ALFA Lab Study Final Report

August 2023

Prepared by:
Jason Schoeneberger
Xiaodong Zhang
Samantha Spinney
Jing Sun
Lauren Kennedy
Samira Rajesh Syal

ICF
1902 Reston Metro Plaza
Reston, VA 20190
www.icf.com

Acknowledgement

This publication was supported by a subagreement from Johns Hopkins University with funds provided by Grant #R305A180154 through the U.S. Department of Education. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Education or Johns Hopkins University.

Many individuals and organizations have contributed to this evaluation study. First, we would like to acknowledge our colleagues from Johns Hopkins University, Principal Investigator (PI) Douglas Mac Iver, and Co-PIs Marcia Davis and Martha Mac Iver for their leadership in the grant and for serving as true intellectual partners who recruited the schools and implemented the ALFA Lab curriculum.

A special thank you to two external research advisers—Christopher Rhoads of the University of Connecticut and Herb Turner from Analytica—for their guidance on the design and implementation of the regression discontinuity design. Essential support was provided by other ICF staff, including Matt McKinney, Kathleen Calcerano, and Danielle Hayden. We also want to acknowledge the University of Kansas and Renaissance Learning for providing support in administering the Adaptive Reading Motivation Measure (ARMM) and STAR Reading™ assessments.

The research team has been inspired by the courage and dedication of the educators and students who participated in the program and data collection by overcoming the unprecedented challenges of the corona virus 2019 (COVID-19) pandemic.

Finally, we would like to thank the Institute of Education Sciences at the U.S. Department of Education for funding this important research.

Table of Contents

Executive Summary

Funded by a research grant from the U.S. Department of Education's Institute of Education Sciences, Johns Hopkins University (JHU) and ICF are conducting a national study of an elective lab course, Accelerating Literacy for Adolescents (ALFA) Lab, which seeks to improve students' reading achievement, particularly for those from economically disadvantaged communities. ALFA Lab is offered as a one-semester extra-help course that the most-challenged readers take in addition to their regular English course during the first or second semester of Grade 9. Making use of small teams rotating through stations, the intent is to enhance student comprehension, motivation, vocabulary development, and fluency. ICF is conducting an external evaluation of ALFA Lab to study (1) the impact of the lab on reading achievement and motivation, (2) how the lab is being implemented, and (3) the cost of implementing the lab. This report provides findings based on data collected from four schools in three cohorts during the 2019–2022 school years.

The study employs a regression discontinuity (RD) design. Students whose pre-ALFA Lab reading scores were below a cutoff point were assigned to ALFA Lab and those with scores above the cutoff were assigned to the business-as-usual (BAU) instruction group. We used a comprehensive online literacy assessment, STAR Reading, to assess reading achievement. To assess student reading motivation, we employed the Adaptive Reading Motivation Measure (ARMM) to measure multiple constructs (e.g., self-efficacy and intrinsic, extrinsic, and social motivation for reading), yielding overall General Reading Motivation (GRM) and Reading Frequency (FRQ) scores. We also examined implementation fidelity and cost. Prominent findings are summarized below.

- Student demographics vary across the four schools. Two schools were diverse while two were predominantly White or Hispanic.
- The four participating high schools made considerable progress implementing the ALFA Lab program. Teachers were implementing ALFA Lab components—such as Daily Launch, Main Station, Wordology, Collaboration Station, and Media Madness—at every class meeting or nearly every class meeting, Additionally, over two-thirds of survey respondents indicated they used the ALFA Lab student worksheets, teacher handbook, and teacher manual as well as reading selections for the Feisty Felines, Heroes, and Galaxy units. All observed ALFA Lab classes demonstrated evidence of implementing most ALFA Lab components included in the Checklist for ALFA Implementation (CFAI).
- Combining data across sites, we found that ALFA students outperformed the BAU group in reading achievement (STAR assessment), reading motivation, and reading frequency (ARMM), although the differences were not statistically significant.
- Mediation analyses exploring whether reading motivation and reading frequency measures explain estimated ALFA Lab impacts on STAR outcome scores found no statistically significant effects, though each explained 30% and 22% of the total effect, respectively. Analyses exploring whether student factors moderated the estimated ALFA Lab impact estimates found no statistically significant effects for STAR or reading motivation outcomes, but did find significant, positive effects for males, non-English Language Learner (ELL) students, non-special education (SPED) students, and Hispanic students.

- The average cost for a one-semester ALFA Lab program is $29,731, with a cost per student of $1,351. The total costs of operating an ALFA Lab program varied largely across the four participating schools, ranging from $22,684 to $90,776. The primary factors driving the difference were the number of sections and number of participating students in each school. The more program sections a school has—with more participating students—the higher the total cost, with a similar per-student cost around $1,200.
- It is important to acknowledge the tremendous challenges the corona virus 2019 (COVID-19) pandemic posed to ALFA Lab implementation and data collection. When the President declared a national emergency on March 13, 2020 because of the pandemic, Cohort 1 schools abruptly ended the academic year on that very day. Posttest outcome assessment was thus delayed for Cohort 1 until participants were sophomores. For Cohort 2, schools implemented a remote-only, synchronous learning schedule or a hybrid model in which general education students alternated between in-person and virtual classes except during COVID-19 outbreaks when instruction became fully remote for a time. For Cohort 3, even with the return to in-person instruction, teachers and principals described challenges with student engagement and participation in academic coursework.

In the course of the study, we have prepared a series of reports including annual implementation reports, an interim outcome report, cost study analysis, and school-level reports to provide ongoing feedback to the intervention team and participating schools. The study team also presented at conferences organized by the American Educational Research Association, the American Evaluation Association, and the Society for Research on Educational Effectiveness.

Introduction

Funded by a research grant from the U.S. Department of Education's Institute of Education Sciences (IES) running from 2018–2023, Johns Hopkins University (JHU) and ICF are conducting a national study of an elective lab course, Accelerating Literacy for Adolescents (ALFA) Lab, which seeks to improve students' reading achievement, particularly for those from economically disadvantaged communities. The lab was originally developed at JHU as one component of the Comprehensive School Reform Design dubbed Talent Development Secondary (TDS). Although TDS has since spun off from JHU as an independent organization, the ALFA Lab course continues to be updated and disseminated by the Center for Social Organization of Schools at JHU. It is offered as a one-semester extra-help course that the most-challenged readers take in addition to their regular English course during the first or second semester of Grade 9. The lab was designed to enhance students' comprehension strategies, motivation, vocabulary development, and fluency. Each lab begins with a Daily Launch and then, during the bulk of the lab session, students rotate in small teams to four different stations (i.e., Main Station, Collaboration Station, Wordology, and Media Madness) where they work both individually and collaboratively to complete activities requiring them to apply comprehension skills, word knowledge, and writing skills.

ICF is conducting an external evaluation of ALFA Lab to study (1) the impact of the lab on reading achievement and reading motivation, (2) how the lab is being implemented, and (3) the cost of implementing the lab. This report provides findings on student reading achievement and motivation based on data collected from three cohorts of students in 2019–2020, 2020–2021, and 2021–2022 school years. School 1 (New York), and School 3 (California) were involved in all three cohorts while School 2 (Connecticut) and School 4 (Alabama) participated in Cohort 3 only.[1]

It is important to acknowledge how the corona virus 2019 (COVID-19) pandemic affected the ALFA Lab implementation and data collection. Because of the President's declaration of a national emergency, schools in Cohort 1 closed months early (in March 2020) and posttreatment outcome data could not be collected until participating freshmen had become sophomores. In Cohort 2, participating freshmen (both ALFA and non-ALFA) either received full-time remote instruction or experienced a hybrid instructional model throughout the year: School 1 (NY) adopted a hybrid model in which all ninth-grade general education students attended class in person just 2 days per week (Mondays and Thursdays) and remotely via videoconference 3 days per week. School 3 (CA) implemented a remote-only, synchronous learning schedule for most students throughout the 2020–2021 school year. A limited number of students were able to attend their virtual classes from the school building (e.g., if they had no internet access at home). In Cohort 3 (the 2021–2022 school year), three of the four schools (School 2 [CT], School 3 [CA], and School 4 [AL]) returned to in-person instruction immediately and School 1 (NY) resumed in-person instruction a couple of months after the school year began. Even with the return to in-person instruction, focus group participants described challenges with student engagement and

[1] A charter high school located in Louisiana participated in cohort 1 of ALFA Lab. Due to low enrollment and the lack of outcome data, the school was excluded from the analysis.

participation in academic coursework. Additionally, participants noted students may have been absent for an extended period due to COVID-19 quarantine policies, none of which was unique to ALFA Lab (see Exhibits B2.1 and B2.2 in Appendix B for attrition rates calculated by including students with missing outcomes and excluding students with missing outcomes, respectively). All of these are major potential threats to "achieved relative strength" or the intended treatment versus comparison contrast on the ALFA Lab key components, which in turn can reduce the differences between the groups on the outcome.

Theory of Change

Figure 1 outlines the theory of change (TOC) associated with ALFA Lab and its intent to impact student achievement. Each of the key pieces can be defined as follows:

- The navy-blue box depicts the basic inputs needed to implement ALFA Lab. (This is the sole box under Inputs.)
- The orange boxes depict key components of the ALFA Lab program. (The three boxes at the bottom of the Key Components column.)
- The glowing light-blue boxes depict factors believed to have an impact on the key components. (This includes the box under the Mediators column and the box under the Moderators column.) Specifically, the light-blue box radiating gray depicts mediators that are believed to have a potential impact on the key components directly (e.g., student-teacher ratios will impact instruction during lab time). The light-blue box radiating green depicts moderators that have the potential to impact how the key components may (or may not) influence direct impacts on student reading achievement.
- The purple boxes depict intermediate outcomes that have the potential to function like mediators, at least in the sense that they have the potential to indirectly impact student reading achievement based on how the key components influence them. In effect, these intermediate outcomes are anticipated to function like mediators that influence student reading achievement through indirect paths from key components. (The three boxes under the Short/Long-Term Outcome column.)
- The navy-blue bracket across the bottom reflects the influence that external events may have had on the ALFA Lab implementation and/or impact estimates (e.g., COVID-19 pandemic, economic downturn, differentially low socioeconomic status in participating schools).

