For example, the students in the grade 3 CCU analyses for year 3 only needed to be enrolled in the CCU schools in the fall of year 3 and did not have to be enrolled in kindergarten in the CCU schools prior to the randomization. The second set of attrition analyses, which are labeled "Non-WWC Attrition Calculations" in the exhibits, show the attrition rates for the samples we used in the CCU analyses and are lower than the WWC attrition calculations.

Using the attrition calculations for the confirmatory analyses (i.e., the WWC Attrition Calculations), the overall and differential (i.e., the difference between the rates for the treatment and control groups) attrition rates at the school level and/or student level are expected to produce an unacceptable level of bias for the analyses for all samples with the exception of grade 3 in year 1 and grade 4 in year 3. When the attrition rates are thought to produce an unacceptable level of bias, the students need to be equivalent at baseline in order for the study to receive a rating of *Meets Group Design Standards with Reservations* (U.S. Department of Education, 2014).

Baseline Comparisons for the OAKS Analyses between the Treatment and Control Groups

The treatment and control students included in the confirmatory analyses were well matched on their prior achievement on the DRA2 and OAKS (see Exhibit 5). For the one-, two-, and three-year analyses, the effect sizes indexing the differences between the groups ranged from -0.10 to -0.03, indicating equivalence at baseline. Additionally, the differences between the treatment and control students were not statistically significant.

The demographic characteristics of the treatment and control students who were part of the samples for the confirmatory analyses, including free/reduced-price lunch status, ELL status, below grade-level reading status, and race/ethnicity were also well matched at baseline. The demographic characteristics of the treatment and control students are shown in Exhibits A-6, A-7, and A-8 in the Appendix for the students included in the one-, two-, and three-year analyses, respectively. None of the baseline differences between students in the treatment and control schools were statistically significant and none of the differences exceeded four percentage points, indicating equivalence.

Exhibit 5. Baseline Comparisons for Treatment and Control Students Included in the OAKS Analyses

Years of Program Participation	Treatment Students			Control Students					Effect Size
	Mean	SD	n	Mean	SD	n	Difference	*p* value	
One year	-0.03	1.00	5,719	0.00	1.00	6,110	-0.03	.82	-0.03
Two years	-0.03	1.00	3,325	0.02	1.00	3,590	-0.05	.71	-0.05
Three years	27.52	7.79	1,077	28.34	8.60	1,208	-0.82	.50	-0.10

Note. The means for the treatment group were calculated by adding the means for the control group (i.e., the unadjusted means) and the differences (i.e., the treatment-control contrasts from the HLM models). The baseline comparisons for the one-year and two-year samples used the DRA2 and OAKS and the scores were standardized within cohort and grade. The baseline comparison for the three-year sample used the DRA2 only. The effect sizes were calculated by dividing the differences by the pooled standard deviations.

Baseline Comparisons for the CCU Analyses between the Treatment and Control Groups

As shown in Exhibit 6, the students in the treatment and control schools in six of the nine CCU analyses were well matched on the baseline CCU measures. The samples that were equivalent were in grade 3 in years 1 and 3, grade 4 in years 1 through 3, and grade 5 in year 2. For these samples, the effect sizes indexing the differences between the groups ranged from -0.12 to 0.14 and none of the differences were statistically significant. According to the WWC (2014) guidelines, effect sizes greater than 0.25 standard deviations indicate that the groups are not equivalent. The effect sizes for the grade 3 sample in year 2 and the grade 5 samples in years 1 and 3 were greater than 0.25 standard deviations. When attrition is high and samples are not equivalent at baseline, which was the case for the grade 3 sample in year 2 and the grade 5 samples in years 1 and 3, the findings cannot meet WWC group design standards with or without reservations.

As shown in Exhibits A-9 to A-17, the demographic characteristics (i.e., free/reduced-price lunch status, ELL status, and race/ethnicity) of the treatment and control students who were part of the samples for the CCU analyses were fairly well matched at baseline. None of the baseline differences between students in the treatment and control schools shown in the exhibits were statistically significant, but some of the differences were nearly 20 percentage points. For example, 30.1 percent of the treatment group for the grade 3 year 1 sample was comprised of free/reduced-price lunch students, while 48.8 percent of the control group for the same sample was comprised of free/reduced-price lunch students. Finally, WestEd compared the baseline scores of the ELL and non-ELL students in the grade 4 samples from years 1 through 3 because the subgroup analyses were statistically significant for this subgroup. The treatment and control ELL and non-ELL students in these samples were well matched on their prior achievement on the CCU Assessments (see Exhibit A-18 in the Appendix). The effect sizes indexing the differences between the treatment and control students in the subgroups ranged from -0.17 to 0.09, indicating equivalence at baseline.

