David P. Weikart
Center for Youth
Program Quality

Quality-Outcomes Study for Seattle Public Schools Summer Programs, 2016 Program Cycle

Submitted to the Raikes Foundation on January 11, 2018

Charles Smith, PhD
Leanne Roy
Stephen C. Peck, PhD
Colin Macleod
Katharine Helegda
John Hughes

301 W. Michigan Ave. | Suite 200 | Ypsilanti, MI 48197 | 734.961.6900 | www.cypq.org
The David P. Weikart Center for Youth Program Quality is a division of the Forum for Youth Investment

The David P. Weikart Center for Youth Program Quality empowers education and human service leaders to adapt, implement and bring to scale best-in-class, research-validated quality improvement systems to advance child and youth development. Afterschool and other out-of-school time systems throughout the United States rely on the Weikart Center's intervention, performance metrics and aligned professional development to drive their continuous improvement efforts. These include an evidence-based intervention model (Youth Program Quality Intervention, or YPQI) and core set of instructional quality metrics (Youth Program Quality Assessment, or Youth PQA). www.cypq.org

The Weikart Center is a division of the Forum for Youth Investment.

Table of Contents

Summary

This quality-outcomes study was designed to both (a) describe performance in Seattle Public Schools (SPS) summer learning programs in ways that are useful to staff and (b) provide evaluative evidence (i.e., validity) for an instructional model that includes challenging academic content and responsive instructional practices.

Results from this study were mainly positive yet partially ambiguous. Summer program offerings were well-attended and characterized by high-quality instructional practices, with a majority of students increasing their literacy and math skills during the program. Findings about the association between exposure to more responsiveness instruction (e.g., quality) and academic skill change were mixed. Results include:

Positive academic skill change was found in the raw data, including for academically at-risk students. Positive change on the academic performance measures used during the summer program was found for 73% of students, and positive change on the academic achievement tests was found for 74% of students from the 2015 to 2016 school year. Standardized effect sizes for the full sample ranged from medium to large (i.e., d_z = .56 - .95) across the seven academic skill measures.

Attendance was regular, and instructional responsiveness was consistently high. Summer program attendance for 21 or more days (out of a total possible 27 days) was observed for 77% of students. Analysis of instructional responsiveness using the Summer Learning PQA revealed three profiles of instructional responsiveness at the point of service: high, medium, and low quality. However, compared to other urban samples, the "low" SPS profile is not very low.

Students in SPS summer programs had similar rates of skill change across profiles of instructional responsiveness in the most rigorous models for 3rd and 4th grade students (N = 535); that is, there was insufficient evidence in support of the hypothesized pattern of differential skill change across profiles of instructional quality. However, these results should be interpreted with caution due to the absence of a true low-quality instructional practices subgroup in the sample. Less statistically rigorous but more theoretically well-specified models for the entire K-4 sample (N = 1060) revealed a positive association between instructional quality and academic skill change, despite the lack of a true low-quality subgroup.

Analyses of academically at-risk students revealed similarly mixed results. In the more statistically rigorous models with grades 3-4, students who entered SPS summer programs below proficient on academic achievement tests for the prior school year (2015-16) showed similar rates of academic skill change across profiles of instruction. In the theoretically well-specified models, academically at-risk students showed greater changes in academic skills in summer programs with higher-quality instructional practices.

Introduction

Since 2013, a collaborative of funders, public and private summer learning service providers, and several technical services organizations have partnered to improve the quality and effectiveness of summer learning services in cities such as Denver, CO; Grand Rapids, MI; Oakland, CA; Seattle, WA; and St. Paul, MN. The method of improving quality and outcomes was a continuous improvement intervention for summer service providers' organizations called the *Summer Learning Program Quality Intervention* (SLPQI). In the SLPQI, summer learning organizations follow a cycle of planning, assessment, and improvement, building the quality of summer instruction toward validated standards and benchmarks over successive cycles.[1] The SLPQI was designed to generate cumulative at-scale improvement in the quality of instructional practices and student academic outcomes in school district and community-based summer learning programs.

Over five summer cycles, an iterative sequence of *design and development studies* were conducted to evaluate (a) SLPQI implementation fidelity and feasibility; (b) adaptations to the design, training, and technical assistance supporting implementation; and (c) the validity of the standard for high-quality instructional practices used in the SLPQI (Smith et al., 2017; Smith et al., 2015; http://cypq.org/SummerLearningPQI). The standard for high-quality instructional practice in the SLPQI is the *Summer Learning Youth Program Quality Assessment* (Summer Learning PQA). The SLPQI and Summer Learning PQA were designed to focus the science and practice of continuous improvement on qualities of the summer curriculum (i.e., responsive practices and challenging content) that build social, emotional, and academic skills.

As part of this broader SLPQI design and development work, during the summers of 2015 and 2016, the Weikart Center, Seattle Public Schools, Schools Out Washington, and the Raikes Foundation collaborated to conduct two studies focused on the following research questions:

- Are Seattle Public Schools (SPS) summer programs high quality and well attended?
- Does participation in a higher-quality summer program predict more academic skill gain during both the summer program and the subsequent school year?
- Do students who enter programs with lower academic skills gain more academic skills in higher-quality programs?

In addition to describing performance in the Seattle Public Schools summer programs, these studies address the validity of the Summer Learning PQA as a standard for high-quality instruction and

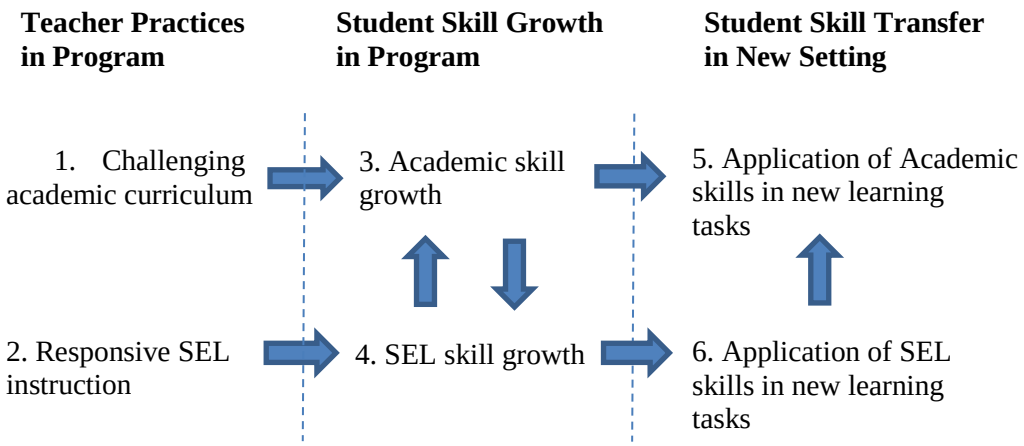
[1] The Summer Learning Program Quality Intervention (SLPQI) is a continuous improvement intervention for summer learning programs that includes four core components: (a) standards and measures for quality of management and instructional practices (i.e., the Summer Learning PQA), (b) training and technical assistance supports, (c) performance data products, and (d) a continuous improvement cycle that fits the prior three elements to local circumstances and resources.

the core performance metric in the SLPQI. Association between the quality of instruction and academic skill development is a critical aspect of validity for that standard and for both the SLPQI and the Summer Learning PQA. Results for the 2015 summer cohort are reported in Smith et al. (2015), and the current report presents results for the 2016 summer cohort.

Theory of Change

Within the 60 SPS summer programs studied in summer 2016, there were important similarities: First, the program-offering design included a curriculum with *challenging academic content* where expert staff led youth through intensive skill-building sequences over many hours of practice. Second, each curriculum emphasized *responsive instructional practices* that were designed to build social and emotional learning (SEL) skills; that is, to help youth be successful at their learning threshold. Sometimes learning a new skill can be frustrating, boring, or anxiety-provoking; however, in SPS summer programs, staff were trained to step in, provide reassurance, and model appropriate thinking and behavior when learning challenges occurred (i.e., to engage in co-regulation[2]). Third, each program offering was based on the theoretically-informed and evidence-based idea that the combination of challenging academic content and responsive instruction can help students grow skills simultaneously in both the targeted academic skill domains and the SEL skill domains that support academic skill learning (e.g., managing emotions, problem solving). Figure 1 illustrates this model of integrated skill learning: growing mastery in academic skills and growing mastery in SEL skills necessary to learn in any content area.

Figure 1. SPS Summer Program Theory of Change



[2] The term co-regulation refers to adult behavior designed to help children and youth successfully self-regulate; for example, to stay focused, keep moving, process emotion, and get the task at hand completed (Murray et al., 2015). Youth with atypical patterns of development due to exposure to trauma or chronic stress may need higher levels of co-regulation from adults. Co-regulation is what happens when staff uses responsive practices to keep the stress and strain of a challenging project curriculum in the optimal range.

This study includes measures for items 2, 3, and 5 of Figure 1. Measures of SEL skills during the program (e.g., emotion expression) or during the subsequent school year (e.g., suspensions) were not included. We assume that item 1, challenging academic content, was available to students and constant across all 60 SPS programs that implemented the same Math and Literacy content curricula.

Literacy activities for different grade levels were drawn from two online literacy curricula from Houghton Mifflin Harcourt: *iRead* and *System44*. Students in kindergarten through second grade used iRead44 and received direct instruction and practice opportunities aimed at improving their reading fluency, sight word accuracy, and comprehension. Students in grades three and four used System44 and received direct instruction and practice opportunities aimed at improving reading rate, expression, accuracy in oral fluency, phrasing, decoding skills, comprehension, and independent reading. During the five-week program, instruction was delivered in small group settings and through adaptive software tailored to student skill levels. Both literacy curricula models included manualized training and coaching for all staff, and fidelity checks were performed by site coordinators.

The math curriculum, *Summer Staircase*, was developed locally by SPS staff and a Seattle-based math education consultant, Math for Love (mathforlove.com), which develops play-based math curricula. The math curriculum, which was developed in 2013, was constructed for each grade band (i.e., K, 1-2, and 3-4) and aligned to the Common Core standards for each grade level. Teacher training was offered before Summer Staircase began, and ongoing support was provided throughout the program. This curriculum, which has been used since 2013, provided opportunities for students to develop their mathematical content knowledge and skills and their perseverance in mathematical practices. In order to create a positive experience for students, games and manipulatives (e.g., pig in pairs and pattern blocks) were a core feature of the curriculum. All in-program assessments were built into, and aligned with, the respective curriculum.

Method

Participants

Summer program staff consisted of 40 individual teachers grouped into 60 teacher teams across 19 schools where summer programs were located. Summer program offerings included 29 Kindegarten-2nd grade offerings and 31 3rd-4th grade offerings.

Students in the study included 39 (4%) students in kindergarten, 173 (15%) in 1st grade, 250 (23%) in 2nd grade, 334 (31%) in 3rd grade, 282 (26%) in 4th grade, and 15 (1%) in 5th grade.[3] The overall

[3] The program targeted K-4, however 15 5th graders entered the program due to an oversight in enrollment.

sample was 48% female, with 50% of participants identified as having limited English proficiency and 21% as having an individualized education plan. Participants were identified as Asian or Other Pacific Islander (23%), Black or African American (31%), Hispanic or Latino (25%), White (9%), Two or More Races (8%), American Indian (< 1%), or Native Hawaiian or Pacific Islander (< 1%).

Measures

The following three measure of instructional responsiveness, seven individual-level measures for academic skill (i.e., five in-program *academic performance* measures and two state *academic achievement* tests), and nine individual-level covariates (e.g., variables that may influence selection into summer learning programs or rates of academic learning) were used in the study. Mean, standard deviation, range, and sample size for all measures are listed in Appendix Table A-1.

Instructional Responsiveness. Instructional responsiveness was assessed using the Summer Learning PQA - Form A, an observation-based measure designed to assess the quality of instructional practices. Three domains from this measure were used in the study: Supportive Environment (e.g., a structured environment facilitated through guidance and encouragement; 23 items), Interaction (e.g., opportunities for leadership and collaboration; 10 items), and Engagement (e.g., opportunities for planning and reflection; 11 items). Trained raters produced complete ratings at two time points that were averaged together to create the three scores (ranging from 1 to 5) for each program.

Sight Words. Sight Words scores refer to the number of target words correctly read by students, using either the iRead or System44 curricula. Teachers conducted sight words assessment at both the beginning (Time 1) and end (Time 2) of the program, and a score was recorded for the number of correct words read from an established list.

Oral Fluency. Oral Fluency scores refer to the number of words per minute correctly identified from a previously read passage, and it was assessed for 3rd and 4th grade students only. Different passages were read for one minute, and the correct number of words per minute was recorded.

Math Scores. Math assessments (aka, *Math Assessment*) were constructed for three grade-level groupings (i.e., K only, grades 1-2, and grades 3-4) and consisted of ten items aligned to the Common Core standards for each grade level. The items were developed by the Summer Staircase math curriculum developer (http://mathforlove.com/) and aligned to the Summer Staircase curriculum. Assessments were collected during the first and last week of the summer session.

Math Content. Teachers completed an observational checklist for each of their students. The checklist is a set of grade-level objectives, accompanied by a rating scale (1 = not ready to begin learning this topic – needs more attention on prerequisites; 2 = ready to begin learning about topic; 3 = making basic progress; 4 = strong knowledge, with some gaps; 5 = student shows mastery or excellent progress). The ratings were submitted at the end of the first three weeks of the program and at the end of the sixth

week. An average score for Common Core Content (aka, *Math Content*) topics was calculated by averaging across relevant objectives for each student.

Math Practices. Teachers completed an observational checklist for each of their students. The checklist is a set of grade-level objectives, accompanied by a rating scale (1 = not ready to begin learning this topic – needs more attention on prerequisites; 2 = ready to begin learning about topic; 3 = making basic progress; 4 = strong knowledge, with some gaps; 5 = student shows mastery or excellent progress). The ratings were submitted at the end of the first three weeks of the program and at the end of the sixth week. An average score for Common Core Practices (aka, *Math Practice*) was calculated by averaging across relevant objectives for each student.