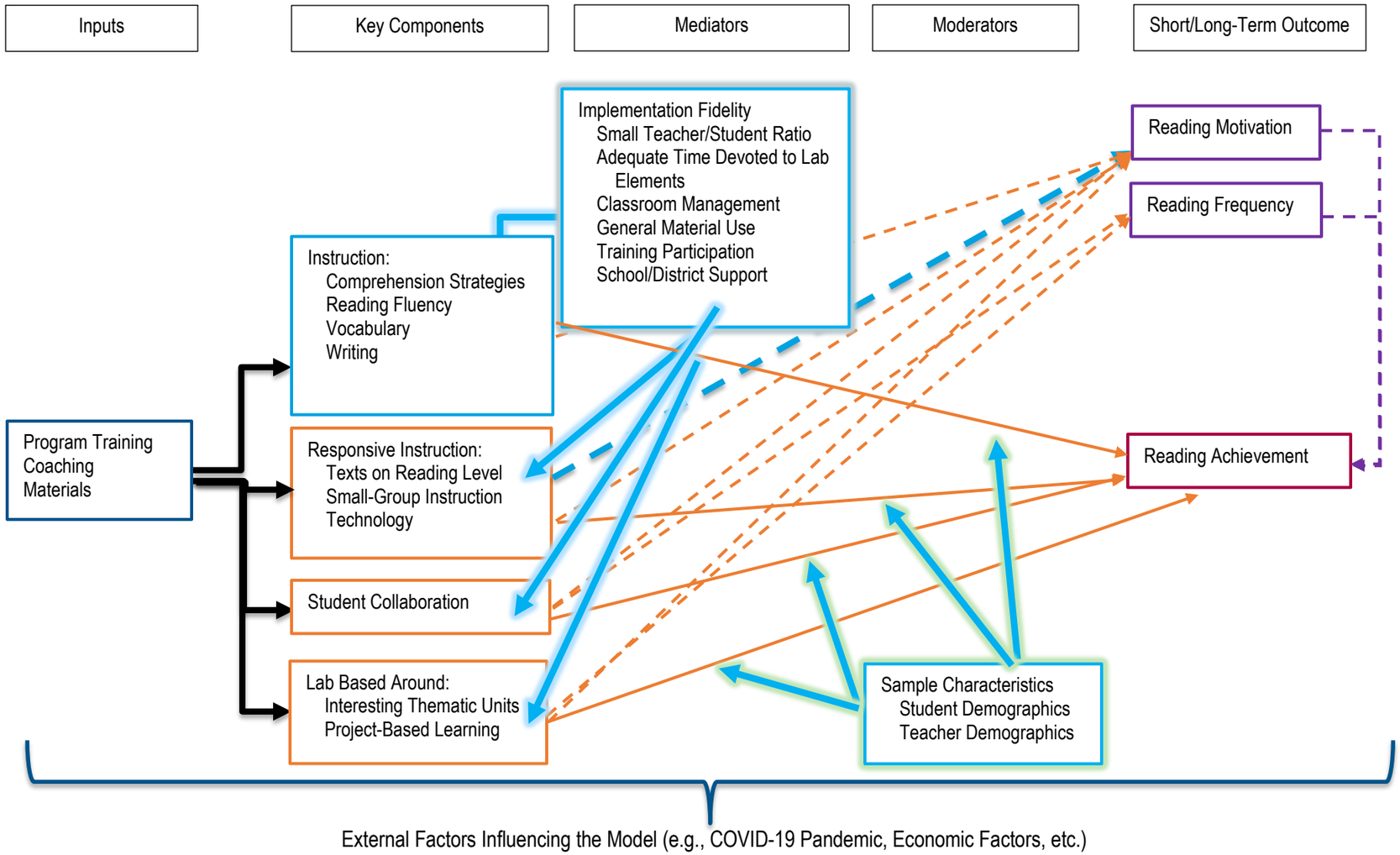


Figure 1. ALFA Lab Theory of Change

Research Questions

Below are the research questions associated with the ALFA Lab evaluation, aligned with the theory of change presented in Figure 1. This report will provide findings regarding research questions 1 through 6 (RQ1–RQ6) based on combined data from three cohorts. RQ1–RQ5 were written in our research grant whereas RQ6 is an additional question the research team explored using the available data.

Confirmatory Question

RQ1: Does the offer/participation in ALFA, compared to business-as-usual (BAU) experiences, improve ninth graders' reading achievement, as measured by the Renaissance STAR assessment?

Exploratory Questions

RQ2: Does offer/participation in ALFA, compared to BAU experiences, improve students' general motivation for reading?

RQ3: Does offer/participation in ALFA, compared to BAU experiences, improve students' reading frequency?

RQ4: How well is the ALFA Lab program implemented with fidelity? What are the factors that facilitate or hinder implementation?

RQ5: What is the cost associated with implementing the ALFA Lab program, and what proportion of the costs are borne by the school?

RQ6: What is the relationship between adolescents with diverse motivational profiles and their reading outcomes?

Methods

Impact Questions (RQ1–RQ3)

This section begins by introducing the regression discontinuity design and describing the sample characteristics. Then we discuss the outcome measures for student reading achievement, motivation, and frequency. Finally, we detail the analytic methods.

Regression Discontinuity Design

To estimate program effects, one must reliably establish an unbiased estimate of the counterfactuals; the outcomes that would have been observed in the absence of the intervention. Regression discontinuity (RD) designs provide a method for obtaining unbiased estimates when random assignment to treatment is not possible by establishing a cutoff point along an assignment variable's continuum to

determine receipt of treatment or membership in a BAU instruction group. In this study, we implemented an RD design by assigning students whose pre-ALFA Lab reading scores were below a cutoff point to ALFA Lab and those with scores above the cutoff to the comparison group. Since ALFA Lab is designed to serve all of a school's most-challenged readers during a single semester of Grade 9,[2] it is not feasible to employ a randomized controlled trial as some of the students most in need would likely be assigned to the BAU group. To the extent possible, the study team adhered to the standards for high-quality RD studies, articulated by the *What Works Clearinghouse (WWC)™ Procedures and Standards Handbook, Version 5.0 (WWC Standards 5.0).*

RD designs generate consistent estimates of the program effect if: (1) the relationship between the outcome and assignment variable is modeled appropriately, (2) the assignment variable is not manipulated to influence assignment to the treatment group, and (3) there is a continuous functional relationship between the outcome and assignment variables. The research team worked with schools to establish eligibility criteria for each cohort's Grade 9 participants based on their prior grade data from locally administered assessments and course grades. Given ALFA Lab's cooperative, student-centered, multiple learning stations approach, most students who displayed severe levels of disruptive behavior in the previous grade were excluded from the study and instead received case-managed behavior supports (and customized literacy supports if needed). Additionally, most students in special education (SPED) whose individualized education program (IEP) specifically excludes them from the mainstream English language arts (ELA) classes were excluded so that they could receive more intensive supports.

The schools included in the study were of varying sizes and student disadvantage. For each cohort and school, Table 1 shows the assessment and cut-score used to determine assignment to the ALFA Lab intervention as well as the number of eligible students that had at least one available outcome score and no missing covariate data (see Appendix B, p. B-4 for attrition details).[3] Note that for School 1 (NY), School 2 (CT), and School 4 (AL), all students assigned to ALFA Lab participated during the first semester of the year. In School 1 (NY), ALFA Lab replaced the first-semester Strategic Reading course for those identified for reading assistance and all students received the full, required English 9 course during the second semester in 90-minute periods. In Schools 2 (CT) and 4 (AL), ALFA Lab functions as an additional course (i.e., does not replace the core English course). In School 3 (CA), ALFA Lab participation replaced one semester of a full year geography course. All students in School 3 (CA) take the core English course for the entire Grade 9 year. School 3 (CA), however, had some students take ALFA Lab during the first semester and some during the second semester because they had more students deemed eligible for assignment than could be served in the first semester's ALFA Lab sections (both sets were included in all

[2] The ideal implementation provides for a school's most-challenged readers to participate in ALFA Lab during the first semester of Grade 9, albeit potentially across multiple sections. However, some participating schools had more students in need than could be served in a single semester. In those instances, students were served in the second semester.

[3] Reasons students at any of the sites were not eligible for assignment include transferring out of the participating school, status as an English Language Learner (ELL) or placement in SPED, participation in an Advancement Via Individual Determination (AVID) program, scheduling conflicts, or a lack of a valid assignment variable measure.

analyses). School 2 (CT) and School 3 (CA) also had some level of non-compliance, where students did not end up in the appropriate group based on their assignment variable, resulting in a Fuzzy RD design.[4]

[4] At School 3 (CA), cohort 2, six students should have been assigned to ALFA Lab but became comparison students. At School 1 (NY), cohort 2, seven students should have been assigned as comparison students but were assigned to ALFA Lab. School 3 (CA) in cohort 3 had two students who should have been assigned as comparison students but were assigned to ALFA Lab, while School 2 (CT) had nine students that should have been assigned to ALFA Lab but became comparison students. These changes in assignment were due to school personnel responding to external information such as superior course grades despite a low prior standardized test score (CA), scheduling conflicts (CT), or teacher shortages forcing classroom collapses (NY). A Fuzzy RD design accounts for the non-compliance when generating impact estimates.

Table 1. Assessment and Cut-Score Information

Cohort	School	Assessment		Sample		
		Name	Cut-Score	n_{Total}	n_{ALFA}	n_{BAU}
1	CA	Degrees of Reading Power	7.5	117	78	39
	NY	Grade 7 ELA Scale Score	581	72	9	63
	Total			***189***	***87***	***102***
2	CA	CAASPP*/Grades Composite	-0.167	213	82	131
	NY	Grade 7 ELA Scale Score	588	93	30	63
	Total			***306***	***112***	***194***
3	AL	Edmentum Scores (Grade 8)	1160	67	22	45
	CA	Block 1 (Grade 8 Fall STAR)	481	157	44	113
		Block 2 (Grade 8 Spring STAR)	588	21	15	6
	CT	STAR Reading Score (Grade 9 at entry)	6.4	90	15	75
	NY	iReady Reading Assessment	600	98	40	58
	Total			***433***	***136***	***297***

*California Assessment of Student Performance and Progress (CAASPP)

Estimation of the intent-to-treat (ITT) RD-based impact of ALFA Lab proceeds using a continuity-based, nonparametric regression approach that obtains an impact estimate as the difference of two separate one-sided, nonparametric (local polynomial) boundary regression functions at the cut-score given a chosen polynomial (e.g., p = 1 for linear), kernel-weighting function (e.g., triangular) and data-driven selected bandwidth *h*. Effectively, data from each side of the cutoff are used to model the assignment variable-outcome relationship, and can be done so within separate, unique bandwidths on each side of the cutoff (i.e., allowing for potentially different slopes). Using the same weights and data-driven bandwidth, the RD effect can also be estimated using the equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 Trt + \beta_3 XTrt + e$$

where Y is the outcome variable, β_0 is the intercept, X is the assignment variable, Trt is a variable denoting treatment or comparison group membership and β_2 is the coefficient associated with the treatment group membership and β_3 is the coefficient associated with the interaction between the assignment variable X and the treatment indicator Trt. When the assignment variable is centered at the cut-point (as it is in our analyses), β_2 can be interpreted as the causal effect of treatment assignment at the cut-point. The Fuzzy design specification incorporates a single binary variable denoting membership in the ALFA Lab (or BAU) groups.

Analytic Process

Smoothing/Binning Data

Both quantile-spaced (QS) bins, which contain the same number of observations in each bin, and evenly-spaced (ES) bins, which partition the assignment score distribution into bins each of the same distance along the assignment score continuum, are used to create binned scatter plots to graphically explore the assignment-outcome relationship. Additionally, two distinct methods for determining the number of bins are utilized. The integrated mean-squared error (IMSE) method minimizes (an asymptotic approximation of) the mean-squared error that results from approximating the individual data points with a local constant. The second bin selection method is the mimicking variance (MV) method, in which the number of bins is chosen so the overall variability of the binned means is close to the variability of the raw data. Plots based on IMSE bins can then be used to identify a well-fitting form of a regression function (Cattaneo, Idrobo & Titiunik, 2020).

Point Estimation

In this study, we explored a number of different RD model specifications as a form of sensitivity analysis, prioritizing the following specifications as our primary model:

- A local linear specification
- A triangular kernel that more heavily weights cases closer to the cutoff (Cattaneo, Idrobo & Titiunik, 2020, p. 42)
- The identification of separate bandwidths on each side of the cut-score designed to minimize the mean-squared error (MSE) of the point estimate (Cattaneo, Idrobo & Titiunik, 2020)
- Effect size estimates are based on calculations aligning with Hedges' g as outlined in the *WWC Procedures and Standards Handbook, Version 5.0*

Inference

Cattaneo, Idrobo, and Titiunik (2020) recommend reporting robust bias-correct confidence intervals for improved coverage and inferential properties in finite samples, along with the point estimate obtained using the MSE-optimal bandwidth (p. 71). A term that estimates misspecification bias is removed from the point estimate, and the standard error is adjusted to account for the additional variation that comes with estimating the misspecification bias. Confidence intervals are computed accordingly. Results presented herein will report the conventional point-estimate, standard error, and confidence intervals as well as the robust bias-corrected point-estimate, standard error, confidence intervals and p-values. As a form of sensitivity analysis, results using coverage error probability (CER)-optimal bandwidths (Cattaneo, Idrobo & Titiunik, 2020, p. 70) are also presented.