Exhibit 6. Baseline Comparisons for the Treatment and Control Students Included in the CCU Analyses, by Grade and Study Year

	Treatment Students			Control Students					
	Mean	SD	n	Mean	SD	n	Difference	*p* value	Effect Size
Grade 3									
Year 1	15.51	7.45	408	14.49	7.35	475	1.02	.59	0.14
Year 2	15.14	7.44	319	18.41	6.86	344	-3.27	.14	-0.46
Year 3	15.08	7.16	418	15.95	7.21	409	-0.87	.61	-0.12
Grade 4									
Year 1	20.72	8.59	265	20.71	10.23	252	0.01	.99	0.00
Year 2	20.35	10.17	308	20.74	9.21	481	-0.39	.89	-0.04
Year 3	20.10	8.77	413	18.98	9.61	541	1.12	.66	0.12
Grade 5									
Year 1	24.33	7.83	239	22.03	9.38	262	2.29	.39	0.26
Year 2	24.21	9.06	301	23.71	9.35	394	0.50	.84	0.05
Year 3	23.48	8.57	408	19.64	8.45	411	3.84	.10	0.45

Note. The means for the treatment group were calculated by adding the means for the control group (i.e., the unadjusted means) and the differences (i.e., the treatment-control contrasts from the HLM models). The effect sizes were calculated by dividing the differences by the pooled standard deviations. In each year, the students in grade 3 completed the Joy Test and the students in grade 4 completed the Ruth Test. In years 1 and 2, the students in grade 5 completed the Ruth Test and in year 3 the students in grade 5 completed the Jackie Test.

Findings

Oregon Assessment of Knowledge and Skills (OAKS) Reading/ Literature Test

The means on the OAKS Reading/Literature test for treatment and control students from the confirmatory analyses are presented in Exhibit 7. The means for treatment and control students ranged from 0.25 to 0.34 and indicated that both groups of students scored slightly above the state average. The results revealed no impact of the A4L Lessons Project on students' achievement on the OAKS. The differences between the treatment and control students on the OAKS after one, two, or three years of program participation were not statistically significant and the magnitude of the differences between treatment and control students were very small (i.e., less than one-tenth of a standard deviation). Because the A4L Lessons Project did not have a significant impact on the full sample of students, we did not conduct subgroup analyses to determine whether the impact of the program varied by the students' ELL status or eligibility for free/reduced-price lunch.

Exhibit 7. Means and Standard Deviations on the Post-Test OAKS Reading/Literature Test for Treatment and Control Students

Years of Program Participation	Treatment Students			Control Students					Effect Size
	Mean	SD	n	Mean	SD	n	Difference	*p* value	
One year	0.25	1.02	5,719	0.28	1.02	6,110	-0.03	.44	-0.03
Two years	0.29	1.07	3,325	0.30	1.06	3,590	-0.01	.75	-0.01
Three years	0.34	1.05	1,077	0.29	1.08	1,208	0.05	.47	0.05

Note. The means for the treatment group were calculated by adding the means for the control group (i.e., the unadjusted means) and the differences (i.e., the treatment-control contrasts from the HLM models). The OAKS scores were standardized by subtracting each score from the state average and dividing by the state standard deviation. The effect sizes were calculated by dividing the differences by the pooled standard deviations.

Histograms for the treatment and control groups included in the confirmatory analyses, showing the distributions of the OAKS scores for students after one, two, or three years of program participation, are presented in Exhibits 8-10 to graphically display the non-significant findings. The distributions are nearly identical, which is consistent with the non-significant mean differences between the treatment and control groups.

Exhibit 8. Histograms Depicting the Distribution of the Post-Test OAKS Reading/Literature Test Scores for Treatment and Control Students After One Year of Program Participation

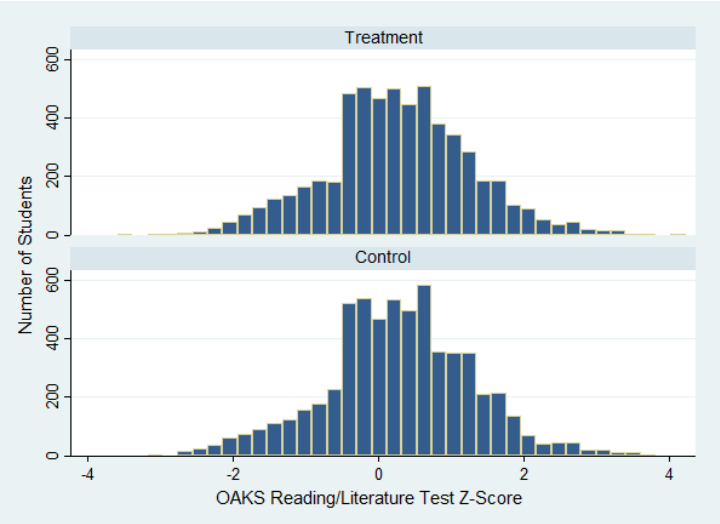


Exhibit 9. Histograms Depicting the Distribution of the Post-Test OAKS Reading/Literature Test Scores for Treatment and Control Students After Two Years of Program Participation