State Math Achievement Test. The Smarter Balanced Math Test was completed in March - June of 2017 by students in grades 3-4 (http://www.k12.wa.us/smarter/).

State Literacy Achievement Test. The Smarter Balanced English Language Arts (ELA) Test was completed in March - June of 2017 by students in grades 3-4. (http://www.k12.wa.us/smarter/).

State Math Proficiency Level. Smarter Balanced Math Test scores were divided into four proficiency levels (i.e., Limited Knowledge, Not Proficient, Proficient, and Advanced) by the Smarter Balanced Assessment Consortium.

State Literacy Proficiency Level. Smarter Balanced ELA scores were divided into four proficiency levels (i.e., Limited Knowledge, Not Proficient, Proficient, and Advanced) by the Smarter Balanced Assessment Consortium.

Covariates. Additional variables were included in statistical models to adjust for certain types of bias; in particular, biased selection of higher performing students into higher quality classrooms. These additional variables included: Attendance (days attending summer program), Grade Level (K-4), Gender (% female), Limited English (% yes), Individualized Educational Plan (IEP) (% yes), Race (% White, Asian, Black, & Hispanic), 2015-2016 State Math Achievement, and 2015-16 State Literacy Achievement.

Data Collection

Observational data collection was conducted by Schools Out Washington (SOWA) using data collection and data management protocols approved by the Weikart Center and SPS. All raters had a reliability endorsement for the Summer Learning PQA. Raters observed each program for one entire 8:30 a.m. to 12:30 p.m. session on each of two days, at least 1.5 weeks apart, and produced one complete rating for each observation day. The first sets of observations were conducted between June 29th and July 21st, 2016, and the second set of observations were conducted between July 16th and July 27th, 2016. SPS staff coordinated collection of student assessment and background information and supplied the Weikart Center with a complete, de-identified data file for analysis.

Analytic Approach

The research questions were addressed using a three-step analytic approach that combined pattern-centered and linear models.[4] First, we used pattern-centered analyses (e.g., cluster analysis) to differentiate among summer learning offerings by identifying subgroups of instructional responsiveness profiles (i.e., quality). The three Summer Learning PQA scale scores were used as input variables to identify instructional responsiveness profiles; that is, subgroups of teachers whose instructional practices were similar. Additional detail is presented in Appendix A.

Next, we used linear models to estimate both academic skill change, and differences in academic skill change, across high, medium, and low instructional responsiveness profile subgroups. In one set of models, we sought to maximize the rigor of inference in relation to the threat of model misspecification and type II error.[5] We used the entire sample (grades K-4) to fit analysis of covariance (i.e., ANCOVA) models that compare rates of individual student growth (e.g., pre-to-post change) across the instructional responsiveness profile subgroups. These models allowed us to test a set of theoretically-driven planned-contrasts that improved both model specification and sample sizes within the subgroups being compared. Technical presentation of methodology and detailed results are presented in Appendix A.

In a second set of multilevel models, we sought to maximize the rigor of inference in relation to the threat of *selection bias* by using multilevel models and propensity matching based on a set of relevant covariates (e.g., state achievement test scores for the prior year). However, inclusion of the covariates limited the sample to grades 3 and 4, reducing the sample of settings to 31 and increasing potential for type II error. Technical presentation of methodology and detailed results are presented in Appendix B, authored by Albright (2017).

Finally, we used each of the modeling approaches to compare the rate of individual growth for students who entered the program at lower levels of academic skill (i.e., academic risk) across the quality subgroups. In addition to the three-step approach, we also conducted attrition analyses to see if students missing at T2 were different from the rest of the sample on the T1 measures. We found no statistically significant differences in 10 of the 15 tests conducted. See Appendix A for further discussion.

[4] This "skill growth by levels of quality" design has been used with some frequency in early childhood evaluations (e.g., Karoly, 2014; Thornburg, Mayfield, Hawks, & Fuger, 2009).

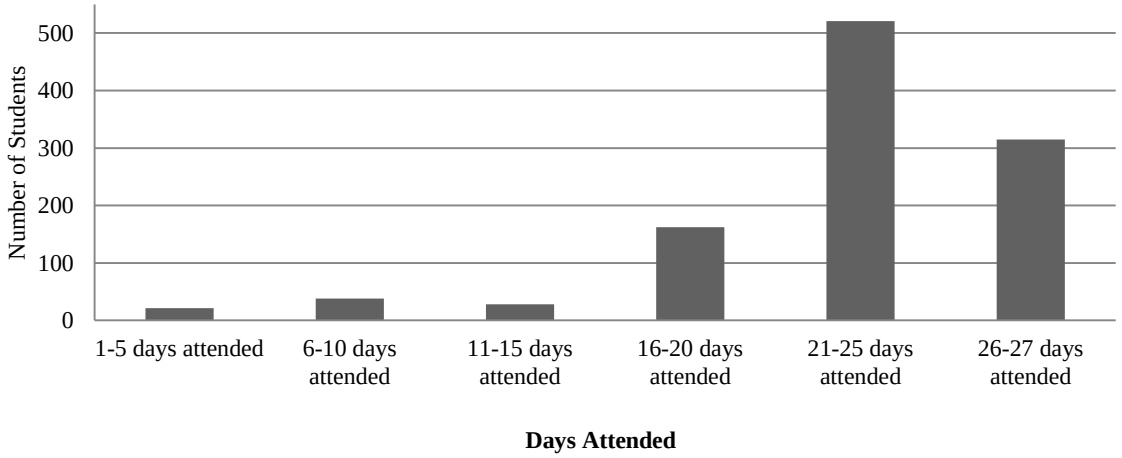
[5] A type II error is to falsely conclude that the null hypothesis is true (e.g., to conclude that the effect does not exist), which can occur when an analysis does not have enough statistical power to detect the effect.

Results

Attendance

The program was well attended, with most students (i.e., 77%) attending between 21 and 27 days of programming. Figure 2 describes the attendance of students.

Figure 2. Attendance in Seattle Public Schools Summer Programs, 2016



Source: Teacher Report

Academic Skill Change

Students in the study were assessed on literacy and math skills at the start and end of the summer program using the five academic performance measures. Using simple subtraction of raw pre-scores from post-scores, a majority of students demonstrated positive skill growth during the summer program period on all five academic performance measures. Similar percentages of students improved on academic achievement tests. Specifically:

- 81% of students showed improvement on *Site Words*, with a standardized pre-to-post effect size[6] of .59 during the summer session.
- 76% of student showed improvement on *Oral Fluency*, with a standardized pre-to-post effect size of .73 during the summer session.
- 64% of students showed improvement on *Math Assessment*, with a standardized pre-to-post effect size of .59 during the summer session.

[6] Cohen's d_z can be used to understand the size of the effect from Time 1 to Time 2 using a quantitative estimate that does not depend on sample size and that can be compared across variables or study samples. Cohen's d_z can be described as the "the standardized mean difference effect size for within-subjects designs" (Lakens, 2013, p. 4). It differs from the standard Cohen's *d* in that it is based on the average difference score between Time 1 and Time 2 divided by the standard deviation of the difference values, whereas the standard Cohen's *d* is the difference in group means divided by the pooled standard deviations of the two groups. Cohen's d_z also corrects for the auto-correlation between the Time 1 and Time 2 scores.

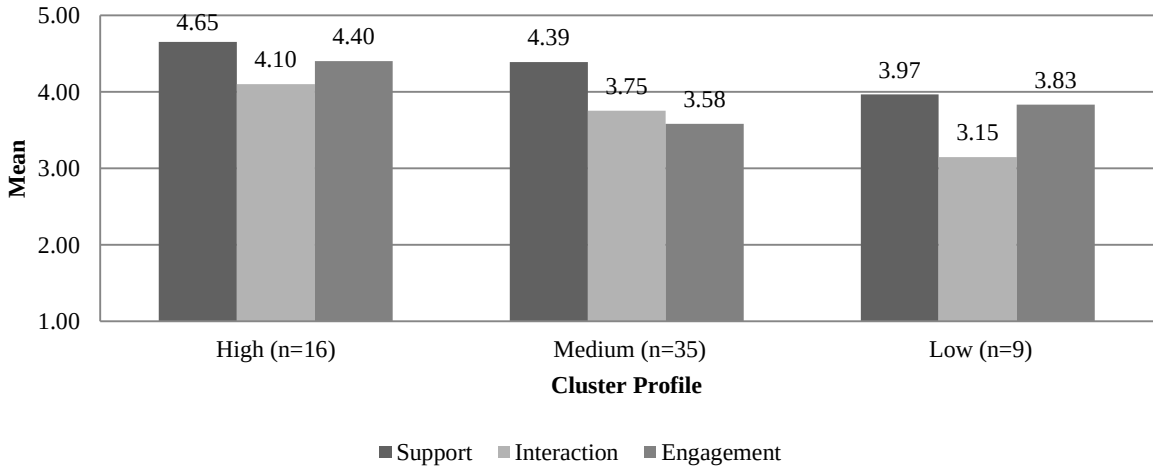
- 71% of students showed improvement on *Math Content*, with a standardized pre-to-post effect size of .95 during the summer session.
- 66% of students showed improvement on *Math Practice*, with a standardized pre-to-post effect size of .95 during the summer session.
- 76% of students improved on *State Math Achievement Test* from the 2015-16 school year to the 2016-17 school year, with a standardized pre-to-post effect size of .61 between annual assessments.
- 73% of students improved on *State Literacy Achievement Test* from the 2015-16 school year to the 2016-17 school year, with a standardized pre-to-post effect size of .56 between annual assessments.

Profiles of Instructional Responsiveness

Three profiles of instructional responsiveness were identified using data from the Summer Learning PQA scores for Supportive Environment, Interaction, and Engagement. These three scores for each of the 60 offerings were subjected to pattern-centered analysis to identify relatively-homogeneous subgroups of programs characterized by similar profiles of instructional responsiveness (see Figure 3).

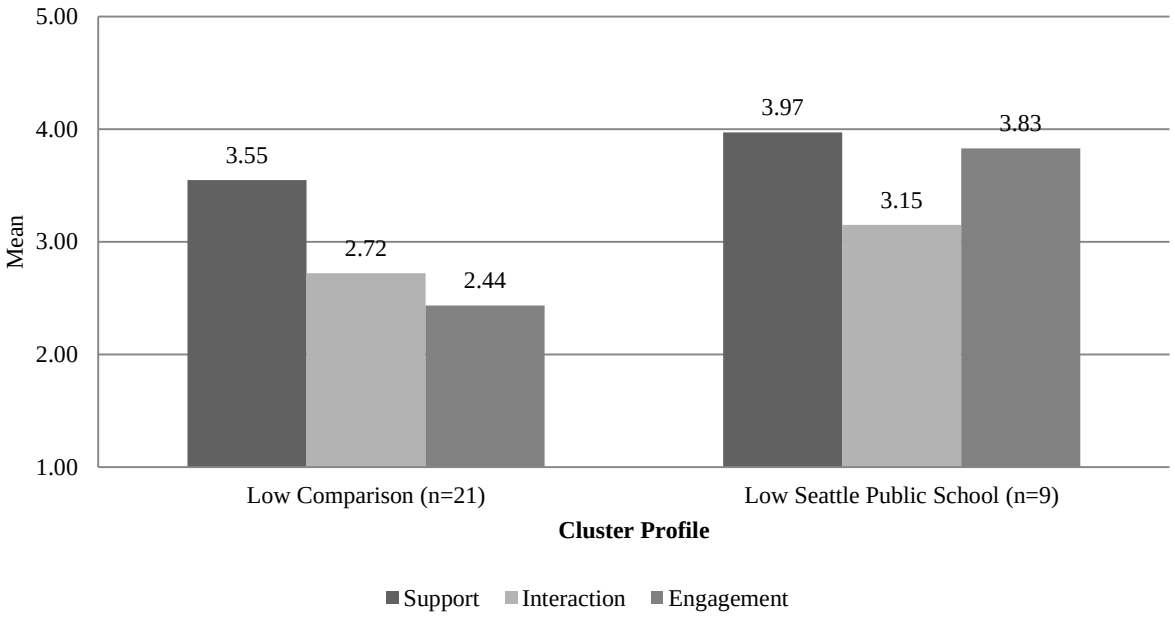
In general, scores for this sample of programs were all quite high. For reference, Figure 4 presents the "low" quality instructional profile from Figure 3 (n = 9) with the low profile from a different two-city sample (n = 21). Although this is good news for SPS summer programs, it also suggests that our ability to detect statistically reliable differences in rates of skill learning by profile may be constrained due to the absence of low-quality offerings.

Figure 3. Instructional Responsiveness Profiles



Source: Summer Learning Program Quality Assessment, 2016 (N = 60 program offerings)

Figure 4. "Low" Instructional Responsiveness Profiles for SPS and Other Cities' Programs



Source: Summer Learning Program Quality Assessment, 2016 (*N* = 60 program offerings); Low Comparison profile excerpted from Figure 4 in Smith et al., 2017.

Academic Skill Gains by Profiles of Instructional Responsiveness

This section presents gain scores on the five measures of academic performance and the two measures of academic achievement for students exposed to one of the profiles of instructional responsiveness. Figures 5-9 present simple gain scores (post-score minus pre-score) for academic performance of students in the program offerings identified by each of the three profiles of instructional responsiveness described Figure 3. Table 1 presents the percentage of students moving from below proficient to proficient or higher on the two academic achievement tests.