Differential Assignment & Overall Impact Estimates

As seen earlier in Table 1, a different assignment variable and cut-score was used for each site (see Appendix A for more details). The assignment variable for each site was standardized through a z-score conversion process using the mean and standard deviation of the assignment variable (Gill et al., 2007; Wong et al., 2013). Subsequently, the z-score value associated with the raw-scale cut-score was subtracted from each z-score so that the cut-score was set at zero (Cattaneo, Titiunik, Vazquez-Bare & Keele, 2016). For this report, an overall estimate of ALFA Lab impact was estimated across sites and cohorts. Follow-up analyses conducted by site will be presented in a later report.

Sample Characteristics

Table 2 displays descriptive characteristics of students by school in the ALFA Lab and BAU groups. ALFA Lab students were mostly male and predominantly White at two sites and mostly Hispanic at the NY and CA sites. Although in general the intervention was not designed for students receiving English Language Learner (ELL) services, some students with these classifications ended up in the treatment group. Relatively few students were classified as requiring SPED in the sample.

Table 2. Demographics by Site and Experimental Group

Demographic	School 1 (NY)				School 2 (CT)				School 3 (CA)				School 4 (AL)				Pooled Total					
	ALFA		BAU		ALFA		BAU		ALFA		BAU		ALFA		BAU		ALFA		BAU		Total	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Gender																						
Female	62	36.5%	172	54.6%	0	0.0%	0	0.0%	109	41.3%	156	50.5%	11	47.8%	20	42.6%	182	33.7%	348	41.5%	530	38.5%
Male	108	63.5%	143	45.4%	15	100.0%	76	100.0%	155	58.7%	153	49.5%	12	52.2%	27	57.4%	290	53.7%	399	47.6%	689	50.0%
Missing	0	0.0%	0	0.0%	0	0.0%	0	0.0%	68	20.5%	91	22.8%	0	0.0%	0	0.0%	68	12.6%	91	10.9%	159	11.5%
Race																						
Asian	1	0.6%	5	1.6%	1	6.7%	1	1.3%	3	1.1%	6	1.8%	0	0.0%	0	0.0%	5	0.9%	12	1.4%	17	1.2%
African American	52	31.1%	54	18.1%	4	26.7%	7	9.2%	77	27.3%	54	16.2%	0	0.0%	0	0.0%	133	24.6%	115	13.7%	248	18.0%
Native American	1	0.6%	1	0.3%	0	0.0%	0	0.0%	12	4.3%	8	2.4%	0	0.0%	2	4.3%	13	2.4%	11	1.3%	24	1.7%
Hispanic	75	44.9%	171	57.2%	2	13.3%	9	11.8%	152	53.9%	216	64.9%	0	0.0%	1	2.1%	229	42.4%	397	47.4%	626	45.4%
Multi	1	0.6%	1	0.3%	1	6.7%	2	2.6%	8	2.8%	15	4.5%	0	0.0%	0	0.0%	10	1.9%	18	2.1%	28	2.0%
White	38	22.8%	69	23.1%	7	46.7%	57	75.0%	22	7.8%	31	9.3%	23	100.0%	44	93.6%	90	16.7%	201	24.0%	291	21.1%
Did not respond	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	2.8%	3	0.9%	0	0.0%	0	0.0%	8	1.5%	3	0.4%	11	0.8%
Missing	2	1.2%	14	4.4%	0	0.0%	0	0.0%	50	15.1%	67	16.8%	0	0.0%	0	0.0%	52	9.6%	81	9.7%	133	9.7%
ELL																						
No	126	75.0%	220	73.1%	15	100.0%	76	100.0%	277	85.5%	368	93.2%	23	100.0%	47	100.0%	441	81.7%	711	84.8%	1152	83.6%
Yes	42	25.0%	81	26.9%	0	0.0%	0	0.0%	47	14.5%	27	6.8%	0	0.0%	0	0.0%	89	16.5%	108	12.9%	197	14.3%
Missing	2	1.2%	14	4.4%	0	0.0%	0	0.0%	8	2.4%	5	1.3%	0	0.0%	0	0.0%	10	1.9%	19	2.3%	29	2.1%
SPED																						
No	147	87.5%	301	100.0%	9	60.0%	70	94.6%	320	98.8%	392	99.2%	16	69.6%	46	97.9%	492	91.1%	809	96.8%	1301	94.4%
Yes	21	12.5%	0	0.0%	6	40.0%	6	7.9%	4	1.2%	3	0.8%	7	30.4%	1	2.1%	38	7.0%	10	1.2%	48	3.5%
Missing	2	1.2%	14	4.4%	0	0.0%	0	0.0%	8	2.4%	5	1.3%	0	0.0%	0	0.0%	10	1.9%	19	2.3%	29	2.1%
Total	170	35%	315	65%	76	84%	15	16%	332	45%	400	55%	23	33%	47	67%	540	28%	838	100%	1378	72%

Attrition Calculations

Study attrition occurs when participating units are lost after being assigned to study conditions (e.g., students transfer to another school, drop out of school, or discontinue participation in the study). Table 3 displays the number of students lost after the assignment process, the raw rates of attrition for each school overall (by condition), and the difference in raw attrition rates between the two conditions. To understand how the attrition process might bias estimates, *WWC Standards 5.0* recommends estimating mean attrition rates at the cut-point for both the treatment and control groups using a method that accounts for the forcing variable. The standards also specify the overall and differential attrition rates thresholds for a study to meet design standards with and without reservations. Attrition rate estimates based on WWC recommendations are presented in Appendix B (pp. B4-B5) and were calculated using the same bandwidths and model specifications that were used to obtain impact estimates. WWC-based attrition rates consider sample loss occurring because of missing data (i.e., no outcome scores). Focusing on the intervention group level, among the BAU group a total of 430 students counted toward attrition, with 137 due to leaving their respective school (after assignment) and 395 because of missing outcome data (16% and 47%, respectively). Among the ALFA Lab group, a total of 358 students counted toward attrition, with 86 due to leaving their school and 300 because of missing outcome data (16% and 56%, respectively).

Table 3. Raw Attrition Rates*

Cohort	Site	Group	Total n	Total Attrition	Attrition %	Leaver n	Leaver %	Missing Outcome n	Missing Outcome %
1	CA	BAU	69	45	65.2	30	43.5	37	53.6
		ALFA	125	79	63.2	26	20.8	62	49.6
	NY	BAU	153	116	75.8	35	22.9	112	73.2
		ALFA	43	39	90.7	8	18.6	38	88.4
2	CA	BAU	150	63	42.0	10	6.7	61	40.7
		ALFA	97	58	59.8	4	4.1	53	54.6
	NY	BAU	81	42	51.9	0	0.0	33	40.7
		ALFA	68	64	94.1	0	0.0	42	61.8
3	AL	BAU	47	5	10.6	2	4.3	1	2.1
		ALFA	23	2	8.7	1	4.3	1	4.3
	CA	BAU	181	84	46.4	60	33.1	82	45.3
		ALFA	110	67	60.9	47	42.7	64	58.2
	CT	BAU	76	14	18.4	0	0.0	12	15.8
		ALFA	15	0	0.0	0	0.0	0	0.0
	NY	BAU	81	61	75.3	0	0.0	57	70.4
		ALFA	59	49	83.1	0	0.0	40	67.8
Total		BAU	838	430	51.3	137	16.3	395	47.1
		ALFA	540	358	66.3	86	15.9	300	55.6

*Raw attrition rates are presented here to assess overall sample loss after assignment. These raw rates are not those used by the WWC to understand the impact of attrition on internal validity. Predicted overall and differential attrition rates at the cut-point are available in Appendix B (pp. B4-B5).

Student Outcome Assessment

While ALFA Lab assignment was based on locally defined criteria and data (i.e. the assignment variable differed across sites), we used the same instruments to measure student outcomes across sites. We used a comprehensive online literacy assessment, STAR Reading, to assess reading achievement. To assess the ALFA Lab impact on student reading motivation and frequency, we employed the Adaptive Reading Motivation Measure (ARMM) that measures multiple constructs (e.g., self-efficacy and intrinsic, extrinsic, and social motivation) and yields General Reading Motivation (GRM) and Reading Frequency (FRQ) scores.

Developed by Renaissance Learning, STAR Reading is a computer-adaptive test of reading comprehension (STAR Reading, n.d.). The 10-minute test consists of 25 items selected from a bank of more than 1,000 multiple-choice vocabulary-in-context questions. The vocabulary-in-context items consist of a single sentence with a blank to indicate a missing word. The student must read and complete the sentence, choosing the correct word from a multiple-choice list of three or four words. Vocabulary-in-context items measure comprehension by requiring students to rely on background information, apply vocabulary knowledge, and use active strategies to construct meaning from the assessment text. For the Grade 9 assessment, Cronbach's alpha reliability and test-retest reliability are both at 0.90 based on the STAR norming sample. As noted earlier (pp. iv-1,) due to the Presidential declaration of a national COVID emergency, Cohort 1 schools closed abruptly on March 13, 2020. Thus, post-intervention assessment of the cohort's participating students was postponed until 2020-2021 when they were 10th graders – in October 2020 at School 3 (CA) and in February 2021 at School 1 (NY). However, in Cohorts 2 and 3, the STAR assessment was administered as planned in the second semester of students' ninth-grade year.

The ARMM is a 20-minute computer-adaptive assessment including 45 items assessing GRM overall and 15 subconstructs—curiosity, involvement, interest, value, challenge, grades, recognition, competition, avoidance, self-efficacy, perceived difficulty, preference for autonomy, social motivation, prosocial goals, and antisocial goals for reading (Davis et al., 2020). The ARMM, developed as part of a separate IES grant-funded project, assesses GRM as indicated by six subfactors: self-efficacy, intrinsic motivation, extrinsic motivation, reading autonomy, (lack of) reading avoidance, and social motivation. The instrument was validated using confirmatory multidimensional item response theory (IRT) using a sample of 7,449 students (Kingston et al., 2017) exhibiting adequate levels of both construct and criterion validity with internal reliability of subscales ranging from 0.70 to 0.84 and an internal reliability of 0.94 for GRM. The ARMM assessment was administered to Cohort 1 students at School 3 (CA) during the fall of 2020 and to Cohort 1 and 2 students at School 1 (NY) beginning in February 2021. Cohort 3 participants at all schools completed the ARMM in April or early May, including fall and spring semester participants at School 3 (CA).

Implementation Questions (RQ4)

Data to address implementation fidelity across the three cohorts of students in 2019–2020, 2020–2021, and 2021–2022 school years were drawn from site visits and teacher surveys. Each visit included an interview or focus group with one administrator, at least one ALFA Lab teacher, and at least one ELA

(non-ALFA Lab) teacher regarding contextual information, literacy instruction, and, if applicable, perceptions of ALFA Lab implementation. Following each site visit, ICF also administered a short survey to ALFA Lab teachers regarding ALFA Lab implementation. Table 4 provides an overview of data collection across the cohorts.

Table 4. Overview of Data Collection in Each Cohort

Data Collection	Cohort 1 (2019–20)	Cohort 2 (2020–21)	Cohort 3 (2021–22)
Site Visits	N=3 (in-person)	N=3 (virtual)	N=5 (virtual)
ALFA teacher interview	Yes	Yes	Yes
Comparison teacher focus group	Yes	No	Yes
Principal interview	Yes	No	Yes
Classroom observation	Treatment and comparison groups with PLATO and CFAI	No	Treatment group with PLATO and CFAI
ALFA teacher survey (N and %)	5 (100%)	3 (100%)	6 (100%)

Note. School 3 (CA) conducted ALFA Lab during both fall and spring semesters, and participated in two site visits—one during November and the other in March—and completed the survey once per semester in Year 2 and Year 3.