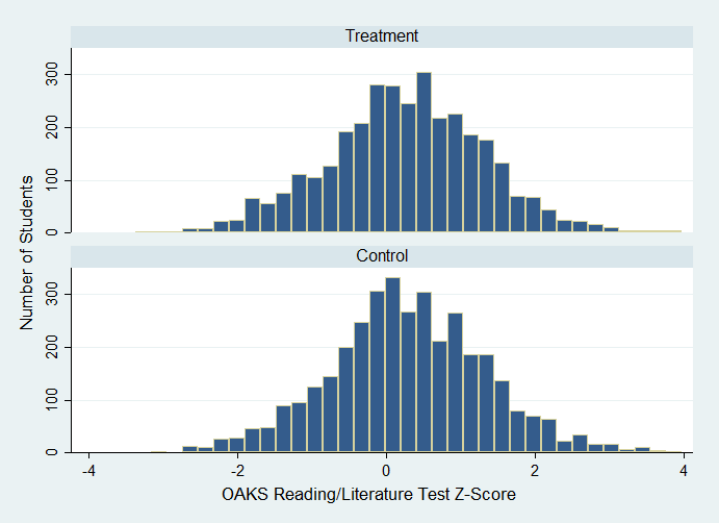
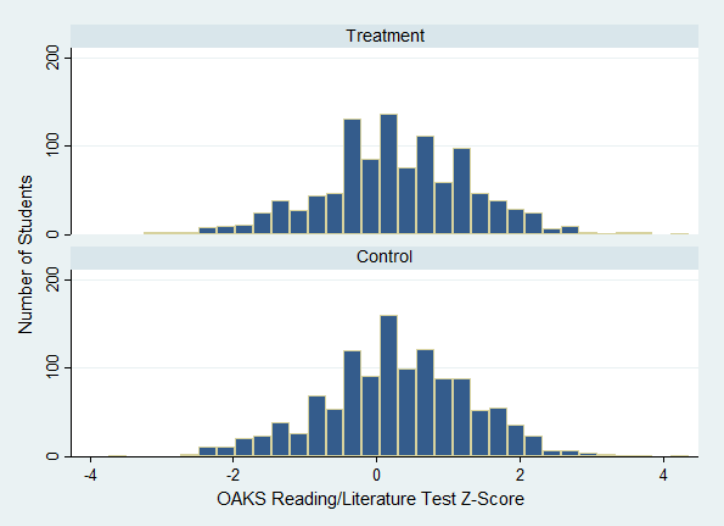


Exhibit 10. Histograms Depicting the Distribution of the OAKS Reading/Literature Test Scores for Treatment and Control Students After Three Years of Program Participation



Comprehensive Cross-Unit (CCU) Assessments

The means for treatment and control students on the post-test CCU Assessments are presented in Exhibit 11 separately by grade and study year. There was a consistent pattern of findings with the students in grade 4. In each study year, the students in grade 4 had significantly higher scores than the control students, indicating a positive impact of the A4L Lessons Project on students' writing achievement. The effect sizes indexing the differences ranged from 0.30 to 0.36, which are considered small positive program effects (Cohen, 1988).

The post-test means for treatment students were higher than the post-test means for control students in grade 3 in each study year. The pattern of findings with the students in grade 3 suggests that the A4L Lessons Project could be having a positive impact on the literacy achievement of students in grade 3. However, these differences were not statistically significant and the effect sizes indexing the differences ranged from 0.14 to 0.24. If these differences held with a larger sample of students, the differences would be reliable. In addition, the differences between the treatment and control students in grade 5 were not statistically significant. The means favored the control students in years 1 and 2 and the treatment students in year 3, indicating there was not a consistent pattern of findings suggesting a positive or negative impact of the A4L Lessons Project in grade 5.