Literacy Academic Performance Change. Figures 5 and 6 show the gain scores for *Sight Words* and *Oral Fluency* assessments, each separated out by subgroups of students exposed to one of the three profiles of instructional quality. In each case, the expected relation between instructional quality and gains in academic performance is demonstrated: Gains are larger in the higher-quality offerings.

Figure 5. Sight Words Gains by Profiles of Instructional Quality

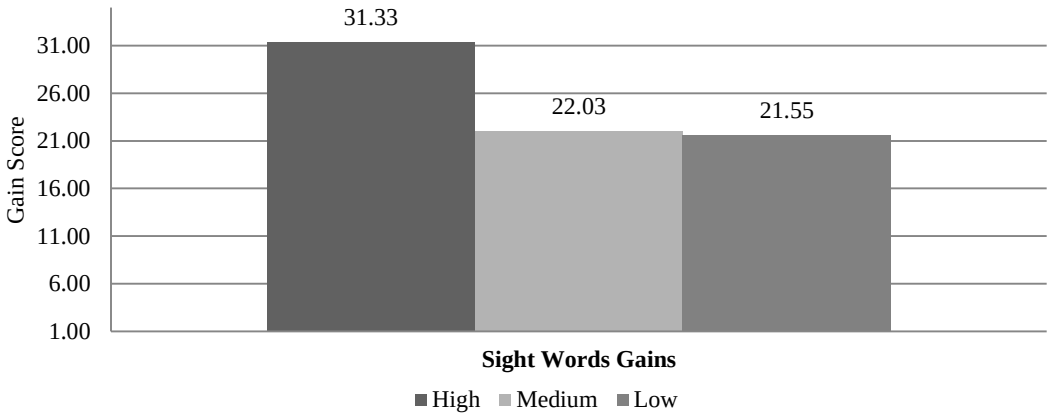
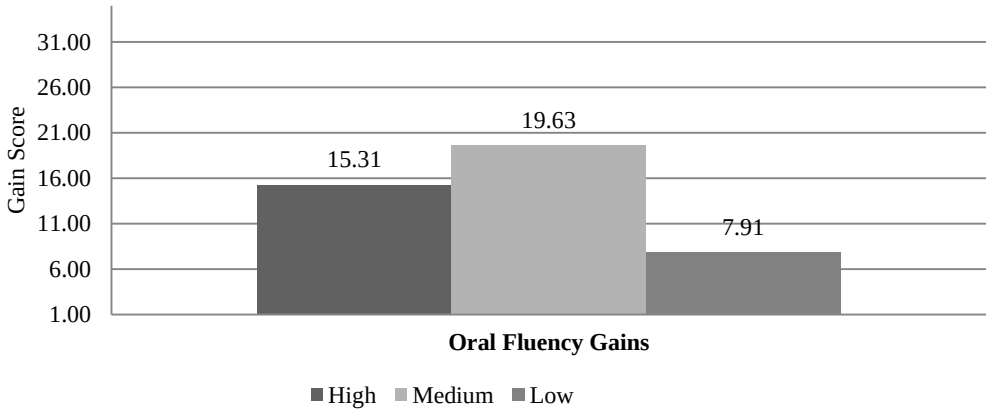


Figure 6. Oral Fluency Gains by Profiles of Instructional Quality



Math Academic Performance Change. Figures 7, 8, and 9 show subgroups of students in each of the three instructional quality profiles. In each case, the expected relationship is demonstrated: Gains are larger in the higher-quality offerings.

Figure 7. Math Assessment Gain Scores by Profiles of Instructional Quality

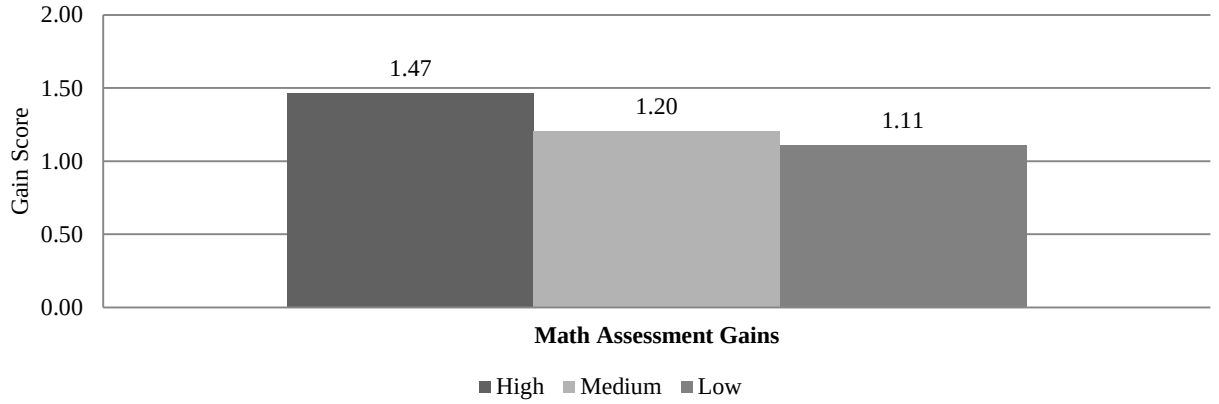


Figure 8. Math Content Gain Scores by Profiles of Instructional Quality

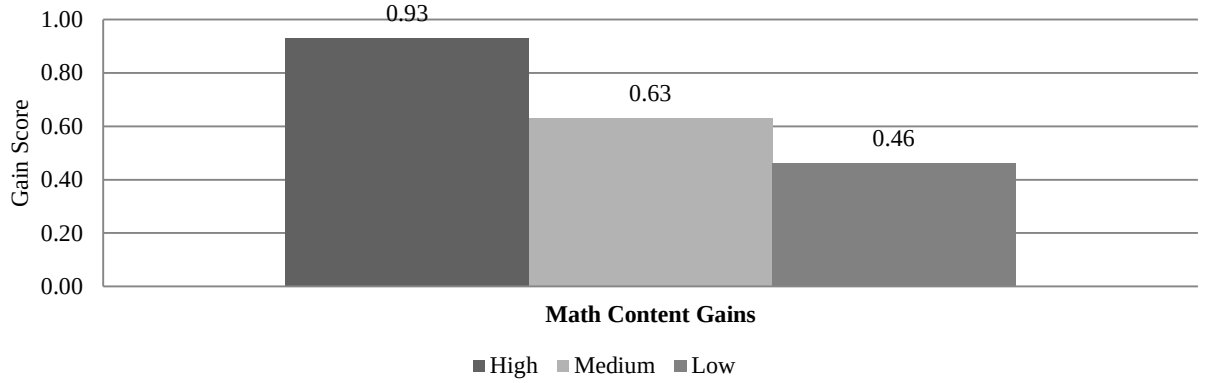
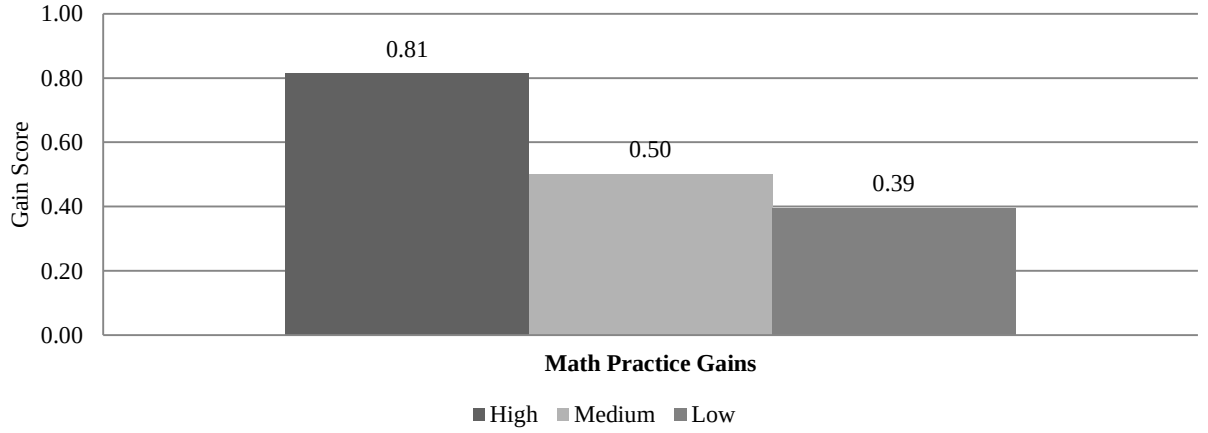


Figure 9. Math Practice Gain Scores by Profiles of Instructional Quality



Academic Achievement Change. Table 1 shows the number of SPS summer program students who moved from below proficient in the 2015 school year to proficient or better in the 2016 school year for both math and literacy achievement. In this case, the greatest gains are demonstrated by students exposed to the lowest-quality instructional profile.

Table 1. Percent of Students Changing from Non-Proficient to Proficient in Math and Literacy Achievement by Profiles of Instructional Quality.

	High	Medium	Low
Math Achievement	9%	6%	15%
Literacy Achievement	13%	9%	15%

Source: Seattle Public Schools

Multivariate Models for Grades K-4 (*N* = 1,069). We tested a series of statistical models that allowed us to specify when students exposed to the medium profile should be grouped with students in the higher or lower profiles (see Appendix A). In these models, statistically significant and positive relationships between exposure to the higher instructional responsiveness and academic skill gains were present for all five academic performance measures. No statistically significant relationship was detected for the two academic achievement measures.

Multivariate Models for Grades 3-4 (*N* = 600). We also tested multilevel statistical models designed to compare gains in student academic performance and academic achievement in the high versus low instructional quality profiles (see Appendix B). These models included two powerful covariates (i.e., the 2015-16 State Math Achievement Test and the 2015-16 State Literacy Achievement Test) in addition to the other six covariates listed in Table A-1. These covariates were used to construct propensity weights for matching students in the high and low instructional quality subgroups, thereby providing the most rigorous test of the association between exposure to high instructional responsiveness and academic skill gains. Overall, these models failed to detect any statistically significant relationship between exposure to higher levels of instructional responsiveness and greater academic skill gains.

Academically At-Risk Students

Finally, for all seven academic skill measures, we tested models that included only academically at-risk students, where academic risk was defined as receiving a below proficient score on the prior year's achievement test. The same pattern of mixed findings was demonstrated. In the more theoretically driven models with grades K-4, academically at-risk students showed greater gains on most academic performance measures where exposed to higher versus lower profiles of instructional responsiveness (see Appendix A). In the more rigorous, propensity-weighted models for grades 3 and 4, no statistically significant differences were detected; that is, more academically at-risk students performed similarly where exposed to higher versus lower profiles of instructional responsiveness (see Appendix B).

Discussion and Recommendations

This quality-outcomes study was designed to both (a) describe performance in Seattle Public Schools summer learning programs in ways that would be useful to staff and (b) provide evaluative evidence (i.e., validity) for an instructional model that includes challenging academic content and responsive instructional practices. In the study, a summer academic curriculum including both challenging content and responsive practices was implemented with elementary-age children. Three setting measures of responsive staff instructional practice and seven mental measures for student academic skill were collected at two time points: five academic performance measures administered near the beginning and end of the summer program and two academic achievement tests taken in the prior (2015-16) and subsequent (2016-17) school years. A series of statistical models were developed to address research questions related to the association between responsive instructional practices and development of academic skills.

Results from this study were mainly positive yet partially ambiguous. Summer program offerings were well-attended and characterized by high-quality instructional practices, with a majority of students increasing their literacy and math skills during the program. Findings about the association between exposure to more responsiveness instruction (e.g., quality) and academic skill change were mixed. Results include:

Positive academic skill change was found in the raw data, including for academically at-risk students. Positive change on the academic performance measures used during the summer program was found for 73% of students, and positive change on the academic achievement tests was found for 74% of students from the 2015 to 2016 school year. Standardized effect sizes for the full sample ranged from medium to large (i.e., d_z = .56 - .95) across the seven academic skill measures.

Attendance was regular, and instructional responsiveness was consistently high. Summer program attendance for 21 or more days (out of a total possible 27 days) was observed for 77% of students. Analysis of instructional responsiveness using the Summer Learning PQA revealed three profiles of instructional responsiveness at the point of service: high, medium, and low quality. However, compared to other urban samples, the "low" SPS profile is not very low.

Students in SPS summer programs had similar rates of skill change across profiles of instructional responsiveness in the most rigorous models for 3rd and 4th grade students (N = 535); that is, there was insufficient evidence in support of the hypothesized pattern of differential skill change across profiles of instructional quality. However, these results should be interpreted with caution due to the absence of a true low-quality instructional practices subgroup in the sample. Less statistically rigorous but more theoretically well-specified models for the entire K-4 sample (N = 1060) revealed a positive

association between instructional quality and academic skill change, despite the lack of a true low-quality subgroup.

Analyses of academically at-risk students revealed similarly mixed results. In the more statistically rigorous models with grades 3-4, students who entered SPS summer programs below proficient on academic achievement tests for the prior school year (2015-16) showed similar rates of academic skill change across profiles of instruction. In the theoretically well-specified models, academically at-risk students showed greater changes in academic skills in summer programs with higher-quality instructional practices.

Strengths and Limitations of the Study

This study has a number of specific strengths and limitations. In terms of strengths, the study includes (a) strong theory (see Figure 1) about how responsive instructional practices affect academic skill growth, (b) measures aligned to several elements of this theory, and (c) rigorous methods of addressing selection bias through the inclusion of relevant covariates and corresponding propensity score matching. Additionally, (d) the combination of observation-based assessment of instructional responsiveness and two time-point academic skill assessment provides a rare opportunity to describe relations between setting quality and skill growth. A further strength of the study is (e) the quality and completeness of the data. The study included high-quality observational data from trained raters integrated with moderately-complete math and literacy skill data from both teacher ratings and state achievement tests at two time points. This level of quality and completeness is unusual in applied research.