The interview/focus group protocols and teacher survey were developed by ICF with input from JHU. The observation protocol consists of two components: Checklist for ALFA Implementation (CFAI) and Protocol for Language Arts Teacher Observations (PLATO). The CFAI was developed by JHU and reviewed by ICF. PLATO is a classroom observation protocol that was developed by Stanford University to measure features of ELA instruction (Grossman et al., 2013). The ICF research team members completed training and were certified on PLATO use to ensure reliability and accuracy in scoring. Each session was observed and scored by two PLATO-certified reviewers.

ICF researchers coded all transcripts and notes according to a range of themes. Observation data were recorded and scored using the CFAI and PLATO protocols (findings are presented in Appendices E and F). ICF researchers conducted a descriptive analysis of the survey results (findings are presented in Appendix D).

Cost Questions (RQ5)

The cost analysis is designed to help schools and districts understand the monetary costs of implementing ALFA Lab. We will examine two types of costs: (1) the total cost of the program (value of all goods and services used to implement the program, including in-kind or donated resources) and (2) the cost of the program borne by the school (expenditures across the lifespan of the program including start-up costs and maintenance costs). Costs were calculated with the ingredients method, which accounts for all resources used to achieve the outcomes (Levin & McEwan, 2001). Both annual costs and per-student costs will be calculated and reported.

The costs for the resources were gathered from the program developer, JHU, and surveys of teachers and school principals. These were supplemented with the national prices, or cost in 2020–2021 dollars. National prices are obtained from sources such as the U.S. Bureau of Labor Statistics and the National Center for Education Statistics. Although the cost analysis is only generalizable to schools in the districts

in our sample, the findings can be applied to many other similar high-poverty urban districts across the country.

The goal of the ingredients method is to identify all resources required to implement a program—including personnel, facilities, equipment, and materials—whether paid for directly or contributed in-kind, and ascertain their costs. A cost analysis should incorporate perspectives from all stakeholders. For this study, the following items were reported for use in the program by the program developer, participating teachers, and principals as part of our data collection for the implementation study, and, therefore, were considered the main ingredients for the ALFA Lab program (Table 5). By collecting the information in real time and as part of the overall data collection, we avoided additional data collection burden as well as recall issues experienced by other cost analyses (Levin et al., 2018).

Table 5. Ingredients in ALFA Lab Cost Analysis

Category	Ingredients
Personnel	Teachers/Lab Assistants
	Principals
	Training Staff
Facilities	Classroom
Equipment	Computers/Electronic Tablets
	Printers
	Projectors
Materials	*Feisty Felines Teacher's Manual*
	Feisty Felines Nonconsumable Resources
	Sets of Feisty Felines Reading Selections & Station Activities
	Heroes Teacher's Manual
	Heroes Nonconsumable Resources
	Sets of Heroes Reading Selections & Station Activities
	Galaxy Teacher's Manual
	Galaxy Nonconsumable Resources
	Sets of Galaxy Reading Selections & Station Activities
Other Inputs	Student Field Trip
	Training Staff Travel
	School Supplies

Teachers and principals also reported the quantities of the ingredients. For example, according to the teacher survey, the four schools offered an average of three ALFA Lab class sections per semester, ranging from one to eight sections. Each class section served an average of 22 students, with class sizes varying between 15 and 24 students. Teachers reported that the ALFA Lab classes usually started at the beginning of the school year. One school reported a longer class length at 70–90 minutes long, and the other three schools reported a regular class time of 45–55 minutes. In addition, schools also reported the quantities of materials, equipment, and facilities used, which are described in detail in the following sections.

To ensure comparability across sites in different locations and to inform replication efforts nationally, the costs presented in this report were estimated based on the following assumptions:

- Average costs were estimated for operating one ALFA Lab (one class of 22 students) in one semester (13 weeks).
- National prices were used to avoid estimating costs that reflect special local-market conditions only. Most of the national prices used in this analysis were obtained from the Center for Benefit-Cost Studies for Education database of educational resource prices available through the cost tool, CostOut (Hollands et al., 2015).
- Facilities and equipment costs were amortized to reflect the life value of these resources beyond the term of this evaluation to provide annualized measures.
- All costs are reported or converted using 2021 dollars.

Reading Motivation Profiles Questions (RQ6)

To examine the relationship of adolescents' motivational profiles and their reading outcomes, motivational profiles were first derived from students' responses to the ARMM. Reading motivation profiles were obtained using latent profile analysis (LPA), a latent variable mixture modeling approach, and the relationship between the resulting profiles and reading outcomes, specifically reading frequency (i.e., ARMM Frequency) and reading achievement (i.e., STAR Reading assessment) were then examined.

LPA is a clustering technique that adopts a person-centered approach to identify subpopulations of a latent construct based on responses to continuous indicators. LPA is advantageous over standard clustering techniques in that it allows for different scales that do not need to be transformed to standardized values. Considering that the ARMM is a computer-adaptive measure where the computer displays subsequent sets of items based on student responses on the initial set of items, it is possible that student scores on each of the 15 constructs have varying distributions. Use of LPA prevents rescaling of student scores on the ARMM.

Moreover, LPA allows for data-driven extraction of profiles with several different possible model configurations, where—depending on the model configuration—the identified classes or profiles can show unique patterns of means and variances. These model configurations can be freely estimated in each class specifically or constrained to be equal across classes; thus, can accommodate models of varying complexity. Model fit criteria, like that of the Bayesian Information Criteria (BIC), can help determine the number of classes or profiles to ensure that model complexity does not compromise model fit (Wardenaar, 2021).

Table 6. Fit Indices for Different Models with Varying Number of Profiles

Model	AIC	BIC	BLRT *p*	Entropy
2 Clusters/Profiles	13628.38	13628.38	0.003	0.92
3 Clusters/Profiles	12502.79	12744.92	0.002	0.95
4 Clusters/Profiles	11990.88	11686.26	0.629	0.91

To proceed with LPA, ARMM data from the treatment and comparison groups from Cohorts 1 and 2 were included in analysis. Table 6 displays the model fit indices for three models with the number of

profiles ranging from 2 to 4. From this table, although the four-profile solution indicated adequate model fit with lower AIC and BIC values, entropy and bootstrap likelihood ratio test (BLRT) statistics indicate that model fit was not adequate. Moreover, when considering profile discrimination, the four-profile solution yielded two redundant profiles, which did not provide theoretical meaningfulness. Finally, the four-profile solution yielded a profile with fewer numbers of cases when compared to the other profiles. For these reasons, the three-profile solution supported by adequate model fit indices was selected because they are theoretically meaningful and were observed to tap into varying dimensions of a higher order dimension of GRM. The three-reading motivation profile solution was used for subsequent analysis to examine their relationship to reading frequency and achievement.

Results

This section presents findings for the research questions RQ1–RQ6. For each impact question in RQ1–RQ3, we describe data compiled across all four schools, starting with the graphical analysis, and followed by the RD estimate results. More details related to model falsification/*WWC Standards 5.0* are included in Appendix B. Supplementary analyses focused on exploring whether RD estimates of ALFA Lab were mediated or moderated by factors as represented in the Theory of Change (Figure 1 on p. 5) are summarized in Appendix C.

Research Question 1 – Confirmatory (STAR Assessment)

Summary of Findings

The evidence associated with the STAR outcome suggests a descriptive advantage for the ALFA Lab group (i.e., estimates were positively associated with ALFA Lab group membership; see Figure 4), but inferentially this advantage was non-significant regardless of estimand (i.e., ITT versus complier average causal effect/Fuzzy RD) or estimation method. Although evidence supported the integrity of the forcing variable, attrition rates were high and beyond the WWC threshold for concern. However, much of the sample loss was due to COVID-19 and impacted both the treatment and control groups equally. Excluding sample loss related to the pandemic, attrition was low. Assessment of the continuity of the relationship between the outcome and forcing variable lent support for continuity as none of the impact estimates at placebo cut-scores were statistically significant. All evidence related to the WWC RD standards can be seen in Appendix B.

Graphical Analysis

Figure 2 plots the standardized and centered running variable across all cohorts and sites along the x-axis and the STAR scale score on the y-axis for students in the analytic sample (i.e., those with complete data on demographic background variables and cohort indicators). The lines are locally estimated scatterplot smoothing (LOESS) lines, constructed to find a curve of best fit without assuming a formal

distributional shape. Examining the lines closely, the ALFA Lab group appears to intersect the dotted line at x=*c* at a point below where the line representing the BAU group intersects the dotted line (i.e., there

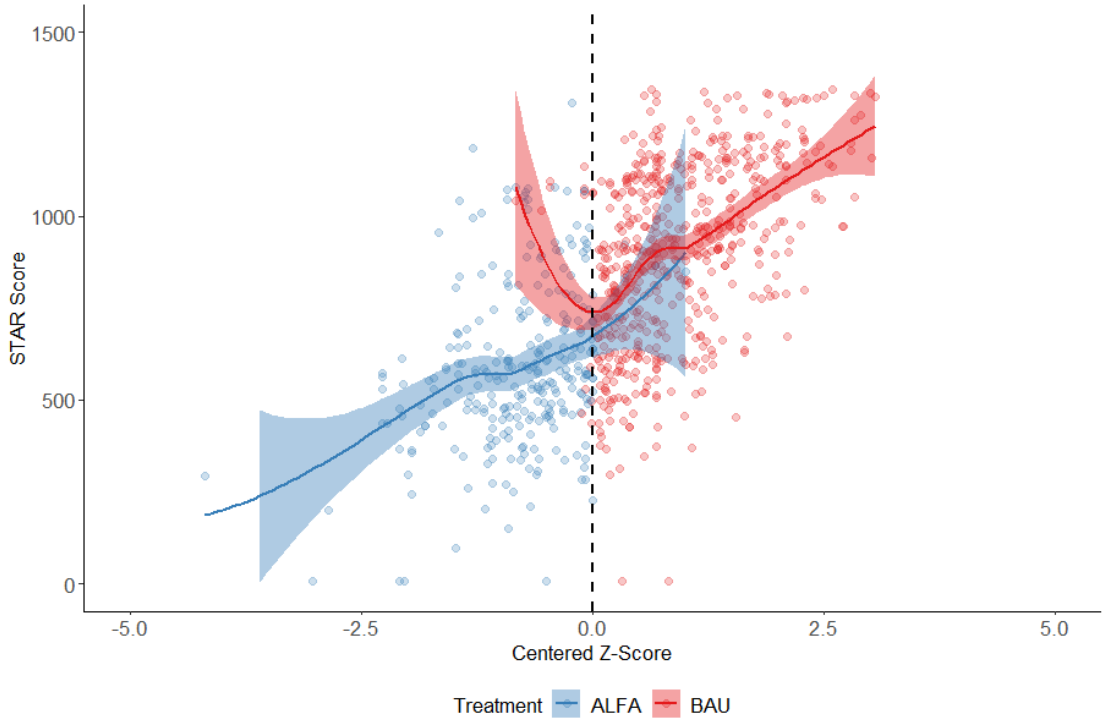


Figure 2. Fuzzy Experimental Group Assignment Scatterplot -- STAR

is no visual evidence supporting positive average treatment effects). Note that Figure 2 displays the data from a Fuzzy RD perspective, where the non-complying units (i.e., the students with assignment score values below the cut-score of zero that were members of the BAU group and vice versa) are presented according to the group the students experienced (i.e., a student with an assignment score below the cut-score of zero who did *not* experience ALFA Lab is considered part of BAU). An ITT perspective would present each student based on the intervention group they *should* have been in under full compliance. Those students shaded in red with assignment scores below zero tended to have relatively higher STAR outcome scores, as shown by the upward curve of the red line to the left of the cut-score. This suggests that the external information that schools paid attention to in "overruling" the assignment variable might have rightly determined that, for these few, their relatively low score on the assignment variable did not reflect their typical level of performance. (That is, for a few students, when a later or different achievement test or superior grades obtained in a later marking period suggested that the student did not really need ALFA, they were left untreated on the assumption that they must not have put forth full effort when taking the assessment test used as the assignment variable for ALFA.)