Exhibit 11. Means and Standard Deviations on the Post-Test CCU Assessments for Treatment and Control Students, by Grade and Study Year

	Treatment Students			Control Students					Effect
	Mean	SD	n	Mean	SD	n	Difference	*p* value	Size
Grade 3									
Year 1	20.30	7.25	408	18.70	7.02	475	1.60	.06	0.22
Year 2	21.85	7.13	319	20.15	7.18	344	1.70	.14	0.24
Year 3	19.74	6.89	418	18.71	7.42	409	1.03	.28	0.14
Grade 4									
Year 1	26.08	7.31	265	23.53	9.64	252	2.55	.007	0.30
Year 2	25.88	8.78	308	22.66	9.22	481	3.22	<.001	0.36
Year 3	24.90	7.54	413	22.20	9.50	541	2.71	<.001	0.31
Grade 5									
Year 1	25.45	7.05	239	25.69	8.47	262	-0.24	.80	-0.03
Year 2	25.23	8.67	301	25.72	8.28	394	-0.49	.63	-0.06
Year 3	22.78	8.34	408	22.07	8.00	411	0.71	.14	0.09

Note. The means for the treatment group were calculated by adding the means for the control group (i.e., the unadjusted means) and the differences (i.e., the treatment-control contrasts from the HLM models). The effect sizes were calculated by dividing the differences by the pooled standard deviations. In each year, the students in grade 3 completed the Joy Test and the students in grade 4 completed the Ruth Test. In years 1 and 2, the students in grade 5 completed the Ruth Test, and in year 3 the students in grade 5 completed the Jackie Test.

We conducted the ELL and free/reduced-price lunch subgroup analyses only with the students in grade 4 because significant impacts were evident for the full sample across years 1 through 3 of the study. The means on the post-test CCUs for treatment and control students in grade 4 are presented in Exhibit 12 separately for ELL and non-ELL students. The results showed that the impact of the program differed significantly across the subgroups of interest in years 1 and 2, but not in year 3. Across the first two years of the study, the impact of the A4L Lessons Project was substantially higher for ELL students, suggesting that the program had a greater impact on this subgroup of students. The effect sizes for ELL students were 0.87 and 0.69 compared to 0.23 and 0.31 for the non-ELL students.

The effect sizes for the ELL students are considered large program impacts by educational research standards and should be viewed extremely cautiously. The findings are based on a very small number of ELLs and the 95 percent confidence interval around these effect sizes is very large. In other words, if the study were conducted a second time, the effect sizes for ELLs could vary substantially from 0.87 and 0.69 by chance alone. In addition, the subgroup analyses showed that the estimated impact of the program on ELL and non-ELL students was nearly identical in year 3. The fact that the ELL effect did not replicate across years weakens our confidence in the findings from years 1 and 2.

The free/reduced-price lunch subgroup analyses with the students in grade 4 did not reveal any statistically significant differences between the free/reduced-price lunch students and the non-free/reduced-price lunch students. In other words, the results indicated that the A4L Lessons

Project had an equally positive impact on free/reduced-price lunch and non-free/reduced-price lunch students.

Exhibit 12. Means and Standard Deviations on the Post-Test CCU Assessments for Treatment and Control Students in Grade 4, by English Language Learner Status

	Treatment			Control					Effect Size
	Mean	SD	n	Mean	SD	n	Difference	*p* value	
Year 1									
English language learner	25.73	6.03	37	19.65	7.63	47	4.34	.003	0.87
Non-English language learner	27.88	6.97	228	26.13	8.05	205			0.23
Year 2									
English language learner	28.30	8.31	71	22.79	7.76	131	3.05	.004	0.69
Non-English language learner	27.90	7.78	237	25.44	8.12	350			0.31
Year 3									
English language learner	25.31	8.47	70	22.34	8.11	172	0.33	.75	0.36
Non-English language learner	27.89	6.82	343	25.26	8.50	369			0.34

Note. The means for the treatment subgroups were calculated by adding the means for the control subgroups and the estimates from the HLM models that contrasted the treatment and control groups and the differential effect of the treatment for the subgroups (i.e., the "Difference" column). The effect sizes were calculated by dividing the differences between the means for the treatment and control groups by the pooled standard deviations.

Conclusions

The confirmatory analyses revealed that the A4L Lessons Project had no impact on students' performance on the OAKS Reading/Literature test. After one, two, and three years of participation in the program, the students who received the intervention scored no higher on the state test than the control students who received the district's typical language arts curriculum. It is likely that the results from the confirmatory analyses will receive a rating of *Meets WWC Group Design Standards without Reservations* from the WWC because the randomization produced equivalent groups at baseline, the study had low attrition, and the outcome measure was a standardized state test with acceptable reliability and validity. The study design used for the confirmatory analyses produced strong causal evidence that the A4L Lessons Project did not produce improvements in students' literacy achievement as assessed by performance on the OAKS Reading/Literature test.