However, this study also includes limitations. Perhaps most importantly, the sample did not include a true low-quality profile of instructional responsiveness. This constraint on the range of the key predictor in all of the models likely resulted in constraints on our ability to address our primary research questions about the quality-outcome relations. Several additional challenges in the analyses should also be mentioned: First, a full moderation analysis that modeled the effects of exposure to high- and low-quality instructional practices for all risk groups (e.g., academically at-risk students and their non-risk peers) was not conducted due to limitations of resources. Second, the moderator selected in all analyses was below academic proficiency on the math or literacy state achievement tests. Although perhaps the most stringent choice as a moderator variable (i.e., less likely to be associated with the quality-outcome relations), it may not have been the best choice. For example, membership in the bottom quartile of summer program academic performance pre-test scores may have provided a more sensitive test of the moderating effects of instructional quality on academic skill gains.

Recommendations

One of the primary recommendations that follows from this study is to extend the analyses by including a no-summer-program-participation control group in the study design. Because SPS summer programs are implementing challenging academic content in combination with highly responsive instructional practices, more powerful and informative estimates of quality-outcome relations could be obtained by conducting statistical analyses similar to those reported here but where including a no-summer-program-participation control group that would practically ensure a true low-quality instructional responsiveness contrast group. Our recommendation to the Evaluation Department on how best to obtain data for such a no-summer-program-participation control group comparison is summarized in Appendix C. Additionally, obtaining data that would allow us to construct a variable indicating which individuals attended two consecutive years of the Summer Staircase program may also increase the power of the evaluation. Finally, adding SEL measures administered during both summer programs (e.g., emotion management) and the school year (e.g., suspensions) could bolster our understanding of the relations among SEL skill growth, academic skill change, and responsive instructional practices.

References

Bergman, L. R., Magnusson, D., & El-Khouri, B. M. (2003). *Studying individual development in an interindividual context: A person-oriented approach*. Mahwah, NJ: Erlbaum.

Karoly, L. A. (2014). *Validation Studies for Early Learning and Care Quality Rating and Improvement Systems: A Review of the Literature*.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers In Psychology, 4*, 1-12.

Murray, D. W., Rosanbalm, K., Christopoulos, C., & Hamoudi, A. (2015). *Self-regulation and toxic stress: Foundations for understanding self-regulation from an applied developmental perspective*. OPRE Report #2015-21, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Smith, C., Helegda, K., Ramaswamy, R., Hillaker, B., McGovern, G., & Roy, L. (2015). *Quality-Outcomes Study for Seattle Public Schools Summer Programs: Summer 2015 Program Cycle*. Retrieved from Ypsilanti, MI: www.cypq.org/publications

Smith, C., Ramaswamy, R.,Helegda, K., Macleod. C., Hillaker, B., Borah, P., Peck. S. C.(2017). *Design Study for the Summer Learning Program Quality Intervention (SLPQI): Final-Year Intervention Design and Evaluation Results.*Retrieved from Ypsilanti, MI: www.cypq.org/publications

Thornburg, K. R., Mayfield, W. A., Hawks, J. S., & Fuger, K. L. (2009). The Missouri quality rating system school readiness study. Columbia, MO: Center for Family Policy & Research.

Vargha, A., Torma, B., & Bergman, L. R. (2015). ROPstat: A general statistical package useful for conducting person-oriented analyses. *Journal for Person-Oriented Research, 1*, 87-98.

Appendix A. Methodology and Results

Descriptive Statistics and Attrition Analyses

Instructional responsive practice measures, student academic skill measures, and the covariates used in the models described below are described in the measures section of the main report. The descriptive statistics associated with these measures are shown in Table A-1.

Table A-1. Descriptive Statistics for all Study Variables.

	N	*M*	*SD*	Minimum	Maximum
Supportive Environment	60	4.39	0.31	3.29	4.83
Interaction	60	3.76	0.48	2.83	5.00
Engagement	60	3.84	0.46	2.69	4.67
Sight Words T1	1067	90.02	41.26	0	398
Oral Fluency T1	594	104.01	45.19	0	216
Math T1	1065	6.54	2.60	0	10
Math Content T1	493	3.09	0.89	1	5
Math Practice T1	430	3.20	0.85	1	5
Math Achievement T1	605	2420.07	81.68	2129	2670
Literacy Achievement T1	581	2397.64	79.93	2179	2608
Sight Words T2	939	115.51	61.40	1	700
Oral Fluency T2	514	120.13	45.46	0	219
Math T2	967	7.86	2.16	0	10
Math Content T2	423	3.75	0.87	1.4	5
Math Practice T2	425	3.88	0.81	1.3	5
Math Achievement T2	776	2442.04	80.97	2160	2728
Literacy Achievement T2	778	2416.89	87.57	2129	2675
Attendance	1093	22.18	5.28	0	27
Grade Level	1093	2.66	1.11	.75	5
Gender (% Female)	1076	49	.50	0	1
Limited English (% Yes)	1076	51	.50	0	1
Individualized Education Plan (% Yes)	1076	21	.41	0	1
Race/Ethnicity (% White)	1093	10	.30	0	1
Race/Ethnicity (% Asian)	1093	23	.42	0	1
Race/Ethnicity (% Black)	1093	31	.46	0	1
Race/Ethnicity (% Hispanic)	1093	25	.44	0	1

In order to evaluate the extent to which missing data for youth participants who dropped out of the study might influence our statistical models, we used Chi-Square and *t*-test analyses to test for significant differences between respondents who participated in program assessments at both Time 1 (T1) and Time 2 (T2) and respondents who participated in program assessments at T1 only (i.e., they were missing all data at T2). The results indicated no significant differences for most study variables – of 15

tests, 5 were statistically significant. Youth who dropped out of the study were more likely to be from programs with lower Interaction scores and in the low profile of instructional responsiveness. Students who dropped out were also less likely to be classified as Asian and scored lower on T1 Sight Words and Math Practices. In short, although we imputed some missing data for the cluster input variables, we made no attempt to adjust for the missing data patterns associated with the covariates and academic skill variables; consequently, it is possible that differential attrition may have biased the parameter estimates in some of the models.

Instructional Responsiveness Profiles

Instructional quality was evaluated using the Summer Learning Program Quality Assessment (PQA) administered during the 2016 summer cycle. Pattern-centered analyses (Bergman, Magnusson, & El-Khouri, 2003) of the Summer Learning PQA scores for Supportive Environment, Interaction, and Engagement (as averaged across two observation periods) were conducted using the ROPstat (2.0) statistical package for pattern-oriented analyses (Vargha, Torma, & Bergman, 2015). After using ROPstat modules for addressing missing data and multivariate outliers, we used Ward's method (with squared Euclidian distances as the dissimilarity measure), followed by *k*-means cluster relocation analyses, to identify profiles of instructional responsiveness (aka, instructional quality profiles).

An optimal cluster solution was determined using a combination of (a) the relative drop in homogeneity coefficients between each solution and a one-cluster solution, (b) a scree plot to evaluate changes in the error sum of squares between solutions and the cumulative explained variance of each solution, and (c) the theoretical interpretation of the profiles associated with each solution. After the optimal sample-level solution was identified (i.e., the six-cluster solution), that initial Ward's solution was subjected to a *k*-means cluster relocation analysis that re-assigned each site to its best matching sample-level profile, thereby correcting for premature classification by the hierarchical (i.e., Ward's) algorithm and further increasing within-group homogeneity. Finally, to minimize complexity, the six profile solution was reduced to three – the high, medium, and low profiles shown in Figure 3 – by combining clusters with similar profile shapes.

Academic Skill Growth Models

In order to evaluate the effects of instructional responsiveness on academic skill growth, a base general linear model was developed and then applied to each of the academic performance and achievement variables. Each of the ANCOVA models consisted of post-test scores as the dependent variable, instructional responsiveness profiles as the independent variable, and the covariates (listed in the measures section of the main report) as control variables. The first covariate was the pre-test score. The models were examined in several ways, including:

- *Estimated Marginal Means* (EMM). These are the covariate-adjusted scores on the dependent variable.
- *Overall model significance*. This tells us whether the covariate-adjusted scores on the dependent variable vary across the instructional quality profiles, at the .05 alpha level.
- *Beta coefficients and alpha levels*. These tell us about the strength of the relation between instructional quality and the covariate-adjusted dependent variable score as well as if this effect estimate differs significantly from 'no effect' at the .05 alpha level.
- *L-Matrix contrasts*. These are contrasts between EMMs for theoretically-specified subsets of the instructional quality profiles.

Summary results for overall model significance are shown in Table A-2. The results indicated that each of the five in-program academic performance score gains differed significantly across instructional profiles, whereas the state academic achievement test score gains did not differ significantly across profiles. Analyses of the state academic achievement test variables had reduced sample sizes, as these measures were only administered to students in grades 3 and 4 (see Table A-1).

Table A-2. Omnibus ANCOVA model results.

Dependent Variable	Omnibus *F*	*p*
Sight Words	3.13	.044
Oral Fluency	7.17	> .001
Math	4.74	.009
Math Content	10.21	> .001
Math Practice	9.81	> .001
Academic Math	0.81	.447
Academic Reading	0.79	.456

The T2 academic skill EMMs for each instructional quality profile are listed in Table A-3. The results indicated that, for the academic performance measures, the greatest gains were seen for youth in either the high profile or a combination of medium and high profiles. The state academic achievement test measures did not vary systematically across the instructional profiles. Figure A-1 presents the EMMs by profile in a line graph so that the relative changes across profiles of instructional quality can be examined on the same scale.

Table A-3. Estimated Marginal Means for T2 Academic Skills by Instructional Quality Profiles.

	Instructional Quality Profile		
Dependent Variable	Low	Medium	High
Sight Words	114.03	113.10	120.79
Oral Fluency	113.42	122.58	119.66
Math	7.87	7.74	8.13
Math Content	3.58	3.73	3.96
Math Practice	3.64	3.95	3.98
Math Achievement	2456.14	2449.48	2451.97
Literacy Achievement	2426.06	2430.31	2435.03

Figure A-1. Estimated Marginal Means Plot for Standardized T2 Academic Skill Variables.

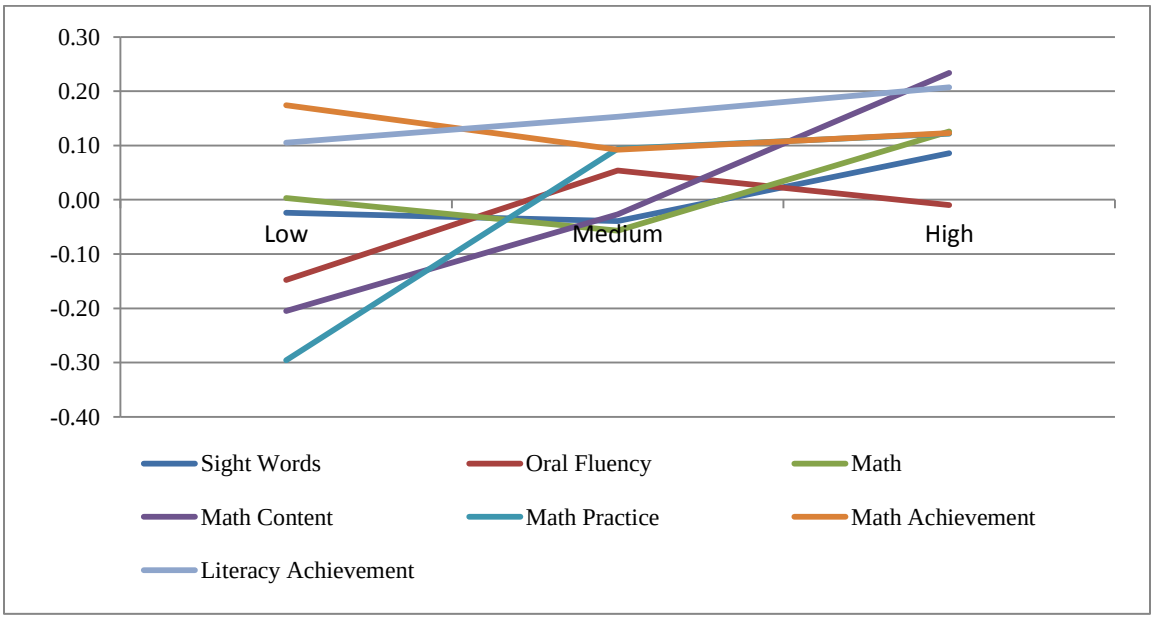


Table A-4 shows unstandardized beta coefficients from the base ANCOVA models corresponding to the default contrast in academic skill gains between youth exposed to "low" versus high-quality profiles of instructional responsiveness. The results indicated significant differences between low- and high-quality profiles for three of the five academic performance measures but neither of the two achievement test measures. However, inspection of the pattern of EMMs across the three profiles of instructional responsiveness suggested that the threshold for sufficient quality to promote skill change may vary across curriculum content (e.g., change in some skills may be facilitated by even minimal forms

of instructional responsiveness, whereas change in other skills may require more well-developed forms of instructional responsiveness). Consequently, we re-ran the ANCOVA models where including a series of planned contrasts corresponding to what appeared as the most meaningful functional difference in instructional quality for each academic variable. As shown in Table A-5, these contrasts were between the high and medium/low (combined) profiles or the high/medium (combined) and low profiles. The results indicated that each of the academic performance score gains differed significantly across the profile contrasts, whereas the academic achievement test gains did not differ across the profile contrasts.

Table A-4. Unstandardized Beta Coefficients for Academic Gains across High versus Low Profiles of Instructional Quality.

Dependent Variable	*b*	*p*
Sight Words	6.76	.123
Oral Fluency	6.25	.026
Math	0.27	.137
Math Content	0.38	> .001
Math Practice	0.34	> .001
Math Achievement	-4.17	.486
Literacy Achievement	8.97	.211

Table A-5. Planned Contrasts for Academic Gains across Higher versus Lower Profiles.