To gain a better view of the potential discontinuity using smoothed scatter plots, we explore both the ES and QS spacing options, along with the IMSE and MV methods to determine the number of bins to use in the smoothing process and plot. The MV method yielded a larger number of bins than the IMSE method and yielded bins that were, on average, smaller in length. Table 7 displays the bins identified—by spacing style, kernel, and method—for the STAR outcome data. Figure 3 contains graphical representations of the four sets of triangular kernel bins from Table 7.

Table 7. Outcome Bin Identification Results, by Spacing and Method – STAR

Bin Type	Kernel	Spacing	n bins (Left / Right)	Mean Length (Left / Right)	Median Length (Left / Right)
MV	Uniform	Equal	22 / 33	[0.19, 0.09]	[0.19 - 0.09]
		Quantile	21 / 26	[0.20, 0.12]	[0.08 - 0.08]
IMSE		Equal	8 / 14	[0.52, 0.22]	[0.52 - 0.22]
		Quantile	18 / 13	[0.23, 0.23]	[0.09 - 0.16]
MV	Triangular	Equal	22 / 33	[0.19, 0.09]	[0.19 - 0.09]
		Quantile	21 / 26	[0.20, 0.12]	[0.08 - 0.08]
IMSE		Equal	8 / 14	[0.52, 0.22]	[0.52 - 0.22]
		Quantile	18 / 13	[0.23, 0.23]	[0.09 - 0.16]

The equally-spaced, mimicking-variance (ESMV) binned plot with a fourth-order polynomial superimposed can be seen in the lower left quadrant of Figure 3, followed by a quantile-spaced, mimicking-variance (QSMV) fourth-order binned plot in the lower right quadrant. The comparison of the ESMW and QSMV plots reveal the greater amount of variability among ALFA Lab data elements as the mean bin values (i.e., dots) shift between the two, whereas BAU mean bins are relatively stable. Moving to the IMSE plots designed to elucidate the overall regression shape, we see that between ES and QS the shape is relatively consistent, though again the ES version stretches farther to the left among the ALFA Lab group. This aligns with what we see in the raw data scatterplot in Figure 2, with very few ALFA Lab units to the far left of the cut-score continuum. In all cases in Figure 4, there appears to be a small advantage for the ALFA Lab group (i.e., the ALFA Lab line is higher on the y-axis at the cut-point relative to the BAU group.)

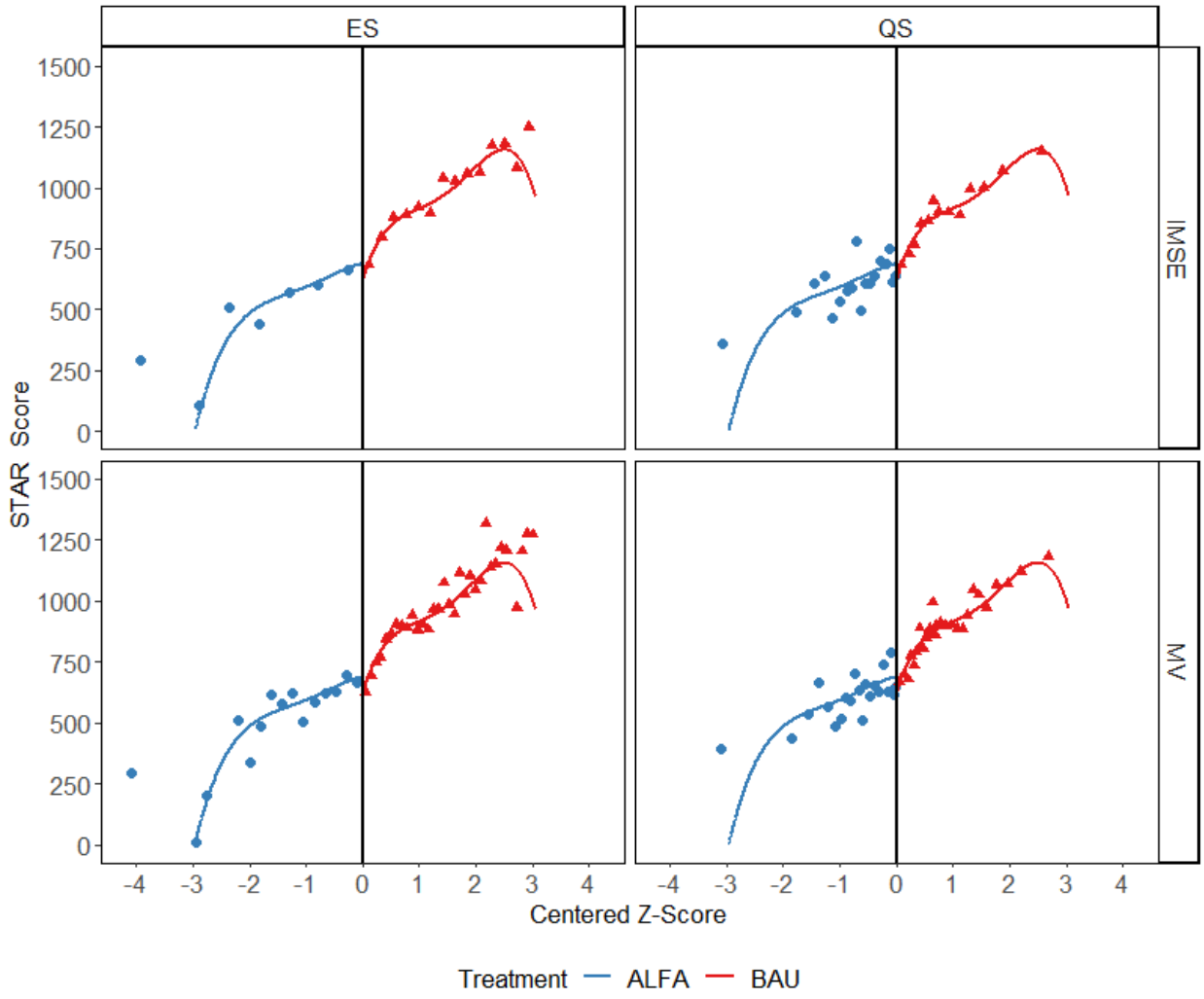


Figure 3. Smoothed Panel Plot (p=4) – STAR RD Estimation Results

RD Estimation Results

Proceeding with the estimation of the ALFA Lab effect under the continuity-based framework, Table 8 displays the ITT and Fuzzy design estimates obtained using STAR as the outcome variable. As a form of sensitivity analysis, we present results using two different bandwidth types (MSE-optimal and CER-optimal), two weighting kernels, two bandwidth selections (Common, meaning the same bandwidth on each side of the cutoff; Separate, meaning a different bandwidth on each side of the cutoff), and two different polynomials ("1" is a linear specification, "2" is quadratic). We prioritize models that use a triangular kernel (to give highest weight to units closest to the cutoff), separate bandwidths, and

polynomials of Order 1.[5] Models for both designs include the following additional covariates: student gender and dummy variables representing individual race categories with "did not respond" counting as its own category, ELL and special education status as well as two dummies representing membership in Cohorts 1 and 2 respectively. We focus on the MSE-optimal point estimates with the triangular kernel, though p-values from corresponding CER-optimal bandwidths are provided as well (see Appendix A for details). Though almost all MSE-optimal ITT and Fuzzy point estimates presented are in favor of ALFA Lab, none of the effects were statistically significant. Evidence related to the assessment of WWC Procedures and Standards Handbook, Version 5.0 can be viewed in Appendix B.

[5] Guidance from Cattaneo, Idrobo, and Titiunik (2020) suggests that in most instances, a polynomial of Order 1 yields adequate asymptotic variances and reduced asymptotic bias, without overfitting that occurs when using higher-order polynomials. Use of a triangular kernel, that weights observations closer to the cut-score more heavily than observations farther away from the cut-score, are MSE-optimal.

Table 8. Continuity-Based (CB) Estimates – STAR

Design	BW Type	Kernel	BW Select	p	n (TX / CT)	Bandwidth	Est (se)	CI	Robust Est (se)	Robust p-value	Robust CI	ES
ITT	MSE	Uniform	Common	1	155 / 100	[-0.47 - 0.47]	19.75 (39.55)	[-57.77, 97.27]	26.14 (45.29)	0.564	[-62.63, 114.90]	0.083
			Separate	1	116 / 133	[-0.36 - 0.60]	27.30 (45.73)	[-62.32, 116.92]	19.53 (52.71)	0.711	[-83.79, 122.85]	0.114
			Common	2	303 / 177	[-0.85 - 0.85]	55.76 (46.60)	[-35.58, 147.09]	64.33 (50.44)	0.202	[-34.54, 163.19]	0.233
			Separate	2	260 / 169	[-0.69 - 0.79]	29.37 (50.06)	[-68.73, 127.48]	23.63 (56.19)	0.674	[-86.49, 133.76]	0.123
		Triangular	Common	1	179 / 111	[-0.51 - 0.51]	10.40 (37.94)	[-63.97, 84.76]	11.58 (45.94)	0.801	[-78.46, 101.62]	0.044
			Separate	1	146 / 152	[-0.44 - 0.70]	26.19 (40.95)	[-54.08, 106.45]	22.95 (48.50)	0.636	[-72.11, 118.01]	0.110
			Common	2	225 / 138	[-0.66 - 0.66]	-4.80 (52.14)	[-107.00, 97.40]	-14.19 (58.94)	0.810	[-129.71, 101.34]	-0.020
			Separate	2	307 / 140	[-0.88 - 0.66]	2.40 (45.72)	[-87.20, 92.01]	1.15 (50.75)	0.982	[-98.32, 100.61]	0.010
	CER	Uniform	Common	1	102 / 82	[-0.34 - 0.34]	-32.51 (52.62)	[-135.63, 70.62]	-30.28 (55.51)	0.585	[-139.08, 78.52]	-0.136
			Separate	1	76 / 94	[-0.26 - 0.43]	34.87 (51.63)	[-66.33, 136.06]	34.87 (54.59)	0.523	[-72.12, 141.86]	0.146
			Common	2	203 / 122	[-0.58 - 0.58]	17.44 (56.23)	[-92.77, 127.65]	21.24 (57.43)	0.712	[-91.33, 133.80]	0.073
			Separate	2	156 / 116	[-0.47 - 0.54]	-15.68 (64.10)	[-141.31, 109.96]	-18.48 (65.52)	0.778	[-146.89, 109.94]	-0.066
		Triangular	Common	1	117 / 86	[-0.36 - 0.36]	2.17 (44.43)	[-84.91, 89.24]	3.61 (48.93)	0.941	[-92.28, 99.51]	0.009
			Separate	1	92 / 107	[-0.31 - 0.50]	17.37 (47.22)	[-75.17, 109.91]	17.61 (50.76)	0.729	[-81.87, 117.09]	0.073
			Common	2	149 / 96	[-0.45 - 0.45]	-11.85 (59.04)	[-127.56, 103.86]	-14.10 (60.99)	0.817	[-133.63, 105.43]	-0.050
			Separate	2	206 / 97	[-0.60 - 0.45]	-5.91 (53.57)	[-110.90, 99.07]	-5.56 (55.27)	0.920	[-113.89, 102.77]	-0.025
Fuzzy*	MSE	Uniform	Common	1	209 / 134	[-0.61 - 0.61]	23.92 (43.95)	[-62.22, 110.06]	56.88 (51.26)	0.267	[-43.58, 157.34]	0.101
			Separate	1	136 / 112	[-0.41 - 0.52]	28.77 (52.41)	[-73.96, 131.50]	32.61 (58.71)	0.579	[-82.45, 147.67]	0.121
			Common	2	271 / 154	[-0.72 - 0.72]	18.57 (65.51)	[-109.83, 146.97]	7.27 (76.75)	0.925	[-143.14, 157.69]	0.078
			Separate	2	260 / 162	[-0.69 - 0.76]	36.23 (66.84)	[-94.77, 167.22]	35.98 (75.76)	0.635	[-112.50, 184.45]	0.153
		Triangular	Common	1	283 / 159	[-0.75 - 0.75]	28.44 (41.29)	[-52.48, 109.37]	50.94 (49.85)	0.307	[-46.76, 148.64]	0.120
			Separate	1	165 / 126	[-0.49 - 0.59]	22.66 (50.23)	[-75.79, 121.10]	22.92 (57.76)	0.692	[-90.28, 136.12]	0.096
			Common	2	310 / 180	[-0.89 - 0.89]	24.89 (60.41)	[-93.51, 143.29]	10.98 (71.05)	0.877	[-128.28, 150.25]	0.105
			Separate	2	313 / 193	[-0.90 - 0.93]	28.41 (59.91)	[-89.02, 145.84]	27.02 (66.24)	0.683	[-102.81, 156.84]	0.120
	CER	Uniform	Common	1	146 / 95	[-0.43 - 0.43]	10.04 (55.06)	[-97.88, 117.95]	29.58 (59.06)	0.617	[-86.18, 145.34]	0.042
			Separate	1	90 / 86	[-0.29 - 0.37]	30.55 (66.87)	[-100.51, 161.61]	32.61 (69.94)	0.641	[-104.47, 169.70]	0.129
			Common	2	167 / 105	[-0.49 - 0.49]	-37.48 (83.09)	[-200.33, 125.38]	-43.68 (86.49)	0.614	[-213.19, 125.83]	-0.158
			Separate	2	155 / 114	[-0.47 - 0.52]	-26.72 (90.85)	[-204.78, 151.35]	-29.34 (93.05)	0.753	[-211.70, 153.03]	-0.113
		Triangular	Common	1	182 / 115	[-0.53 - 0.53]	16.52 (48.31)	[-78.17, 111.21]	27.88 (53.82)	0.604	[-77.60, 133.36]	0.070
			Separate	1	112 / 93	[-0.35 - 0.42]	6.30 (60.05)	[-111.39, 124.00]	7.39 (64.18)	0.908	[-118.39, 133.17]	0.027
			Common	2	209 / 134	[-0.61 - 0.61]	-11.18 (74.55)	[-157.30, 134.93]	-16.68 (78.84)	0.832	[-171.21, 137.84]	-0.047
			Separate	2	212 / 136	[-0.62 - 0.63]	-8.17 (74.31)	[-153.81, 137.47]	-8.45 (76.72)	0.912	[-158.82, 141.93]	-0.034