The generalizability of the findings from the confirmatory analyses relying on the OAKS Reading/Literature test is unknown. The study was conducted with 32 elementary schools in one school district in Oregon and it is unclear whether the findings would replicate in other contexts (e.g., districts with different student populations). Additionally, the teachers' role in implementing the A4L Lessons Project is critical and the impacts could change with a different group of teachers participating in the program. Furthermore, the program was not implemented with complete fidelity. For example, not all of the classrooms in the treatment schools implemented all of the Lessons comprising an A4L Lessons Unit and not all of the teachers in the treatment schools attended at least four professional learning community (PLC) sessions each school year. It is conceivable that the student impacts reported may have been different were the program implemented with higher fidelity to the program design. Finally, the results based on the CCU Assessments indicate that findings from the confirmatory analyses may not apply to all achievement domains and the program may have impacts on outcomes other than state standardized tests.

The results of the exploratory analyses relying on the CCU Assessments revealed that the A4L Lessons Project had a positive impact on the literacy and life skills of students in grade 4 across all three years of the project. The research design for the exploratory analyses was much less rigorous than the design for the confirmatory analyses because only 12 schools (six treatment and six control schools) participated in the CCU portion of the study and the attrition rates differed substantially across the treatment and control groups at the student and school levels. As a result, the findings from the CCU analyses would likely not receive a rating of *Meets WWC Group Design Standards with or without Reservations*, thereby disallowing causal inferences to be drawn regarding program impacts. Additionally, the CCU findings suggest that the program may be particularly effective at improving the literacy achievement of ELL students in grade 4. However, these findings are based on a small number of ELL students unevenly distributed across schools and should be viewed extremely cautiously.

Although there was a trend for the treatment students in grade 3 to have higher CCU scores than the control students after participating in the program, the treatment-control group differences were not statistically significant. Similarly, the analyses showed that the program did not have a statistically significant impact on the CCU scores of students in grade 5.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57,* 289–300.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

John, O. P., & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339–369). New York, NY: Cambridge University Press.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms.* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92.

Marsh, L. C., & Cormier, D. R. (2002). Spline regression models. *Quantitative Applications in the Social Sciences, 137*. Thousand Oaks, CA: Sage Publications.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Oregon Department of Education. (2011). *2011-2012 technical report: Oregon's Statewide Assessment System—test development volume 2*. Retrieved from http://www.ode.state.or.us/search/page/?=1305

Oregon Department of Education. (2012). *A best practices guide for districts: Regarding when to administer the Oregon Assessment of Knowledge and Skills (OAKS)*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/admin/best_practices_guide.pdf

Oregon Department of Education. (n.d.). *Reading/literature test specifications and blueprints, 2012–2014 grade 5*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/dev/testspecs/asmtrdtestspecsg5_2012-2014.pdf

Pearson Education Inc. (2009). *K–8 technical manual: Developmental Reading Assessment*® (2nd ed.). Upper Saddle River, NJ: Pearson.

Price, C. (2014). *Causal inference at student and cluster levels.* Presentation at the i3 Project Directors Meeting, Washington, DC.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014-4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Seidel, S., Tishman, S., Winner, E., Hetland, L., & Palmer, P. (2009). *The qualities of quality: Understanding excellence in arts education*. Cambridge, MA: Project Zero Publications.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Silva, E. (2009). *Measuring skills for the 21st century*. Washington, DC: Education Sector.

U.S. Department of Education. (2014). *What Works Clearinghouse: Procedures and standards handbook* (Version 3.0). Washington, DC: Institute of Education Sciences.

Appendix

Exhibit A-1. Baseline Comparisons on the School-Level Measures for Treatment and Control Schools

	Treatment Schools		Control Schools				
	Mean	SD	Mean	SD	Difference	*p* value	Effect Size
School enrollment	559.38	138.41	589.88	164.50	-30.50	.57	-0.20
Percentage of English language learners	19.04	14.58	20.35	18.37	-1.31	.82	-0.08
Percentage free/reduced-price lunch	41.77	23.46	40.63	26.42	1.13	.90	0.05
Percentage minority	48.30	12.84	47.84	18.54	0.47	.93	0.03
OAKS Reading/Literature test percentage meets and above 2011	87.73	6.80	87.16	8.69	0.56	.84	0.07
OAKS Writing test percentage meets and above 2011	58.32	13.72	53.46	14.17	4.86	.33	0.35

Note. Treatment $n = 16$; Control $n = 16$. The p values were calculated using t tests. The effect sizes were calculated by dividing the differences by the pooled standard deviations.