Dependent Variable	Planned Contrast	*F*	*p*
Sight Words	High vs. Low/Med	5.02	.025
Oral Fluency	High/Med vs. Low	11.00	> .001
Math	High vs. Low/Med	6.26	.013
Math Content	High vs. Low/Med	19.64	> .001
Math Practice	High/Med vs. Low	19.62	> .001
Math Achievement	High/Med vs. Low	1.18	.278
Literacy Achievement	High/Med vs. Low	1.22	.269

Academically At-Risk Students

The ANCOVA models were also run separately two more times for subgroups of youth considered academically ‘at risk’ in math or reading. ‘At risk’ was defined as being in the lowest two proficiency levels based on the state academic achievement test scores from the 2015-16 year. The results of these additional models are summarized in Tables A-6 and A-7.

Table A-6: Model Summaries for Students with Academic Risk in Math.

Dependent Variable	Omnibus *F*	*p*	High vs. Low Quality beta	*p*	Planned Contrast	*p*
Sight Words	2.88	.057	-10.51	.017	High vs. Low/Med	.024
Oral Fluency	3.13	.045	8.56	.038	High/Med vs. Low	.014
Math	0.33	.719	0.20	.550	High vs. Low/Med	.441
Math Content	4.90	.009	0.44	.002	High vs. Low/Med	.007
Math Practice	4.24	.017	0.37	.009	High/Med vs. Low	.004
Math Achievement	1.38	.254	-12.33	.118	High/Med vs. Low	.099
Literacy Achievement	0.06	.945	2.80	.749	High vs. Low/Med	.738

Table A-7. Model Summaries for Students with Academic Risk in Literacy.

Dependent Variable	Omnibus *F*	*p*	High vs. Low Quality beta	*p*	Planned Contrast	*p*
Sight Words	4.07	.018	-12.64	.005	High/Med vs. Low	.006
Oral Fluency	1.53	.219	4.27	.232	High/Med vs. Low	.107
Math	0.24	.784	0.01	.977	High vs. Low/Med	.758
Math Content	5.07	.008	0.43	.002	High/Med vs. Low	.003
Math Practice	3.49	.034	0.31	.035	High/Med vs. Low	.010
Math Achievement	1.07	.346	-10.01	.181	High/Med vs. Low	.145
Literacy Achievement	0.04	.962	-1.43	.873	High/Med vs. Low	.810

Appendix B. Methodology and Results for Multi-level Models with Propensity Matching for Grades 3-4

The following is a report for Seattle Public Schools authored by Jeremy Albright (November, 2017).

Data and Methodology

The intent of this analysis is to test four sets of hypotheses:

1. Does summer program quality impact proficiency on standardized tests at the end of the program, controlling for scores at the start of the program?
2. Does summer program quality impact proficiency on standardized tests at the end of the program, controlling for baseline scores, previous school year's state standardized assessments, race, grade, gender, English proficiency, and free/reduced lunch status?
3. Are scores on state standardized assessments taken the year after the summer program significantly better in high quality programs, controlling for prior year state assessments, baseline summer proficiency scores, race, grade, gender, English proficiency, and free/reduced lunch?
4. For the subset of low achieving students, does program quality moderate the relationship between prior year assessments and current year assessments?

The pre/post summer program tests that will be analyzed are oral fluency, sight words, and math. In addition, state assessments for reading and math will also be analyzed. English proficiency is coded dichotomously, as is eligibility for free/reduced school lunch. The race categories are White, African American, Hispanic, Asian, and Other/Unknown. Site quality is classified as high, medium or low based on a previous latent profile analysis.

There was one outlier in the summer 2015-2016 pre-program math scores. This value has been recoded as missing for the analysis.

Data were collected from 59 different programs, and the first hypothesis is tested using all available data. The subsequent hypotheses are limited to 3rd and 4th graders due to the unavailability of state assessment data. Summary statistics for the variables available for the entire set of data are the following:

Table B-1. Summary Statistics, All Cases

Variable	Categories	Observed	% Missing	Mean or N	SD or %
Site Quality, n (%)	High	1069	0	324	30.31
Site Quality, n (%)	Medium			563	52.67
Site Quality, n (%)	Low			182	17.03
Sight Words (Pre), M (SD)		1043	2.43	89.84	41.34
Sight Words (Post), M (SD)		934	12.63	115.53	61.51

Oral Fluency (Pre), M (SD)	589	44.9	103.96	45.02
Oral Fluency (Post), M (SD)	510	52.29	119.89	45.44
Math Assessment (Pre), M (SD)	1041	2.62	6.56	2.70
Math Assessment (Post), M (SD)	962	10.01	7.86	2.16

Summary statistics for the variables in the analysis of 3rd and 4th graders are:

Table B-2. Summary Statistics - 3rd and 4th Graders Only

Variable	Categories	Observed	% Missing	Mean or N	SD or %
Site Quality, n (%)	High	600	0	149	24.83
Site Quality, n (%)	Medium			296	49.33
Site Quality, n (%)	Low			155	25.83
Grade Level, n (%)	4	600	0	276	46.00
Gender, n (%)	Female	600	0	299	49.83
Race, n (%)	White	600	0	56	9.33
	African American			185	30.83
	Hispanic			156	26.00
	Asian			145	24.17
	Other/Unknown			58	9.67
Limited English, n (%)	ESL	600	0	282	47.00
Individualized Ed. Plan, n (%)	IEP	600	0	151	25.17
Sight Words (Pre), M (SD)		583	2.83	84.01	30.61
Sight Words (Post), M (SD)		527	12.17	105.68	35.26
Oral Fluency (Pre), M (SD)		549	8.5	104.59	45.87
Oral Fluency (Post), M (SD)		486	19	120.28	45.69
Math Assessment (Pre), M (SD)		580	3.33	6.38	2.89
Math Assessment (Post), M (SD)		535	10.83	7.59	2.38
State Math 2015-2016, M (SD)		585	2.5	2419.72	80.60
State Math 2016-2017, M (SD)		539	10.17	2453.23	81.06
State Reading 2015-2016, M (SD)		562	6.33	2396.78	79.48
State Reading 2016-2017, M (SD)		538	10.33	2428.57	89.39

All hypotheses will be tested using linear mixed models that include a random effect for the site. The first set of hypotheses will include the pre-program tests as covariates and will be fit to all available students. The subsequent hypotheses are fit only to 3rd and 4th graders and will incorporate a larger set of covariates. Rather than include all of the possible confounders in the models, the adjustment will take place by weighting on the basis of propensity scores fit using Gradient Boosted Machines (GBMs). The benefit of GBMs versus logistic regression in calculating propensity scores is that the former implicitly introduce far more nonlinearities to the model, thereby generating more accurate predicted probabilities. The propensity scores are then converted to weights that are intended to bring the distribution of

confounders between treatment levels closer to each other. A weighted mixed model is then fit with site quality as the sole predictor and program again included as a random effect.

Results

Within-Program Change

The first hypotheses ask whether post-program scores on the three assessments - oral fluency, sight words, and math - are significantly different in high versus low or medium quality settings. The following figures display the unadjusted comparisons, which show significant improvements from pre to post for all three evaluations.

Figure B-1. Mean Sight Words Assessment

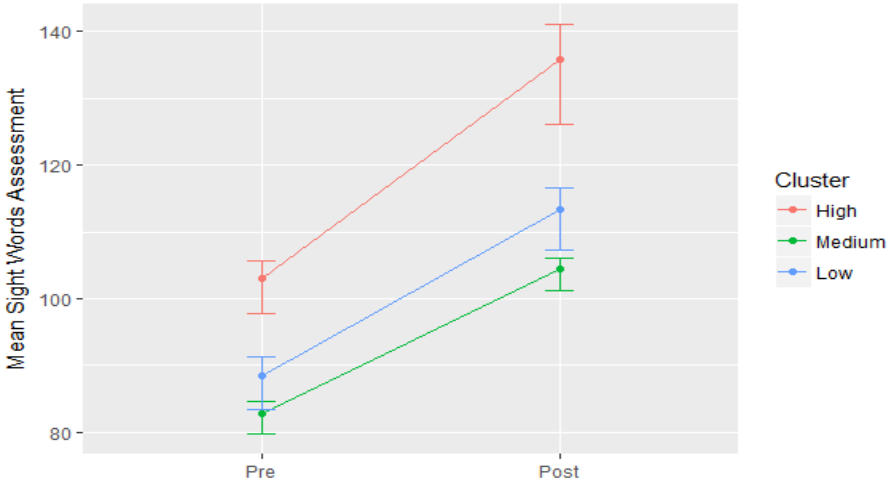


Figure B-2.Mean Oral Fluency Assessment

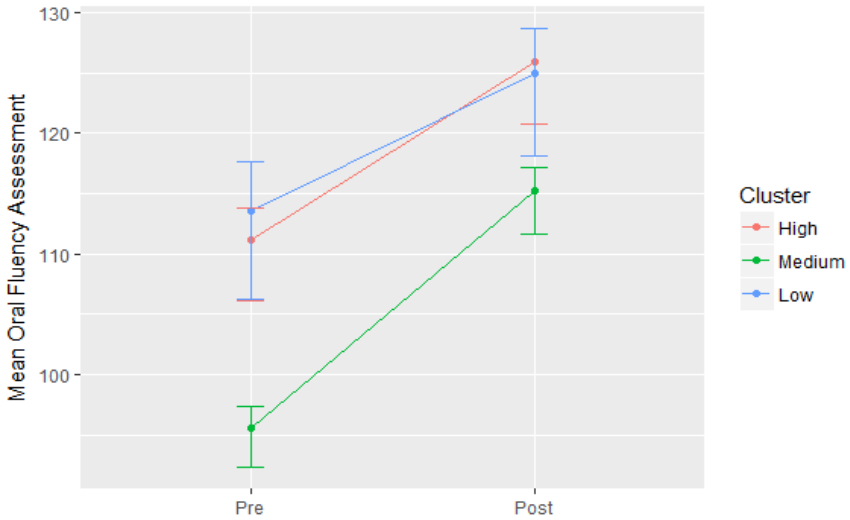
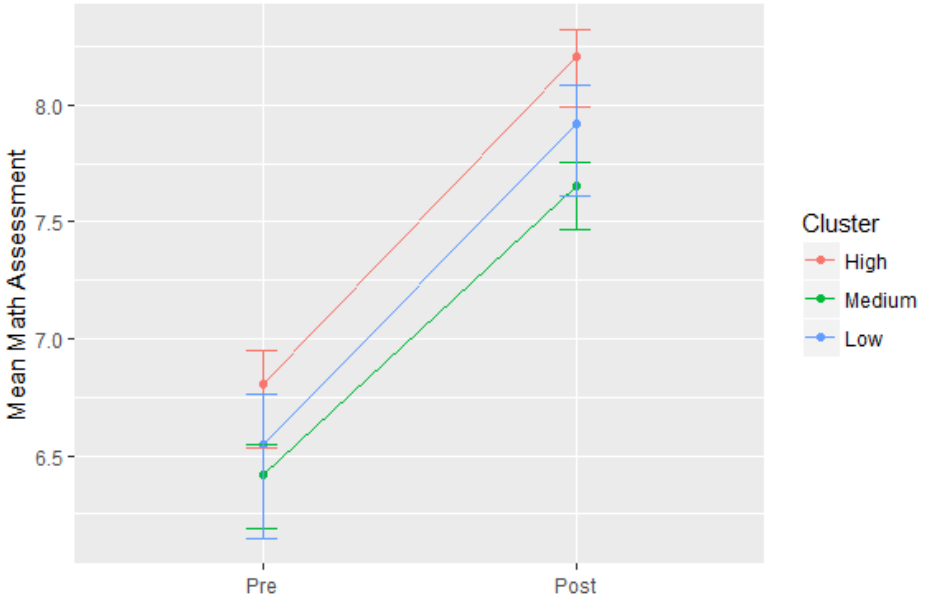


Figure B-3. Mean Math Assessment



The following table shows the results for the statistical analysis of each outcome. The rows labeled "Pre-Program Evaluations" in the table correspond to the respective assessment (i.e. pre sight words for modeling post sight words, pre oral fluency for modeling post oral fluency). The effect of a unit increase in pre-program sight words scores is to increase post-program scores by 1.07, which is statistically significant. The mean scores on post-program evaluations are 4.57 lower in the medium group relative to the high quality group, and 5.65 lower in the low quality group relative to the high quality group, but these comparisons are not statistically significant. Looking at oral fluency, each unit increase in pre-program scores is associated with a .95 increase in post-program scores, which is statistically significant. Mean oral fluency scores are 3.75 higher in the medium group relative to the high group and 6.24 lower in the low group relative to the high group, but these comparisons are not statistically significant. Finally, each unit increase on pre-program math scores is associated with a .435 increase in post-program scores, which is significant. Mean post-program math scores are .36 lower in the medium group relative to the high group and .26 lower in the low quality group relative to the high quality group, but these comparisons are not significant.

Table B-3. Post program Outcomes

	Dependent variable:		
	Sight Words	Oral Fluency	Math
	(1)	(2)	(3)
Pre-Program Evaluations	1.069***	0.923***	0.435***
	(0.029)	(0.022)	(0.022)
Quality = Medium	−4.574	3.745	−0.362
	(8.576)	(4.579)	(0.221)
Quality = Low	−5.652	−6.241	−0.255
	(11.729)	(5.617)	(0.306)
Constant	20.077**	23.726***	5.226***
	(7.614)	(4.525)	(0.233)
Observations	933	508	958
Log Likelihood	−4,545.251	−2,236.933	−1,908.257
Akaike Inf. Crit.	9,102.503	4,485.866	3,828.514
Bayesian Inf. Crit.	9,131.533	4,511.249	3,857.703

Note: *p<0.05; **p<0.01; ***p<0.001

Within Program Change - 3rd and 4th graders

It is possible to extend the previous analysis to include additional covariates. However, due to constraints in the availability of state assessment data, the analysis is limited to students in the 3rd and 4th grades. Propensity score weighting will be used to adjust for pre-treatment differences given the expansion of the covariate set. The variables going into the propensity score model are grade, gender, race, English proficiency, free/reduced lunch, pre-treatment scores for the summer assessments, and 2015-2016 state reading and math assessments.