*See Appendix B, Exhibit B5.1 for first-stage regression output related to the participation indicator.

Figure 4 plots the data within the bandwidth, superimposing the local linear regression lines estimated using a triangular kernel and linear specification for the ITT and Fuzzy designs.

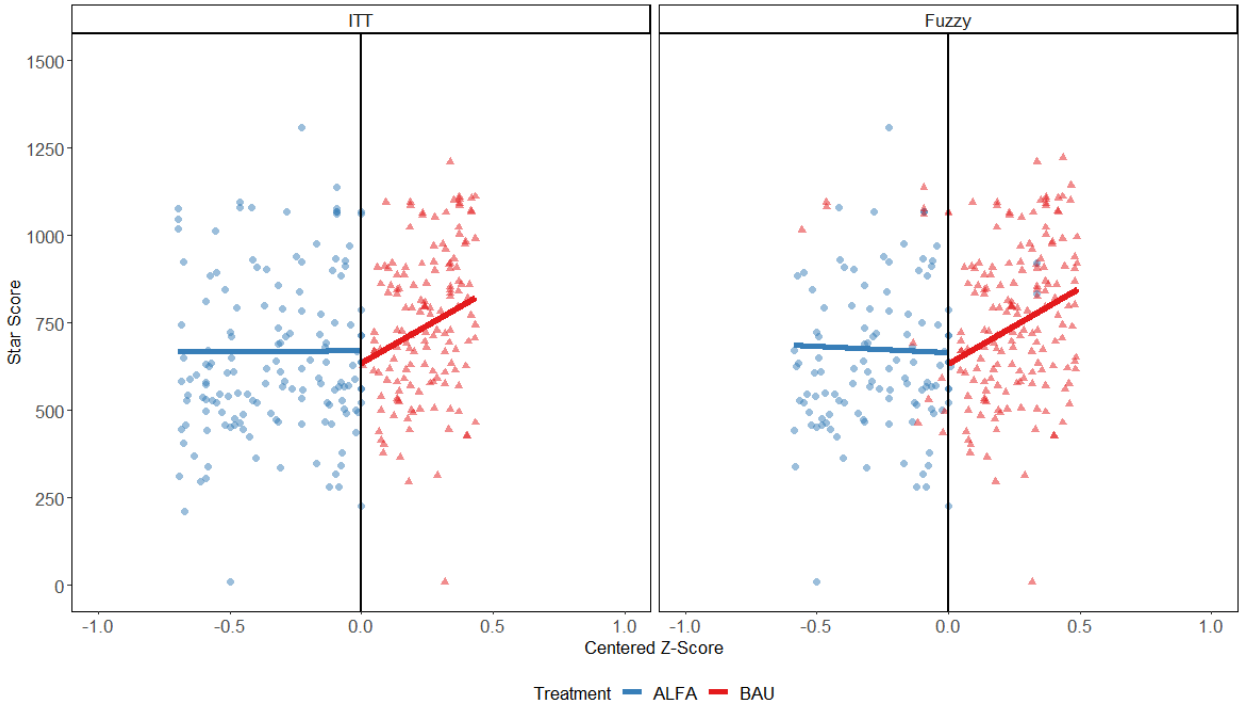


Figure 4. Estimated ITT and Fuzzy RD Effect Plots – STAR

Research Question 2 – Exploratory (ARMM General Reading Motivation)

Summary of Findings

The evidence associated with the ARMM General Reading Motivation (GRM) outcome suggests a descriptive advantage for the ALFA Lab group (i.e., estimates were positively associated with ALFA Lab group membership; see Figure 7 on p. 29), but inferentially this advantage was non-significant regardless of estimand (i.e., ITT versus complier average causal effect/Fuzzy RD) or estimation method. Attrition rates were again high and beyond the WWC threshold for concern, though excluding sample loss due to COVID-19 again renders attrition rates low. Assessment of the continuity of the relationship between the outcome and forcing variables indicated a lack of continuity as the rate of significant impact estimates associated with placebo cut-scores were higher than acceptable WWC thresholds. All evidence related to the WWC RD standards can be seen in Appendix B.

Graphical Analysis

Figure 5 plots the standardized and centered running variable across all cohorts and sites along the x-axis and the GRM scale score on the y-axis for students in the analytic sample (i.e., those with complete

data on demographic background variables and cohort indicators). The lines are LOESS lines, constructed to find a curve of best fit without assuming a formal distributional shape. Examining the lines closely, the ALFA Lab group appears to intersect the dotted line at x=*c* at a point roughly equivalent to where the line representing the BAU group intersects the dotted line (i.e., there is no visual evidence supporting positive average treatment effects). Note that Figure 5 displays the data from a Fuzzy RD perspective, where the non-complying units (i.e., the students with assignment score values below the cut-score of zero that were members of the BAU group and vice versa) are presented according to the groups students experienced (i.e., a student with an assignment score below the cut-score of zero that did *not* experience ALFA Lab is considered part of BAU). An ITT perspective would present each student based on the intervention group they *should* have been in under full compliance. Those students shaded in red with assignment scores below zero tended to have relatively higher GRM outcome scores, as shown by the upward curve of the red line to the left of the cut-score.

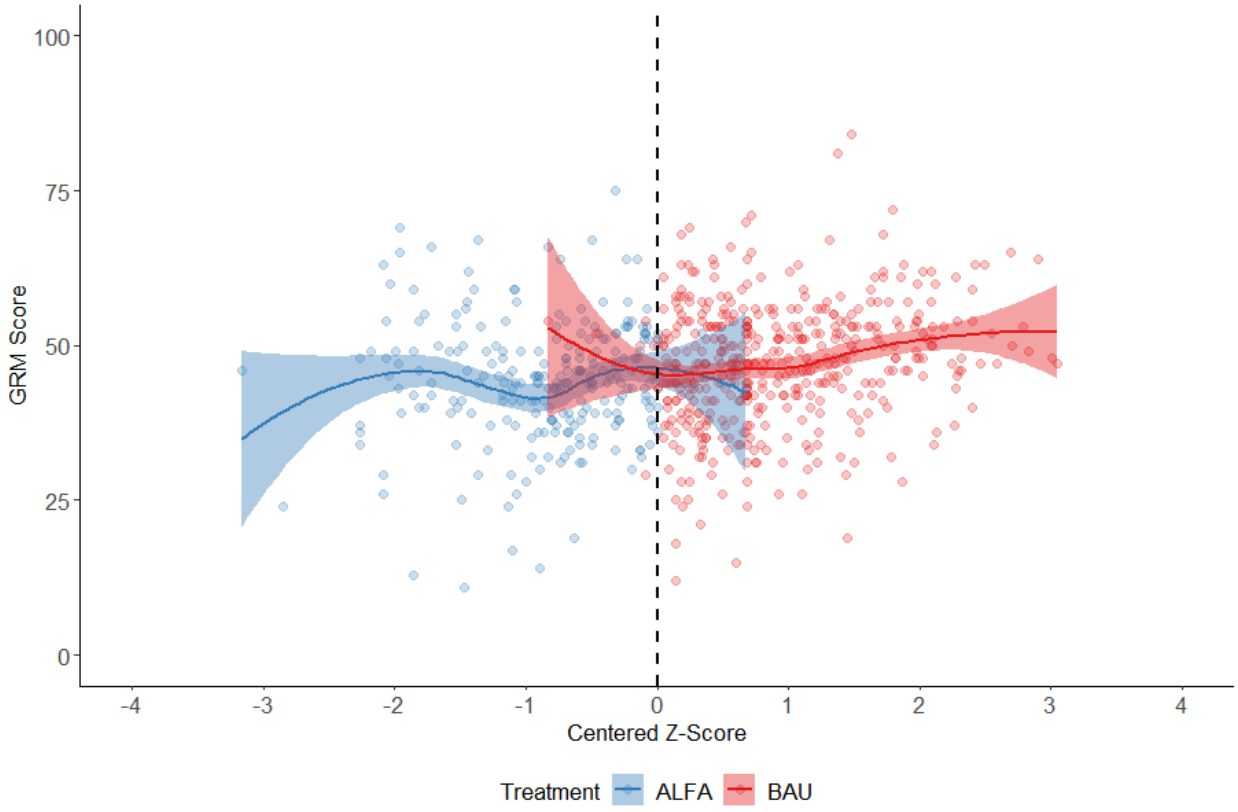


Figure 5. Fuzzy Experimental Group Assignment Scatterplot – ARMM GRM

We again explore both the ES and QS spacing options, along with the IMSE and MV methods to determine the number of bins to use in the smoothing process and plot. The MV method yielded a larger number of bins than the IMSE method and yielded bins that were, on average, smaller in length. Table 9 displays the bins identified—by spacing style, kernel, and method—for the GRM outcome data. Figure 6 contains graphical representations of the four sets of triangular kernel bins from Table 9.