Exhibit A-2. Descriptions of the Eight Score Reporting Categories (SRCs) Comprising the OAKS Reading/Literature Test

SRC	Description
SRC 1 – Vocabulary	In this skill area, students use appropriate strategies to determine the meaning of unknown words. For the items on the state assessment, students are asked to focus primarily on context clues. Passages providing context clues include well-known, high frequency words that explain the meaning of the target word. The clues may be stated directly in a phrase or in sentences before or after use of the target word, or they may be found through careful reading of the entire text. At some grade levels, students may also be asked to use context clues to determine the meanings of words with multiple meanings or of phrases, such as idioms and figurative expressions.
SRC 2 – Read to perform a task	When reading to perform a task, students use skimming and scanning techniques to search for information in what is termed "practical" text. Depending on the grade level, practical text may include charts, schedules, directions, recipes, forms, maps, graphs, or job and consumer-related materials. The reader's purpose is to look for information in order to do something. At grade 8 and at the high school level, questions ask students to synthesize information and reach logical conclusions, not simply to understand the selection's content.
SRC 3 – Demonstrate general understanding	Students show a general understanding by accurately responding to questions about material that is explicitly stated in the text. After reading informational text, students might be asked to identify an article's topic statement, recall the correct sequence of events, or identify important details that were stated in the reading passage. Similarly, after reading literary text, students might be asked questions about the sequence of events in the plot or asked to identify details or events that were critical to the development of the plot.
SRC 4 – Develop an interpretation	To develop an interpretation, students must look beyond what is explicitly stated in a selection and show a more complete understanding of what was read. For informational text, questions include drawing inferences about the author's meaning, making predictions about forthcoming information in the text or events that are likely to occur in the future, and drawing conclusions about reasons for actions when those reasons are not explicitly stated. For literary text, students make predictions about events likely to happen later in the story, interpret the story to uncover its themes, and draw conclusions about traits present in the character and motivations for his or her actions.
SRC 5 – Examine content and structure: Informational text	Examining content and structure requires students to critically analyze and evaluate text. Students stand apart from the text, consider it objectively, and evaluate its quality and effectiveness. For informational text, questions ask students to consider the author's purpose and style. Depending on the grade level, students may be asked about instances in which the author has relied on facts or opinion; which arguments or statements have support; whether the passage has evidence of bias; and what structural elements are present in the work. At the upper grades, students may be asked to compare information and make connections across parts of a text or between texts. This reporting category is not assessed at grade 3.
SRC 6 – Examine content and structure: Literary text	Examining content and structure requires students to critically analyze and evaluate text. Students stand apart from the text, consider it objectively, and evaluate its quality and effectiveness. For literary text, students evaluate the use of literary elements and devices and the impact and purpose of their use within a selection. Questions may ask students to examine selections to determine their mood or tone and to determine how authors achieved that mood or tone. Students may be asked literary genre questions at specific grades (poetry at grade 6 and drama at the high school level, for example). At the upper grades, students may be asked to compare the treatment of themes and make connections between two literary selections. This reporting category is not assessed at grades 3 and 4.
SRC 8 – Reading informational text SRC 9 – Reading literary text	In addition to the overall reading score and subscores in the score reporting categories, beginning in 2012-2013, students will receive subscores in Reading Informational Text and Reading Literary Text. This breakdown can be used to compare student performance on the items associated with literary selections to performance on items relating to informational texts. This type of analysis at the individual or group level can be used to help inform instruction.

Note. The text on SRCs 1 through 6 has been reproduced from Oregon Department of Education's (2011) Technical Report on Oregon's Statewide Assessment System. The text on SRCs 8 and 9 has been reproduced from Oregon Department of Education's (n.d.) Reading/Literature Test Specifications and Blueprints 2012-2014. There is little publically available material on SRCs 8 and 9 because they were new in 2012-13.

Exhibit A-3. Number of Students Included in and Excluded from the Grade 3 CCU Analyses and the Attrition Rates

	Study Year					
	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
WWC Attrition Calculations						
Students randomized	592	646	536	629	514	626
Students randomized with complete data	362	409	262	250	293	238
Attrition rate	38.9%	36.7%	51.1%	60.3%	43.0%	62.0%
Non-WWC Attrition Calculations						
All students in grade 3	594	660	505	616	539	631
Students with complete data in grade 3	408	475	319	344	418	409
Attrition rate	31.3%	28.0%	36.8%	44.2%	22.4%	35.2%

Exhibit A-4. Number of Students Included in and Excluded from the Grade 4 CCU Analyses and the Attrition Rates