The following are graphs that visualize the unadjusted differences (without the propensity score weights) in outcomes. Examining the distributions before any adjustments are made provides a baseline against which the adjusted results can be compared. As the figures show, there are not large differences in post-program test scores between clusters. The median student performs best in the low quality sites for sight words and in the high quality sites for math, but the boxes overlap substantially and therefore do not indicate substantial variation from one site to the next. If selection bias is a problem, students have not selected into the different quality levels in a manner that causes average scores to appear different.

Figure B-4. Sight words (Post)

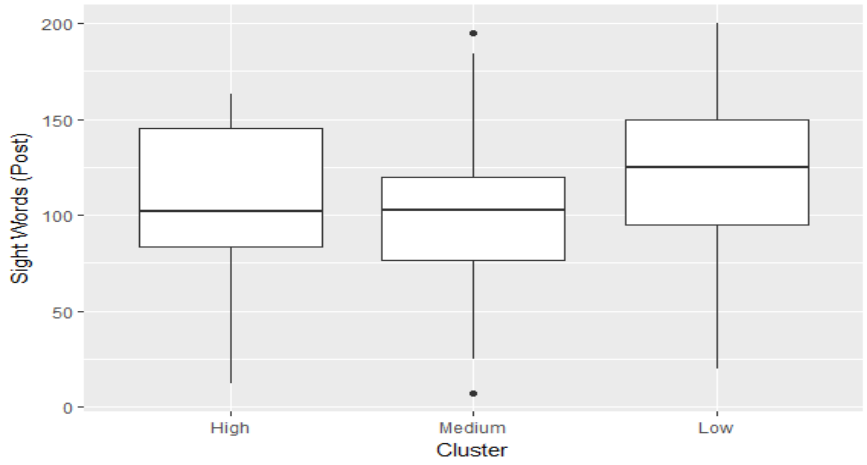


Figure B-5. Oral Fluency (Post)

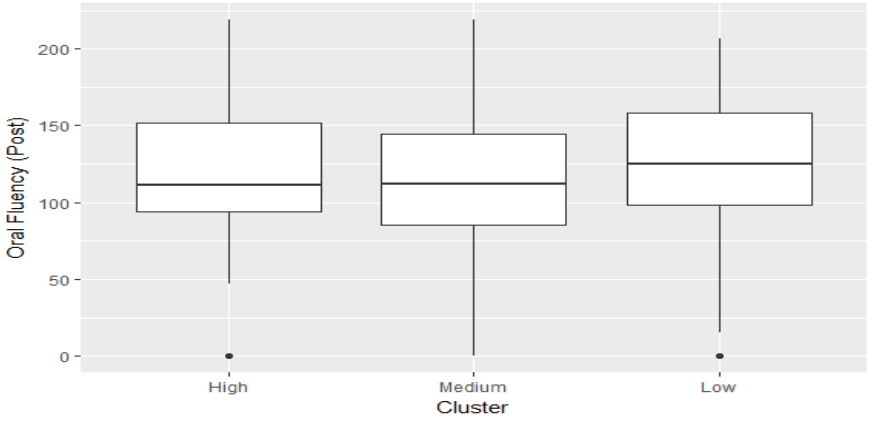
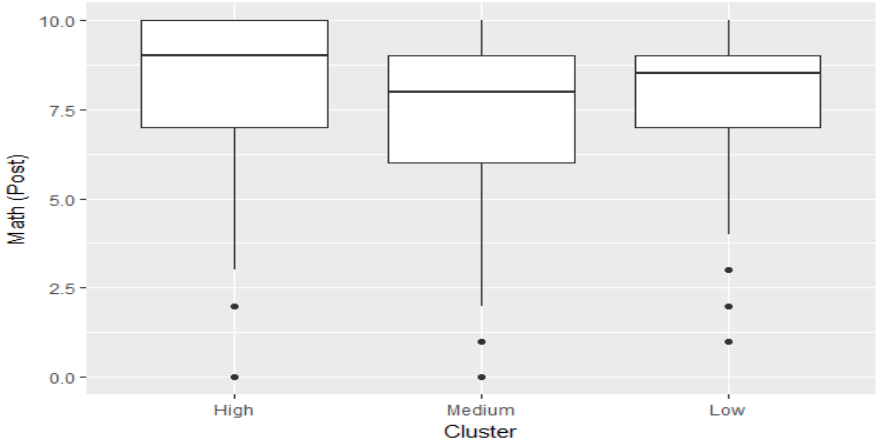


Figure B-6. Math Assessment (Post)



The following boxplots incorporate the propensity scores as weights. The boxes once again show a good deal of overlap between site quality. The weighting actually causes the median math scores to become nearly indistinguishable, which is consistent with the interpretation that the small differences observed in the unadjusted figures were due to selection bias.

Figure B-7. Sight Words (Post) Box Plot

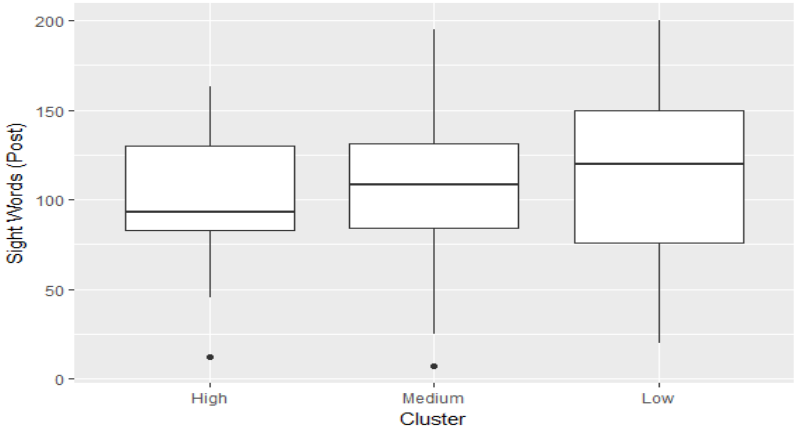


Figure B-8. Oral Fluency (Post) Box Plot

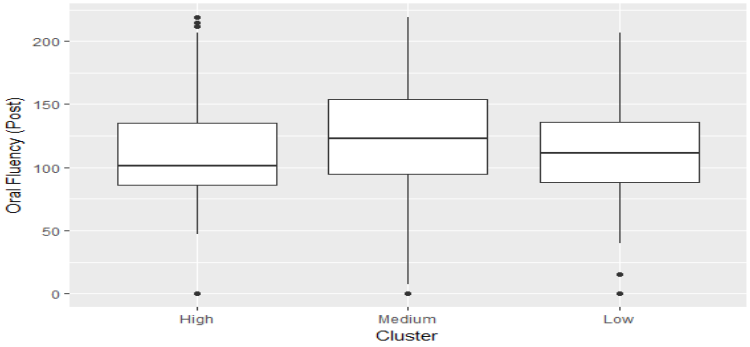
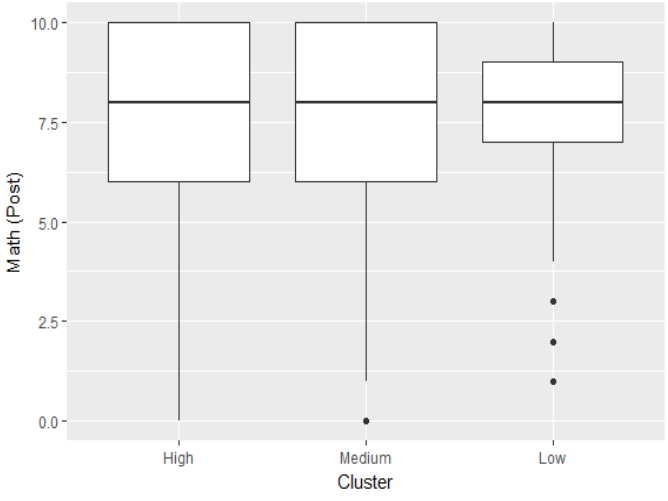


Figure B-9. Math Assessment (Post) Box Plot



The following table shows the results of the mixed models estimated on the subset of matched comparisons. The results can be interpreted as traditional (weighted) regression estimates, though the standard errors are different due to the inclusion of the random effect for site. Adjustments for possible pre-treatment confounders have been made by the weights, so the only predictors included are dummies for medium and low site quality (high quality is the baseline). The results show that the *mean* sight words score (as opposed to the median highlighted in the boxplots) is 2.73 lower in the medium group relative to the high group, while the mean low score is 4.13 higher than the high quality sites. For oral fluency, the mean score is 8.89 higher in the medium group relative to the high group and 4.04 higher in the low group relative to the higher group. The mean math score is .11 lower in the medium group relative to the high group and .11 higher in the low group relative to the high group. None of these comparisons are statistically significant.

Table B-4. Post program Outcomes - Propensity Score Matched Subsample

	Sight Words	Oral Fluency	Math
	(1)	(2)	(3)
Quality = Medium	−2.725	8.888	−0.113
	(11.697)	(11.028)	(0.332)
Quality = Low	4.126	4.040	0.105
	(13.784)	(12.864)	(0.350)
Constant	104.954***	115.014***	7.718***
	(9.700)	(9.087)	(0.244)
Observations	527	486	535
Log Likelihood	−2,552.511	−2,570.745	−1,294.865
Akaike Inf. Crit.	5,115.023	5,151.491	2,599.731
Bayesian Inf. Crit.	5,136.359	5,172.422	2,621.142

Note: *p<0.05; **p<0.01; ***p<0.001

As a robustness check in case the propensity score model was poorly specified, the following table presents results based on a traditional unweighted linear mixed model that adjusts for covariates by directly including them in the model. The models again include a random effect for site but can otherwise be interpreted as a standard regression model.

The covariate-adjusted mean sight words score was 5.42 higher in the medium group relative to the high quality group, and the low quality group had mean scores that were 11.29 higher than the high quality group. The medium quality group had oral fluency scores that were 5.13 higher relative to the high group, and the low quality group had oral fluency scores that were 4.17 lower relative to the higher group. The medium group's mean math scores were .03 lower relative to the high quality group, and the

low quality sites had scores that were .05 lower than high quality sites. However, none of these results are statistically distinguishable from zero.

Table B-5. Post-Program Outcomes - Propensity Score Matched Subsample

	Sight Words	Oral Fluency	Math
Quality = Medium	5.422	5.129	-0.028
	(9.479)	(4.413)	(0.300)
Quality = Low	11.286	-4.174	-0.052
	(11.305)	(5.307)	(0.354)
Grade Level	4.516**	-1.127	0.090
	(1.597)	(1.979)	(0.174)
Female	-0.453	-0.893	0.175
	(1.576)	(1.916)	(0.170)
African American	0.215	0.209	-0.081
	(3.025)	(3.718)	(0.322)
Hispanic	0.886	0.167	-0.408
	(2.961)	(3.718)	(0.323)
Asian	0.454	-1.496	0.168
	(3.133)	(3.820)	(0.333)
Other	-0.098	0.447	-0.109
	(3.445)	(4.304)	(0.372)
ESL	3.397	-1.374	-0.105
	(1.769)	(2.141)	(0.191)
Special needs	-4.225*	3.392	-0.157
	(1.995)	(2.372)	(0.210)
Sight Words – Pre	0.479***	0.222***	0.007
	(0.042)	(0.047)	(0.004)
Oral Fluency – Pre	0.205***	0.774***	0.001
	(0.029)	(0.032)	(0.003)
Math Assessment 2015-2016	-0.006	-0.017	0.006**
	(0.017)	(0.021)	(0.002)
Reading Assessment 2015-2016	0.008	0.053**	0.001
	(0.016)	(0.019)	(0.002)
Constant	14.476	-64.009	-12.923***
	(37.559)	(44.829)	(3.924)
Observations	467	450	473
Log Likelihood	-1,985.169	-1,964.055	-967.552
Akaike Inf. Crit.	4,006.337	3,964.110	1,971.105
Bayesian Inf. Crit.	4,080.971	4,038.077	2,045.968

Note: *p<0.05; **p<0.01; ***p<0.001

Change in State Assessments - 3rd and 4th Graders

The next analysis treats the 2016-2017 school year state assessments as the outcome. Visualizations of the unadjusted differences (without the propensity score weights) in outcomes are shown in the following boxplots. The medians are roughly similar between site quality groupings, and the boxes overlap substantially. There is not much evidence of a selection mechanism causing better students to end up in one site relative to the others.