Table 9. Outcome Bin Identification Results, by Spacing and Method – ARMM GRM

Bin Type	Kernel	Spacing	n bins (Left / Right)	Mean Length (Left / Right)	Median Length (Left / Right)
MV	Uniform	Equal	13 / 20	[0.24, 0.15]	[0.24 - 0.15]
		Quantile	17 / 18	[0.19, 0.17]	[0.10 - 0.13]
IMSE		Equal	5 / 5	[0.63, 0.61]	[0.63 - 0.61]
		Quantile	16 / 6	[0.20, 0.51]	[0.10 - 0.35]
MV	Triangular	Equal	13 / 20	[0.24, 0.15]	[0.24 - 0.15]
		Quantile	17 / 18	[0.19, 0.17]	[0.10 - 0.13]
IMSE		Equal	5 / 5	[0.63, 0.61]	[0.63 - 0.61]
		Quantile	16 / 6	[0.20, 0.51]	[0.10 - 0.35]

The ESMV binned plot with a fourth-order polynomial superimposed can be seen in the lower left quadrant of Figure 6, followed by a QSMV fourth-order binned plot in the lower right quadrant. The comparison of the ESMW and QSMV plots reveals the greater amount of variability among ALFA Lab data elements as the mean bin values (i.e., dots) shift between the two, whereas BAU mean bins are relatively stable. Moving to the IMSE plots designed to elucidate the overall regression shape, we see that between ES and QS the shape (near-linear) is relatively consistent, though again the ES version has more space between points among the ALFA Lab group. This aligns with what we see in the raw data scatterplot in Figure 5, with very few ALFA Lab units to the far left of the cut-score continuum. In all cases in Figure 6, there appears to be a negligible advantage for the ALFA Lab group (i.e., the ALFA Lab line is at the same point on the y-axis at the cut-point relative to the BAU group).

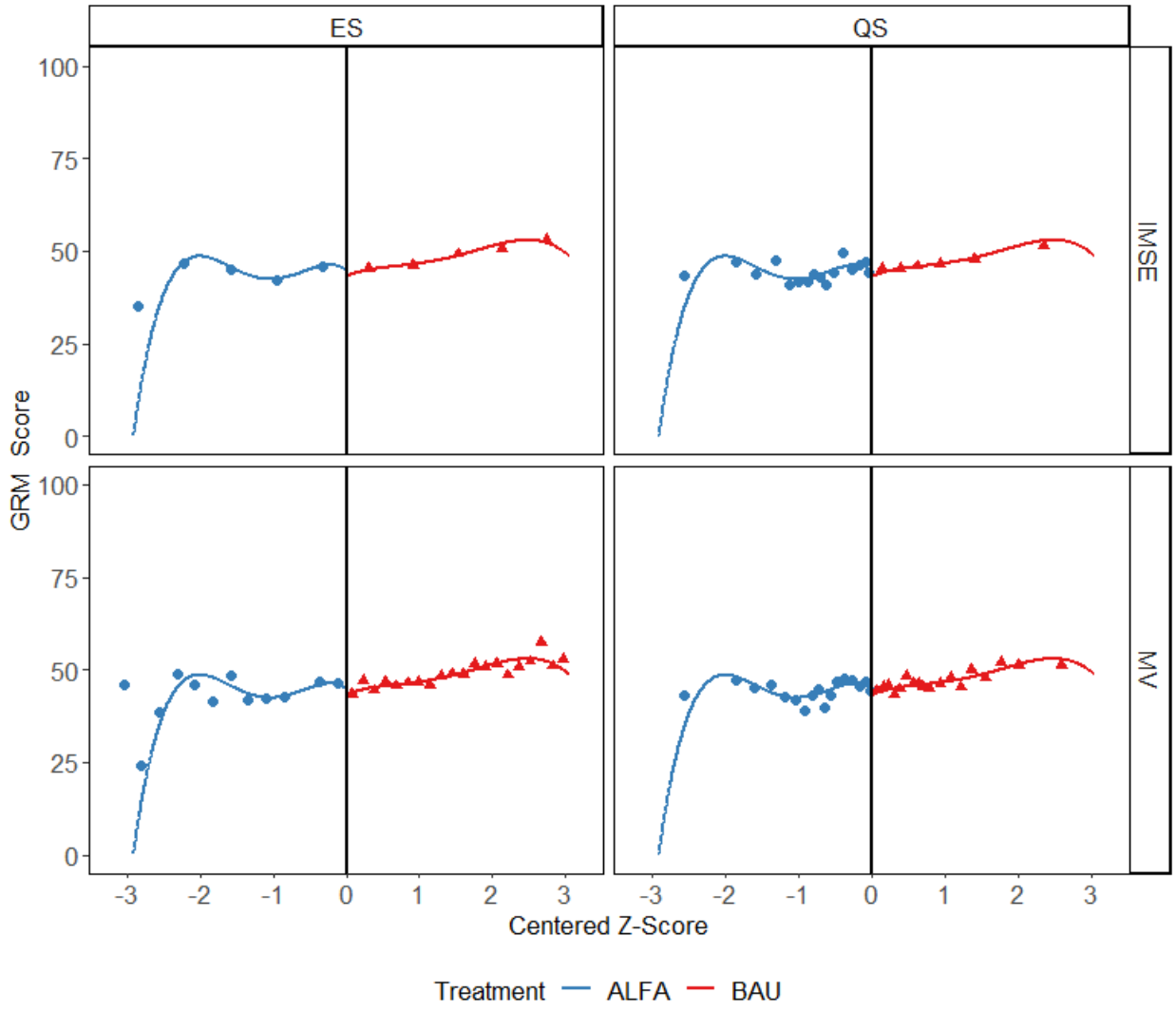


Figure 6. Smoothed Panel Plot (p=4) – GRM RD Estimation Results

RD Estimation Results

Table 10 on p. 28 displays the ITT and Fuzzy design estimates obtained using ARMM GRM as the outcome variable under the continuity-based framework. We present results using two different bandwidth types (MSE-optimal and CER-optimal), two weighting kernels, two bandwidth selections (Common, meaning the same bandwidth on each sit of the cutoff; Separate, meaning a different bandwidth on each side of the cutoff), and two different polynomials ("1" is a linear specification, "2" is quadratic). We again prioritize models that use a triangular kernel (to give highest weight to units closest to the cutoff), separate bandwidths, and polynomials of Order 1. Models for both designs include the following additional covariates: student gender and dummy variables representing individual race categories with "did not respond" counting as its own category, ELL and SPED status as well as two dummies representing membership in Cohorts 1 and 2, respectively. We focus on the MSE-optimal point estimates yielded with the triangular kernel, though p-values from corresponding CER-optimal bandwidths are provided as well (see Appendix A for details). Focusing on the ITT estimates, all but two

estimates were positive in favor of ALFA Lab, though none were statistically significant. The Fuzzy design results included only a single negative estimate with all others again favoring ALFA Lab (though none statistically significant).

Table 10. Continuity-Based (CB) Estimates – GRM

Design	BW Type	Kernel	BW Select	p	n (TX / CT)	Bandwidth	Est (se)	CI	Robust Est (se)	Robust p-value	Robust CI	ES
ITT	MSE	Uniform	Common	1	181 / 108	[-0.65 - 0.65]	3.12 (1.96)	[-0.72, 6.96]	3.32 (2.18)	0.127	[-0.94, 7.59]	0.318
			Separate	1	176 / 79	[-0.62 - 0.49]	1.17 (2.04)	[-2.83, 5.16]	1.30 (2.29)	0.570	[-3.19, 5.79]	0.119
			Common	2	123 / 73	[-0.42 - 0.42]	0.71 (3.49)	[-6.14, 7.55]	0.35 (3.86)	0.928	[-7.22, 7.92]	0.072
			Separate	2	266 / 65	[-0.98 - 0.35]	1.77 (2.53)	[-3.19, 6.72]	2.31 (2.71)	0.394	[-3.00, 7.61]	0.180
		Triangular	Common	1	145 / 80	[-0.50 - 0.50]	0.71 (2.14)	[-3.49, 4.91]	0.23 (2.50)	0.925	[-4.67, 5.14]	0.072
			Separate	1	224 / 73	[-0.72 - 0.43]	1.05 (1.94)	[-2.75, 4.86]	0.82 (2.24)	0.714	[-3.57, 5.22]	0.107
			Common	2	232 / 128	[-0.76 - 0.76]	0.46 (2.45)	[-4.35, 5.27]	0.07 (2.74)	0.980	[-5.31, 5.45]	0.047
			Separate	2	265 / 122	[-0.97 - 0.71]	0.69 (2.35)	[-3.92, 5.30]	0.53 (2.59)	0.839	[-4.55, 5.60]	0.070
	CER	Uniform	Common	1	131 / 76	[-0.47 - 0.47]	1.24 (2.25)	[-3.17, 5.65]	1.38 (2.35)	0.557	[-3.23, 6.00]	0.127
			Separate	1	128 / 65	[-0.45 - 0.35]	0.74 (2.39)	[-3.95, 5.44]	0.84 (2.52)	0.740	[-4.11, 5.78]	0.076
			Common	2	83 / 54	[-0.29 - 0.29]	-2.52 (3.96)	[-10.28, 5.24]	-2.75 (4.07)	0.499	[-10.74, 5.23]	-0.258
			Separate	2	192 / 49	[-0.67 - 0.24]	3.30 (2.73)	[-2.04, 8.65]	3.46 (2.80)	0.216	[-2.02, 8.94]	0.337
		Triangular	Common	1	103 / 65	[-0.36 - 0.36]	0.76 (2.27)	[-3.69, 5.21]	0.54 (2.49)	0.828	[-4.35, 5.43]	0.078
			Separate	1	149 / 59	[-0.52 - 0.31]	0.61 (2.13)	[-3.56, 4.77]	0.51 (2.30)	0.825	[-4.00, 5.02]	0.062
			Common	2	150 / 88	[-0.52 - 0.52]	0.16 (2.84)	[-5.41, 5.74]	0.03 (2.96)	0.992	[-5.78, 5.84]	0.016
			Separate	2	184 / 79	[-0.67 - 0.49]	0.81 (2.64)	[-4.36, 5.98]	0.77 (2.74)	0.778	[-4.60, 6.14]	0.082
Fuzzy*	MSE	Uniform	Common	1	153 / 90	[-0.54 - 0.54]	2.61 (2.58)	[-2.45, 7.66]	2.29 (2.97)	0.440	[-3.53, 8.11]	0.266
			Separate	1	230 / 73	[-0.74 - 0.44]	1.80 (2.55)	[-3.20, 6.80]	1.04 (2.97)	0.727	[-4.78, 6.86]	0.184
			Common	2	148 / 87	[-0.51 - 0.51]	-1.12 (4.05)	[-9.06, 6.81]	-2.42 (4.81)	0.615	[-11.85, 7.01]	-0.115
			Separate	2	272 / 77	[-1.02 - 0.47]	2.65 (3.41)	[-4.04, 9.34]	3.73 (3.75)	0.320	[-3.62, 11.08]	0.271
		Triangular	Common	1	153 / 90	[-0.55 - 0.55]	1.31 (2.60)	[-3.78, 6.40]	0.98 (2.97)	0.741	[-4.83, 6.80]	0.134
			Separate	1	230 / 76	[-0.74 - 0.47]	1.42 (2.52)	[-3.51, 6.36]	1.32 (2.87)	0.646	[-4.31, 6.95]	0.146
			Common	2	234 / 129	[-0.77 - 0.77]	0.65 (3.24)	[-5.71, 7.00]	0.04 (3.67)	0.991	[-7.15, 7.23]	0.066
			Separate	2	264 / 127	[-0.96 - 0.74]	0.91 (3.13)	[-5.22, 7.04]	0.64 (3.44)	0.854	[-6.11, 7.38]	0.093
	CER	Uniform	Common	1	114 / 68	[-0.39 - 0.39]	-0.23 (2.93)	[-5.97, 5.50]	-0.35 (3.13)	0.911	[-6.48, 5.78]	-0.024
			Separate	1	152 / 59	[-0.54 - 0.31]	2.13 (2.82)	[-3.40, 7.67]	1.78 (3.03)	0.557	[-4.15, 7.71]	0.218
			Common	2	102 / 65	[-0.35 - 0.35]	5.15 (5.88)	[-6.37, 16.67]	5.00 (6.19)	0.419	[-7.13, 17.13]	0.526
			Separate	2	218 / 65	[-0.70 - 0.33]	1.85 (3.73)	[-5.47, 9.16]	2.19 (3.87)	0.571	[-5.39, 9.78]	0.189
		Triangular	Common	1	114 / 68	[-0.39 - 0.39]	0.48 (2.87)	[-5.13, 6.10]	0.31 (3.09)	0.919	[-5.74, 6.37]	0.049
			Separate	1	152 / 65	[-0.54 - 0.34]	0.90 (2.72)	[-4.44, 6.24]	0.83 (2.93)	0.778	[-4.91, 6.57]	0.092
			Common	2	150 / 89	[-0.53 - 0.53]	0.19 (4.03)	[-7.70, 8.08]	-0.04 (4.24)	0.993	[-8.34, 8.27]	0.019
			Separate	2	183 / 87	[-0.66 - 0.51]	0.93 (3.66)	[-6.25, 8.11]	0.85 (3.81)	0.824	[-6.62, 8.31]	0.095

*See Appendix B, Exhibit B.5.2 for first-stage regression output related to the participation indicator for the General Reading Motivation outcome.