	Study Year					
	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
WWC Attrition Calculations						
Students randomized	523	619	592	646	536	629
Students randomized with complete data	244	216	258	364	313	367
Attrition rate	53.3%	65.1%	56.4%	43.7%	41.6%	41.7%
Non-WWC Attrition Calculations						
All students in grade 3	520	625	591	626	540	661
Students with complete data in grade 3	265	252	308	481	413	541
Attrition rate	49.0%	59.7%	47.9%	23.2%	23.5%	18.2%

Exhibit A-5. Number of Students Included in and Excluded from the Grade 5 CCU Analyses and the Attrition Rates

	Study Year					
	Year 1		Year 2		Year 3	
	Treatment	Control	Treatment	Control	Treatment	Control
WWC Attrition Calculations						
Students randomized	541	622	523	619	592	646
Students randomized with complete data	222	218	251	303	329	283
Attrition rate	59.0%	65.0%	52.0%	51.1%	44.4%	56.2%
Non-WWC Attrition Calculations						
All students in grade 3	538	638	503	599	609	642
Students with complete data in grade 3	239	262	301	394	408	411
Attrition rate	55.6%	58.9%	40.2%	34.2%	33.0%	36.0%

Exhibit A-6. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Students Included in the One-Year OAKS Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	39.2%	2,239	40.2%	2,456	-1.0%	.99
English language learner	15.2%	868	18.4%	1,122	-3.2%	.89
Below grade-level reader	20.4%	1,169	20.7%	1,265	-0.3%	.72
African American/Black	2.9%	164	2.3%	138	0.6%	.15
Asian/Pacific Islander	15.9%	910	13.2%	809	2.7%	.54
Hispanic/Latino	23.3%	1,334	27.1%	1,654	-3.7%	.91
White	51.3%	2,934	49.8%	3,042	1.5%	.85
Other	6.6%	377	7.6%	467	-1.1%	.13

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-7. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Students Included in the Two-Year OAKS Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	38.8%	1,291	39.4%	1,413	-0.5%	.99
English language learner	14.3%	476	18.1%	649	-3.8%	.96
Below grade-level reader	22.7%	756	22.4%	803	0.4%	.70
African American/Black	2.9%	98	1.8%	66	1.1%	.01
Asian/Pacific Islander	16.3%	541	13.0%	468	3.2%	.50
Hispanic/Latino	23.1%	769	26.9%	964	-3.7%	.99
White	51.0%	1,696	50.6%	1,816	0.4%	.89
Other	6.6%	221	7.7%	276	-1.0%	.18

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-8. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Students Included in the Three-Year OAKS Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	36.5%	393	37.4%	452	-0.9%	.96
English language learner	12.6%	136	14.5%	175	-1.9%	.58
Below grade-level reader	30.7%	331	27.2%	328	3.6%	.38
African American/Black	2.9%	31	1.9%	23	1.0%	.14
Asian/Pacific Islander	15.6%	168	12.8%	155	2.8%	.79
Hispanic/Latino	23.6%	254	26.7%	323	-3.2%	.79
White	51.1%	550	50.5%	610	0.6%	.75
Other	6.9%	74	8.0%	97	-1.2%	.49

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-9. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Third Graders Included in the Year 1 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	30.1%	123	48.8%	232	-18.7%	.44
English language learner	17.9%	73	30.5%	145	-12.6%	.41
African American/Black	1.2%	5	3.2%	15	-1.9%	.17
Asian	22.1%	90	17.9%	85	4.2%	.79
Hispanic/Latino	20.8%	85	34.7%	165	-13.9%	.44
White	52.5%	214	37.7%	179	14.8%	.07
Other	3.4%	14	6.5%	31	-3.1%	.08

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-10. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Third Graders Included in the Year 2 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	33.9%	108	37.8%	130	-3.9%	.66
English language learner	20.7%	66	23.0%	79	-2.3%	.39
African American/Black	0.9%	3	2.6%	9	-1.7%	.36
Asian	26.6%	85	26.2%	90	0.5%	.86
Hispanic/Latino	22.3%	71	27.6%	95	-5.4%	.54
White	41.4%	132	36.3%	125	5.0%	.28
Other	8.8%	28	7.3%	25	1.5%	.47

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-11. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Third Graders Included in the Year 3 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	33.7%	141	45.7%	187	-12.0%	.33
English language learner	19.9%	83	30.3%	124	-10.5%	.16
African American/Black	1.2%	5	1.5%	6	-0.3%	.84
Asian	23.4%	98	22.2%	91	1.2%	.58
Hispanic/Latino	20.8%	87	28.9%	118	-8.0%	.45
White	48.3%	202	38.9%	159	9.5%	.15
Other	6.2%	26	8.6%	35	-2.3%	.20