Figure B-10. Sight Words (Post)

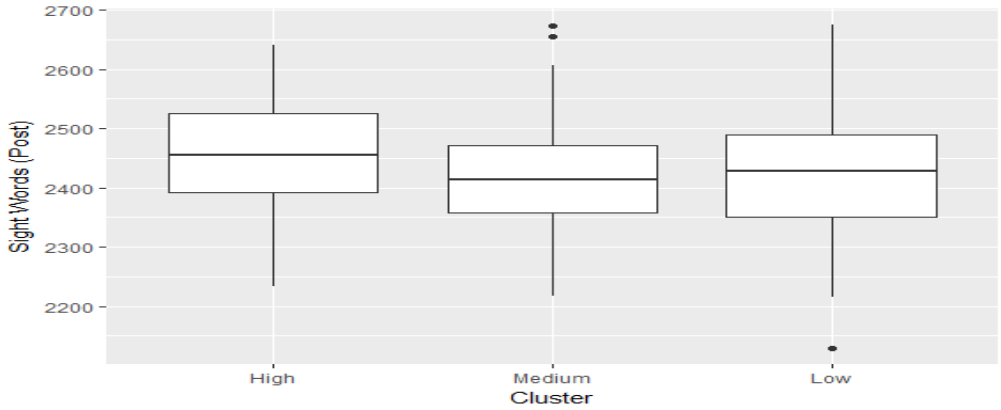
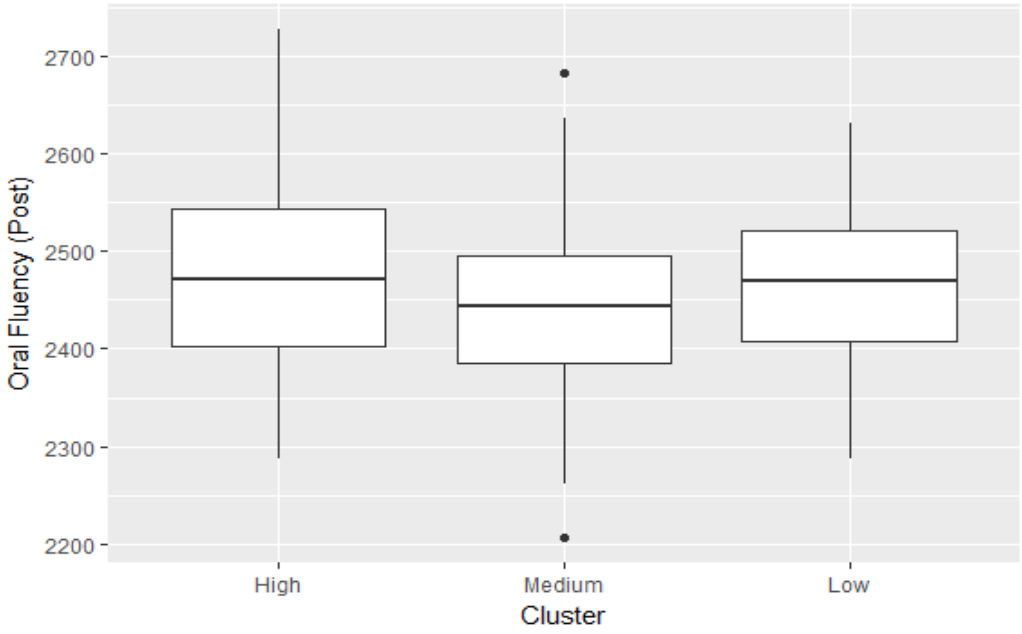


Figure B-11. Oral Fluency (Post)



Weighting the data to maximize comparability between sites does very little to change these distributions, as shown in the following figures.

Figure B-12. Sight Words (Post)

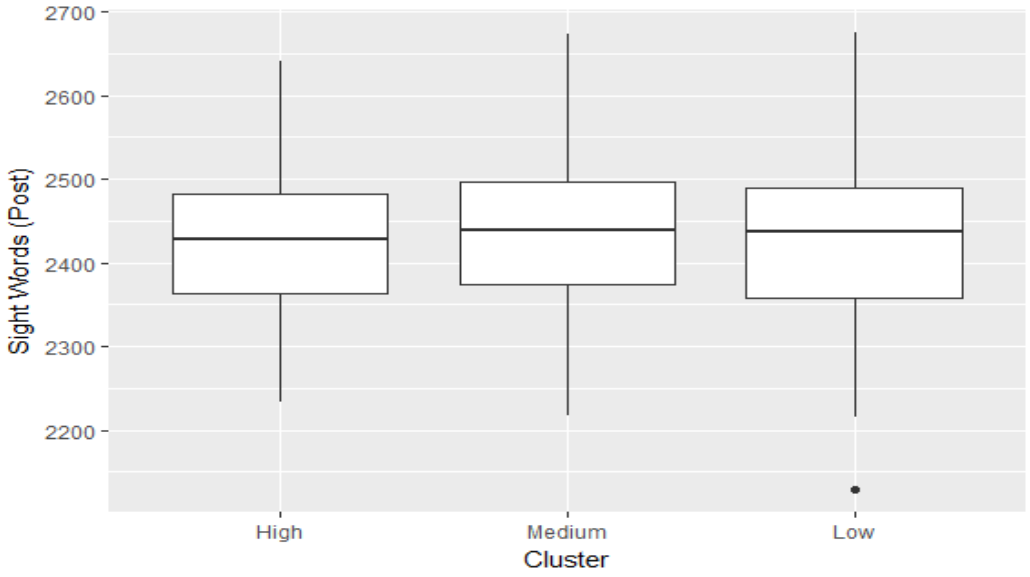
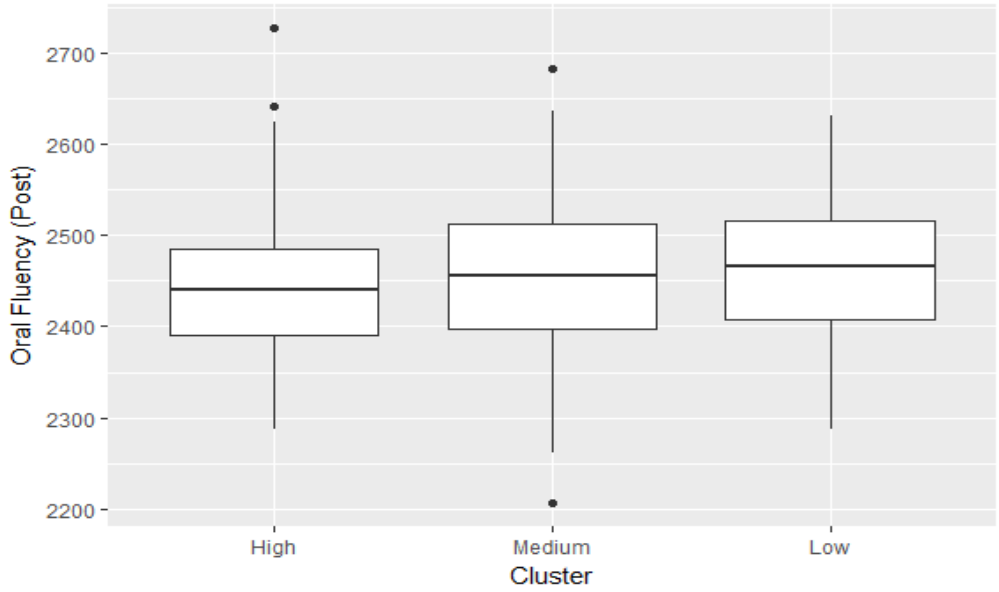


Figure B-13. Oral Fluency (Post)



The following table shows that site quality does not statistically distinguish post-state assessment scores in a propensity score-weighted mixed model. The mean reading score was 5.2 higher in the medium quality group relative to the high quality group, and the low site quality group had scores that were 3.09 lower on average relative to the high quality case. For math, the average score was 3.28 higher

in the medium group relative to the high quality group, and scores were 10.21 higher for the low quality group relative to the high quality group.

Table B-6. State Assessments - 3rd and 4th Graders

	Dependent variable:	
	Reading	Math
	(1)	(2)
Quality = Medium	5.201	3.276
	(16.269)	(15.545)
Quality = Low	−3.086	10.207
	(17.778)	(17.262)
Constant	2,428.728***	2,447.141***
	(12.46)	(12.093)
Observations	538	539
Log Likelihood	−3,264.962	−3,197.727
Akaike Inf. Crit.	6,359.923	6,405.455
Bayesian Inf. Crit.	6,561.363	6,426,904

Note: *p<0.05; **p<0.01; ***p<0.001

The following table presents results that adjust for covariates using a traditional mixed model without relying on propensity scores. The typical reading score in the medium group was 2.95 lower relative to the high group, and the low group was 6.01 points lower than the high quality group. For math, medium site scores were 1.28 lower on average compared to high quality groups, and low quality sites had scores that were 6.55 higher on average relative to high quality groups.

Table B-7. State Assessments-3rd and 4th Graders

	Reading	Math
	(1)	(2)
Quality = Medium	-2.947	-1.284
	(7.336)	(5.909)
Quality = Low	-6.009	6.548
	(8.434)	(6.768)
Grade Level	-3.209	-11.550*
	(5.268)	(4.530)
Female	16.463**	-7.845
	(5.171)	(4.443)

African American	-20.076*	-15.530
	(9.614)	(8.284)
Hispanic	-16.829	-13.965
	(9.719)	(8.421)
Asian	-10.695	3.573
	(9.996)	(8.605)
Other	-7.211	-22.973*
	(11.402)	(9.913)
ESL	-3.699	-6.287
	(5.964)	(4.884)
Special needs	-0.608	-4.596
	(6.419)	(5.503)
Sight Words – Pre	-0.032	0.022
	(0.071)	(0.090)
Oral Fluency – Pre	0.261***	-0.025
	(0.071)	(0.059)
Math -Pre	0.495	3.019**
	(1.133)	(0.964)
Math Assessment 2015-2016	0.313***	0.611***
	(0.054)	(0.045)
Reading Assessment 2015-2016	0.516***	0.143***
	(0.050)	(0.043)
Constant	432.192***	996.212***
	(116.443)	(98.762)
Observations	470	471
Log Likelihood	−2,503.051	−2,439.778
Akaike Inf. Crit.	5,042.102	4,917.557
Bayesian Inf. Crit.	5,116.852	4,996.499

Note: *p<0.05; **p<0.01; ***p<0.001

Moderation Analysis - 3rd and 4th Graders

The final analysis considers whether the association between pre-treatment and post-treatment scores is affected by site quality for low achieving 3rd and 4th graders. There were two math related outcomes, the post-program evaluation and the 2016-2017 state math assessment. The test scores have been rescaled to be z-scores to facilitate interpretation of the interactions. The following figures display the relationship between the pre and post scores for both outcomes separated by site quality. The lines represent the least squares fit to the respective site quality type weighted using propensity scores, and the shaded areas represent 95% confidence intervals. Evidence of moderation - site quality mattering more for low baseline students compared to higher baseline students - would appear if the slopes were dramatically different from each other.

Figure B-14. State Math Assessment 2016-2017 (Z-score)

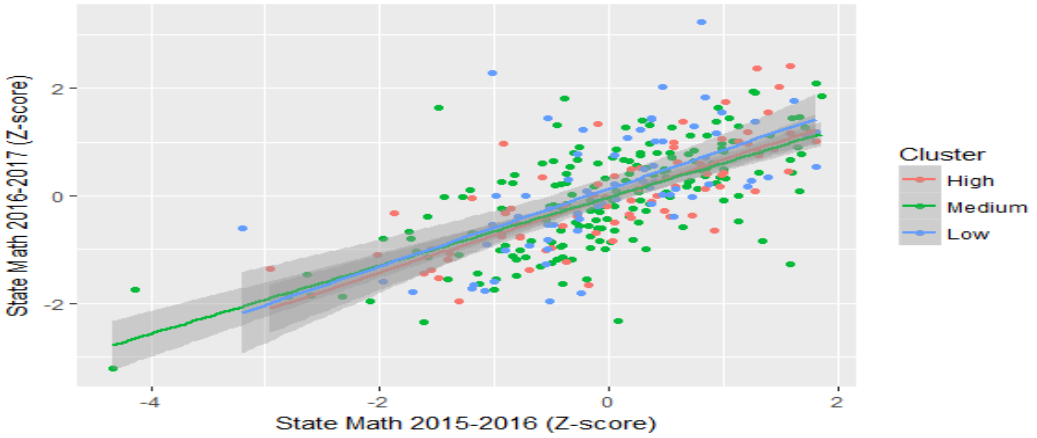
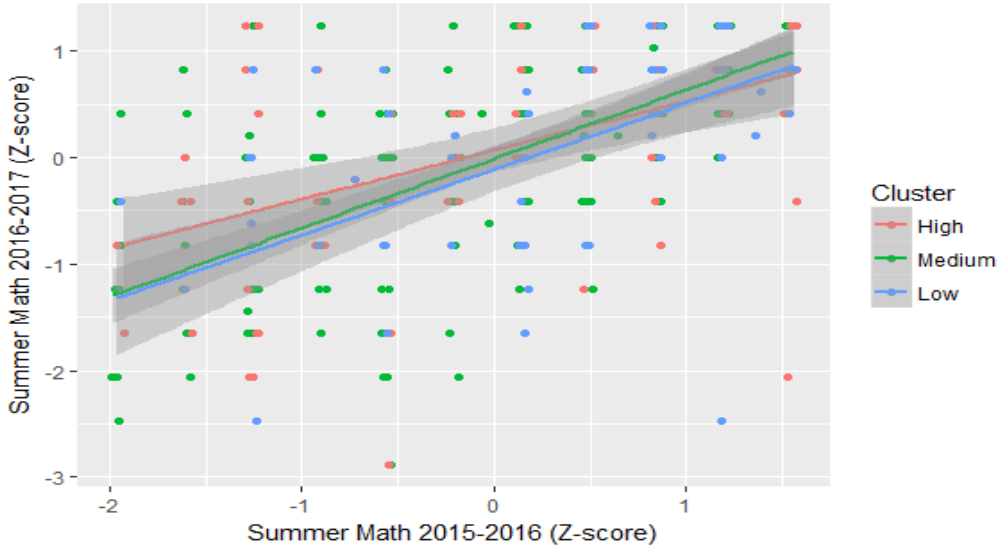


Figure B-15. Summer Math Assessment 2016-2017 (Z-score)



The figures do not show much difference in slopes for the state assessments, and, as the following table shows, the interactions do not reach statistical significance. The "main effects" in the model are the slopes for each variable when the other variable in the interaction equals zero. For example, the medium-quality groups have state assessment scores that are .04 higher than the high-quality group *when pre-treatment scores are at their means only*, and scores for low-quality groups are .18 higher on average than high-quality groups *when pre-treatment scores are at their mean only*. An increase of one in pre-treatment scores on state math assessments leads to an increase of .684 in post-treatment scores *for high-quality groups only*. The interactions show that the effect of pre-treatment scores is smaller for the medium-and-low-quality groups compared to the high-quality group. If the effect of pre-treatment math assessment scores is an increase of .684 for the high-quality group, it is an increase of .684 - .018 = 0.666 for the medium-quality group and an increase of .684 - .065 = 0.619 for the low-quality group. The significance on the interaction terms tests if .684 is significantly different from 0.666 (medium) and if .684 is significantly different from 0.619 (low). Neither yields a p-value less than .05.

Table B-8. Math Outcomes - Low Proficiency Only

	Dependent variable:	
	State Math	Summer Math
	(1)	(2)
Quality = Medium	0.043	-0.066
	(0.115)	(0.157)
Quality = Low	0.179	-0.094
	(0.127)	(0.185)
Pre Score	0.684***	0.525***
	(0.070)	(0.07)
Pre X Quality = Medium	-0.018	0.116
	(0.094)	(0.108)
Pre X Quality = Low	-0.065	0.109
	(0.099)	(0.122)
Constant	-0.068	0.039
	(0.089)	(0.126)
Observations	326	310
Log Likelihood	-370.337	-393.903
Akaike Inf. Crit.	756.675	803.806
Bayesian Inf. Crit.	786.970	833.699

Note: *p<0.05; **p<0.01; ***p<0.001

Looking at summer math scores, there are no significant differences between medium-and-high-quality group (-.066) nor low-and-high-quality group (-.094) *when pre-treatment state assessments are at their mean*. The effect of increasing pre-treatment math scores by one is to increase post-treatment math scores by .525 *for the high group only*. The interactions show that an increase of one on the pre-treatment test leads to an increase of .525 + .116 = 0.641 in post-treatment scores for the medium-quality group and an increase of .525 + .109 = 0.634 in the low-quality group. The difference between .525 and 0.641 is not significant, nor is the difference between .525 and 0.634.

A similar analysis was performed for the state reading assessments and the summer Sight Words and Oral Fluency tests. The following figures show scatterplots and best weighted linear fits by cluster. The slopes are very similar for the reading scores, which do not indicate an interaction. The Sight Words score shows similar slopes for high- and-medium-quality sites but a noticeably flatter slope for the low-quality sites. Finally, the Oral Fluency lines generally overlap, but the medium-quality group demonstrates a tendency to be steeper.

Figure B-16. State Reading 2016-2017 (Z-score)

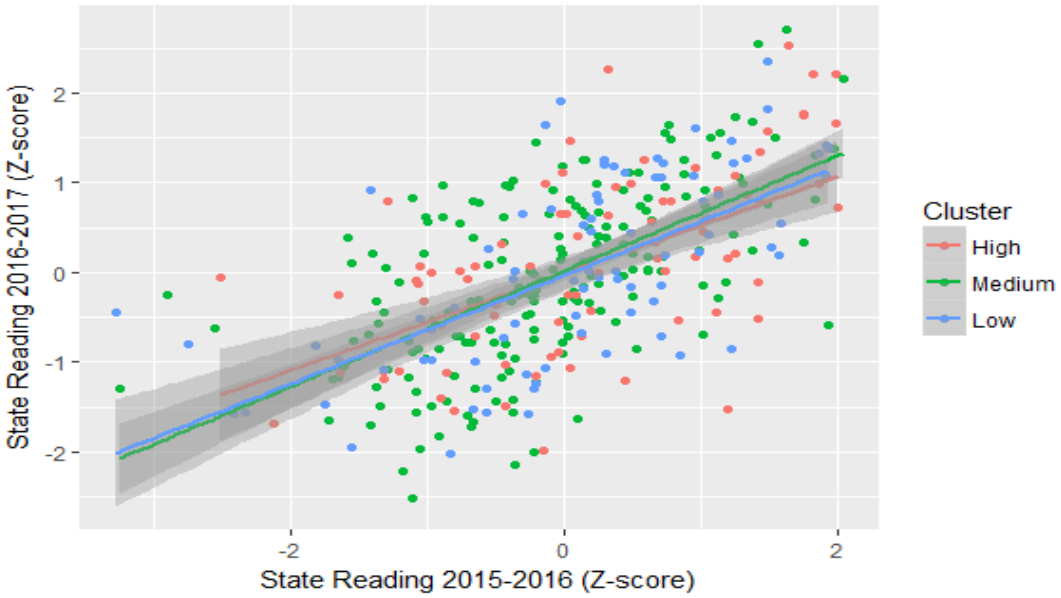
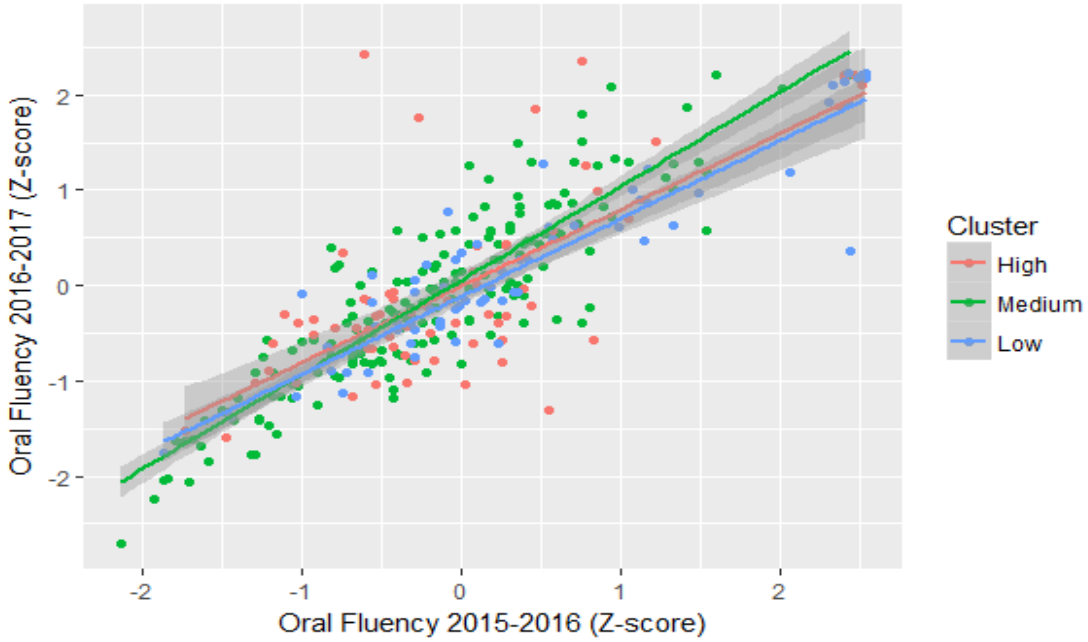


Figure B-17. Sight Words Post (Z-score)

Figure B-18. Oral Fluency 2016-2017 (Z-score)



The final table shows results from propensity score-weighted mixed models of reading outcomes. The main effects for site quality on state reading assessments are not significant, while the main effect for pre-treatment assessments is significant. This latter term predicts an increase of .55 in post-treatment scores for each unit increase in pre-treatment scores *for high quality sites only*. The interactions show how much the slope for pre-treatment scores changes for medium and low quality groups relative to the high quality group. A unit increase in pre-program state reading scores leads to an increase of .547 + .083 = 0.63 for medium sites and .547 + .023 = 0.57 for low sites, but these are not significantly different from the .547 slope in the high group.

The second model considers summer sight words scores. There is no significant difference in post-treatment scores between medium and high quality sites, nor between low and high quality sites, when pre-treatment sight words scores are at their mean. The effect of increasing pre-treatment scores by one is to increase post-treatment scores by .443 in the high quality sites. The interactions here are significant, though in a manner that is a little different from what was implied by the previous figure. The model shows that the pre-treatment scores slope is significantly *higher* for the medium group versus the high group as well as for the low group versus the high group. The interpretation of the model in the table is that the effect of a unit change in pre-treatment scores for medium groups is to increase the post-treatment scores by 443 + .228 = 0.671, which is statistically significant from the .443 for the high quality sites. The effect of a unit increase in pre-treatment scores for the low quality group is to increase post-treatment scores by .443 + .171 = 0.614, which is significantly higher than the .443 for the high quality sites.

The explanation for why the model differs from what the figure implied is the inclusion of the control for site in the form of a random effect. The results from a model without the random effect (i.e. as just a regression model) found no significant effect for the medium site X pre-treatment scores interaction but a significant *negative* interaction term, consistent with the figure, for the low site X pre-treatment scores interaction.

Finally, there are no significant differences in oral fluency scores between medium and high quality sites, nor between low and high quality sites, when pre-treatment scores are at their mean. The effect of a unit increase in pre-treatment scores is to increase post-treatment scores by .758 in the high quality group, a statistically significant amount. The effect of a unit increase in pre-treatment scores for the medium group is to increase post-treatment scores by .758 + .213 = 0.971, which is significantly more than the .758 in the high group. The effect of a unit increase in pre-treatment scores for the low quality sites is to increase post-treatment scores by .758 + .060 = 0.818, which is not significantly different from the .758 in the high quality group.

Table B-9. Reading Outcomes- Low Proficiency Only

	State Reading	Summer Sight Words	Summer Oral Fluency
Quality = Medium	0.044	0.001	0.105
	(0.139)	(0.277)	(0.131)
Quality = Low	0.015	0.098	-0.077
	(0.149)	(0.322)	(0.154)
Pre Score	0.547***	0.443***	0.758***
	(0.076)	(0.051)	(0.064)
Pre X Quality = Medium	0.083	0.228***	0.213*
	(0.111)	(0.068)	(0.084)
Pre X Quality = Low	0.023	0.171*	0.060
	(0.103)	(0.080)	(0.091)
Constant	-0.024	-0.065	-0.039
	(0.106)	(0.228)	(0.108)
Observations	357	347	330
Log Likelihood	-463.345	-280.572	-264.910
Akaike Inf. Crit.	942.689	577.144	545.820
Bayesian Inf. Crit.	973.711	607.939	576.213

Note: *p<0.05; **p<0.01; ***p<0.001

Appendix C. Plan for Quasi-Experimental Test of SPS Summer Program Participation with a No-Program Control Group

(The following is a memo sent to John Hughes, from Charles Smith, on June, 09, 2016)

Here are some thoughts on adding a sample of youth for a no-program comparison. Per the additional thinking below, what you want to request from the Evaluation Department is: Two consecutive years of data for the 2nd and 3rd graders who attended in Summer 16 (n=584) and a matched comparison based on 2015-16 data. Additional reasoning to get us to that request is below.

1. Assumptions:

- Summer 16 (last year) and Summer 17 (current year) will be the two highest impact (highest fidelity) years of Summer Staircase
- Summer Staircase program lead wants to evaluate:
    - Program impact on school year achievement for attenders of Summer Staircase
    - Program impact on school year achievement for multi-year attenders of Summer Staircase
- WC is currently addressing the impact question by comparing Summer Staircase students who were in high fidelity cohorts (High PQA score) to students who were in lower fidelity (low PQA score) cohorts. Summer Staircase program lead wants to add a "no-program" group (or arm) to the study.

2. Data request strategy:

- Summer staircase program lead wants to make a request to SPS evaluation department that would allow addition of the no-program arm of the study
- Table 1 describes the current evaluation data that has been assembled on Summer Staircase as part of the WC evaluation to date
- The simplest (and cheapest) way to add a no-program arm is to identify a comparison group of students:
    - Who did not attend Summer Staircase in 2015, 2016, or 2017
    - Who can be matched (on achievement and background data) to the 1,094 Summer 16 sample using the 2015-16 school year data
- This allows us to conduct the following program vs. no program comparisons:
    - Summer Staircase 2016 sample vs matched comparison on School year 16-17 achievement (which we will receive in September of 2017)

  - Summer Staircase 2016 sample vs matched comparison on School year 17-18 achievement (regardless of attendance in Summer Staircase 2017)- this would be conducted in September of 2018
  - Summer Staircase two year attenders (summer 16 and summer 17) vs matched comparison on 17-18 achievement- this would be conducted in September of 2018
- Given testing procedures in SPS:
  - By selecting only 2nd and 3rd grade students in the summer 16 sample (N=584), we would have the following comparisons:
    - 16-17 school year achievement for 3rd and 4th graders
    - 17-18 school year achievement for 4th and 5th graders
- **Summary** – Create matched no-program sample for summer 16 sample of 2nd and 3rd graders (N=584 out of 10,940) using 2015-16 school year data as the baseline for the match, then compare achievement of those groups of students in the subsequent 16-17 and 17-18 school years

3. Other questions to ask SPS Evaluation Department:

- Can SPS provide us with an identified comparison group of students– can they do propensity score or other matching method to identify a comparison group?
- If no to above, can they supply us with 15-16 school year data for all of the 2nd and 3rd grade students in the district in that year so we can create the comparison group

Table C-1. Summer Staircase Sample and Data

	Summer 15	School Year 15-16	Summer 16	School Year 16-17	Summer 17	School Year 17-18
Sample for Summer Staircase	345 of summer 16 students		10,940 students		Est. 500 of summer 16 students	
Outcome data – pre/post in summer; school year achievement	2 outcome variables in math and lit		5 outcome variables in math and lit	Achievement data		Achievement data
Mod/med data – Teacher quality in summer; student background	Teacher quality for 28 cohorts at 9 sites	Student background	Teacher quality for 60 cohorts at 19 sites			
Comparison Group		*Identify comparison group matched to summer 16 program group using 2015-16 school year data*				