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-12. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Fourth Graders Included in the Year 1 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	31.3%	83	41.7%	105	-10.3%	.55
English language learner	14.0%	37	18.7%	47	-4.7%	.62
African American/Black	3.4%	9	2.4%	6	1.0%	.56
Asian	22.3%	59	17.9%	45	4.4%	.88
Hispanic/Latino	21.1%	56	29.4%	74	-8.2%	.71
White	46.8%	124	43.3%	109	3.5%	.57
Other	6.4%	17	7.1%	18	-0.7%	.80

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-13. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Fourth Graders Included in the Year 2 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	38.3%	118	44.3%	213	-6.0%	.57
English language learner	23.1%	71	27.2%	131	-4.2%	.49
African American/Black	1.0%	3	2.3%	11	-1.3%	.31
Asian	22.4%	69	21.8%	105	0.6%	.79
Hispanic/Latino	26.9%	83	32.2%	155	-5.3%	.52
White	46.1%	142	36.4%	175	9.7%	.14
Other	3.6%	11	7.3%	35	-3.7%	.11

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-14. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Fourth Graders Included in the Year 3 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	27.4%	113	46.0%	249	-18.7%	.40
English language learner	16.9%	70	31.8%	172	-14.8%	.21
African American/Black	1.0%	4	2.2%	12	-1.2%	.31
Asian	26.4%	109	20.3%	110	6.1%	.50
Hispanic/Latino	19.4%	80	33.8%	183	-14.5%	.69
White	44.6%	184	38.3%	207	6.3%	.38
Other	8.7%	36	5.4%	29	3.4%	.12

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-15. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Fifth Graders Included in the Year 1 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	31.4%	75	50.8%	133	-19.4%	.57
English language learner	14.6%	35	15.3%	40	-0.6%	.57
African American/Black	1.7%	4	3.4%	9	-1.8%	.30
Asian	14.2%	34	16.4%	43	-2.2%	.50
Hispanic/Latino	21.3%	51	31.3%	82	-10.0%	.63
White	56.5%	135	43.1%	113	13.4%	.14
Other	6.3%	15	5.7%	15	0.6%	.87

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-16. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Fifth Graders Included in the Year 2 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	35.2%	106	50.8%	200	-15.5%	.54
English language learner	18.3%	55	22.3%	88	-4.1%	.99
African American/Black	3.0%	9	2.8%	11	0.2%	.71
Asian	25.6%	77	17.5%	69	8.1%	.70
Hispanic/Latino	22.6%	68	37.3%	147	-14.7%	.49
White	46.5%	140	36.8%	145	9.7%	.38
Other	2.3%	7	5.6%	22	-3.3%	.05

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-17. Baseline Comparisons on the Demographic Characteristics for Treatment and Control Fifth Graders Included in the Year 3 CCU Analyses

	Treatment Students		Control Students			
	%	n	%	n	Difference	*p* value
Free/reduced-price lunch	26.0%	106	44.8%	184	-18.8%	.40
English language learner	11.3%	46	20.9%	86	-9.7%	.33
African American/Black	2.2%	9	2.7%	11	-0.5%	.72
Asian	23.5%	96	20.2%	83	3.3%	.72
Hispanic/Latino	17.9%	73	33.3%	137	-15.4%	.37
White	51.7%	211	36.7%	151	15.0%	.10
Other	4.7%	19	7.1%	29	-2.4%	.15

Note. Other includes American Indian/Alaskan Native and Multi-Racial. The *p* values were calculated using multi-level logistic regression models that accounted for the hierarchical structure of the data.

Exhibit A-18. Means and Standard Deviations on the Pre-Test CCU Assessments for Treatment and Control Students in Grade 4, by English Language Learner Status

	Treatment			Control					Effect
	Mean	SD	n	Mean	SD	n	Difference	*p* value	Size
Year 1									
English language learner	14.66	5.56	37	15.03	6.11	47	0.23	.91	-0.06
Non-English language learner	22.60	8.18	228	23.20	9.32	205			-0.07
Year 2									
English language learner	14.45	8.14	71	15.77	7.66	131	-0.37	.82	-0.17
Non-English language learner	22.34	9.22	237	23.29	8.41	350			-0.11
Year 3									
English language learner	16.01	7.85	70	15.36	7.12	172	0.55	.71	0.09
Non-English language learner	22.20	8.50	343	22.10	9.01	369			0.01

Note. The means for the treatment subgroups were calculated by adding the means for the control subgroups and the estimates from the HLM models that contrasted the treatment and control groups and the differential effect of the treatment for the subgroups (i.e., the "Difference" column). The effect sizes were calculated by dividing the differences between the means for the treatment and control groups by the pooled standard deviations.

730 Harrison Street
San Francisco, California 94107-1242