Figure 7 plots the raw data within the bandwidth, superimposing the local linear regression curves estimated using a triangular kernel.

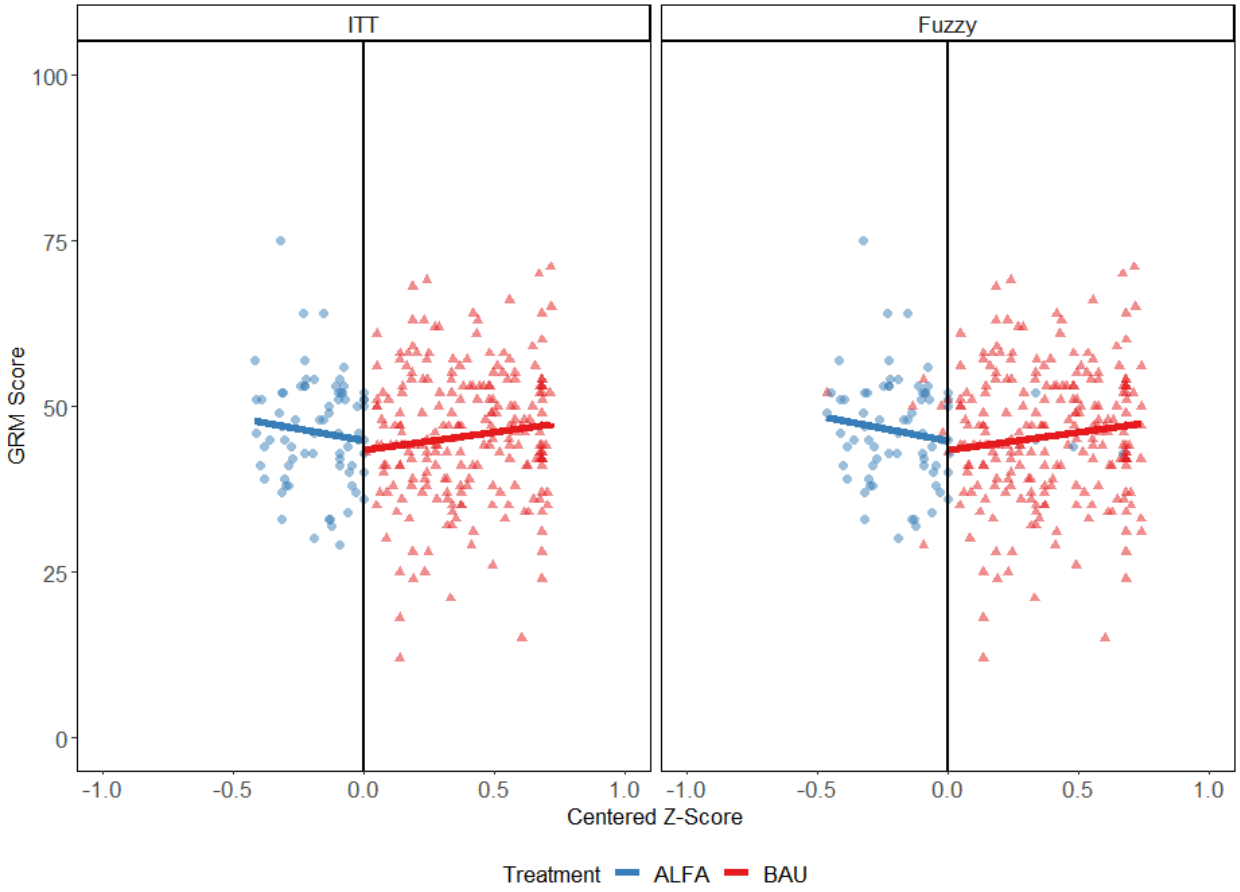


Figure 7. Estimated ITT and Fuzzy RD Effect Plots – GRM

Research Question 3 – Exploratory (ARMM Reading Frequency)

Summary of Findings

The evidence associated with the ARMM General Reading Frequency (FRQ) outcome suggests a descriptive advantage for the ALFA Lab group (i.e., estimates were positively associated with ALFA Lab group membership; see Figure 10 on p. 34), but inferentially this advantage was non-significant regardless of estimand (i.e., ITT versus complier average causal effect/Fuzzy RD) or estimation method. Attrition rates were again high and beyond the WWC threshold for concern. Assessment of the continuity of the relationship between the outcome and forcing variables lent support for continuity as none of the impact estimates at placebo cut-scores were statistically significant. All evidence related to the WWC RD standards can be seen in Appendix B.

Graphical Analysis

Figure 8 plots the standardized and centered running variable across all cohorts and sites along the x-axis and the FRQ scale score on the y-axis for students in the analytic sample (i.e., those with complete data on demographic background variables and cohort indicators). The lines are LOESS lines, constructed to find a curve of best fit without assuming a formal distributional shape. Examining the lines closely, the ALFA Lab group appears to intersect the dotted line at x=*c* at a point slightly above where the line representing the BAU group intersects the dotted line. Note that Figure 8 displays the data from a Fuzzy RD perspective, where the non-complying units (i.e., the students with assignment score values below the cut-score of zero that were members of the BAU group and vice versa) are presented according to the groups students experienced (i.e., a student with an assignment score below the cut-score of zero that did *not* experience ALFA Lab is considered part of BAU). An ITT perspective would present each student based on the intervention group they *should* have been in under full compliance. Those students shaded in red with assignment scores below zero tended to have relatively higher FRQ outcome scores, as shown by the upward curve of the red line to the left of the cut-score.

Figure 8. Fuzzy Experimental Group Assignment Scatterplot – ARMM FRQ

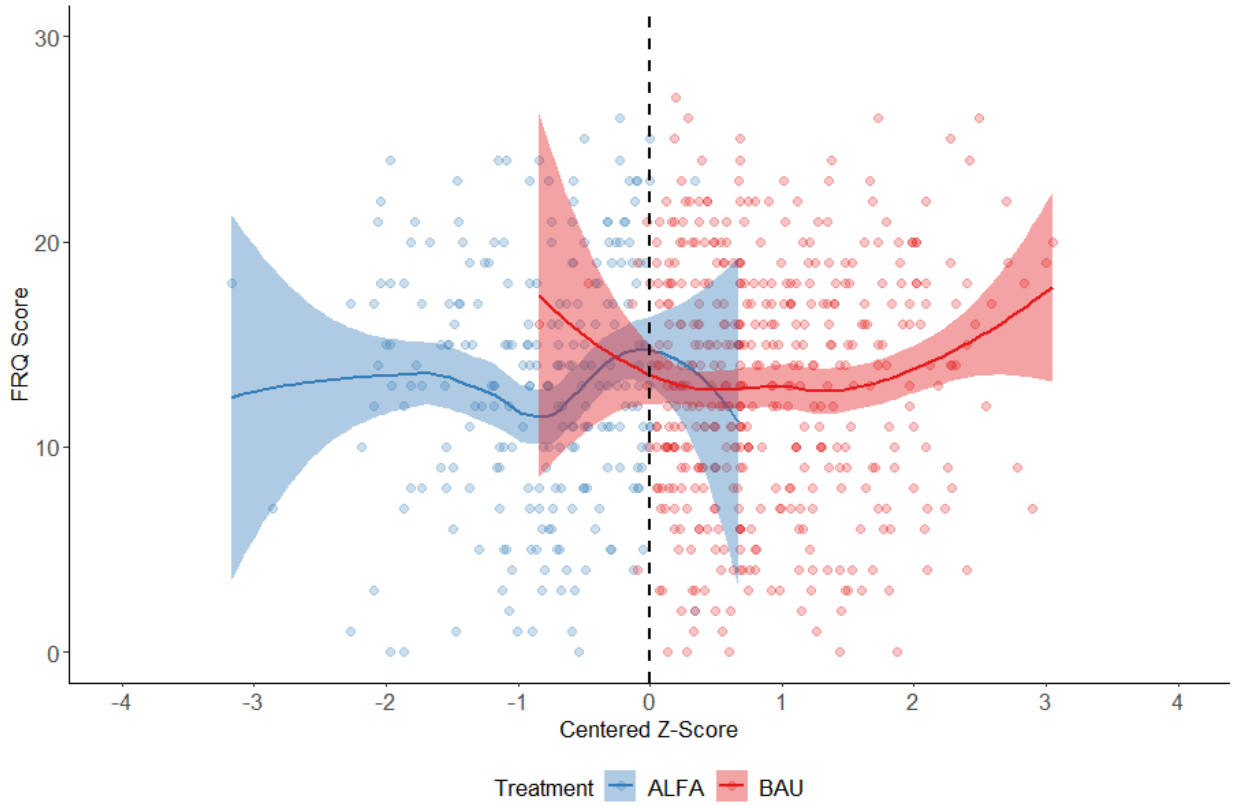


Table 11 displays the ES and QS spacing options, along with the IMSE and MV methods, to determine the number of bins to use in the smoothing process and plot for the Reading Frequency outcome. The MV method yielded a larger number of bins than the IMSE method and yielded bins that were, on average, smaller in length. Table 11 also displays the bins identified—by spacing style, kernel, and method. Figure 9 on p. 32 contains graphical representations of the four sets of triangular kernel bins from the bottom half of Table 11.

Table 11. Outcome Bin Identification Results, by Spacing and Method – ARMM FRQ

Bin Type	Kernel	Spacing	n bins (Left / Right)	Mean Length (Left / Right)	Median Length (Left / Right)
MV	Uniform	Equal	15 / 17	[0.21, 0.18]	[0.21 – 0.18]
		Quantile	17 / 17	[0.19, 0.18]	[0.10 – 0.13]
IMSE		Equal	6 / 4	[0.53, 0.76]	[0.53 – 0.76]
		Quantile	5 / 7	[0.63, 0.44]	[0.34 – 0.24]
MV	Triangular	Equal	15 / 17	[0.21, 0.18]	[0.21 – 0.18]
		Quantile	17 / 17	[0.19, 0.18]	[0.10 – 0.13]
IMSE		Equal	6 / 4	[0.53, 0.76]	[0.53 – 0.76]
		Quantile	5 / 7	[0.63, 0.44]	[0.34 – 0.24]

The ESMV binned plot with a fourth-order polynomial superimposed can be seen in the lower left quadrant of Figure 9, followed by a QSMV fourth-order binned plot in the lower right quadrant. The comparison of the ESMW and QSMV plots reveals the greater amount of variability among ALFA Lab data elements as the mean bin values (i.e., dots) shift between the two, whereas BAU mean bins are relatively stable. Moving to the IMSE plots designed to elucidate the overall regression shape, we see that between ES and QS the shape is relatively consistent, though again the ES version has more space between points among the ALFA Lab group. This aligns with what we see in the raw data scatterplot in Figure 8, with very few ALFA Lab units to the far left of the cut-score continuum. In all cases in Figure 9, there appears to be a small advantage for the ALFA Lab group (i.e., the ALFA Lab line is slightly higher at the cut point on the y-axis relative to the BAU group.)

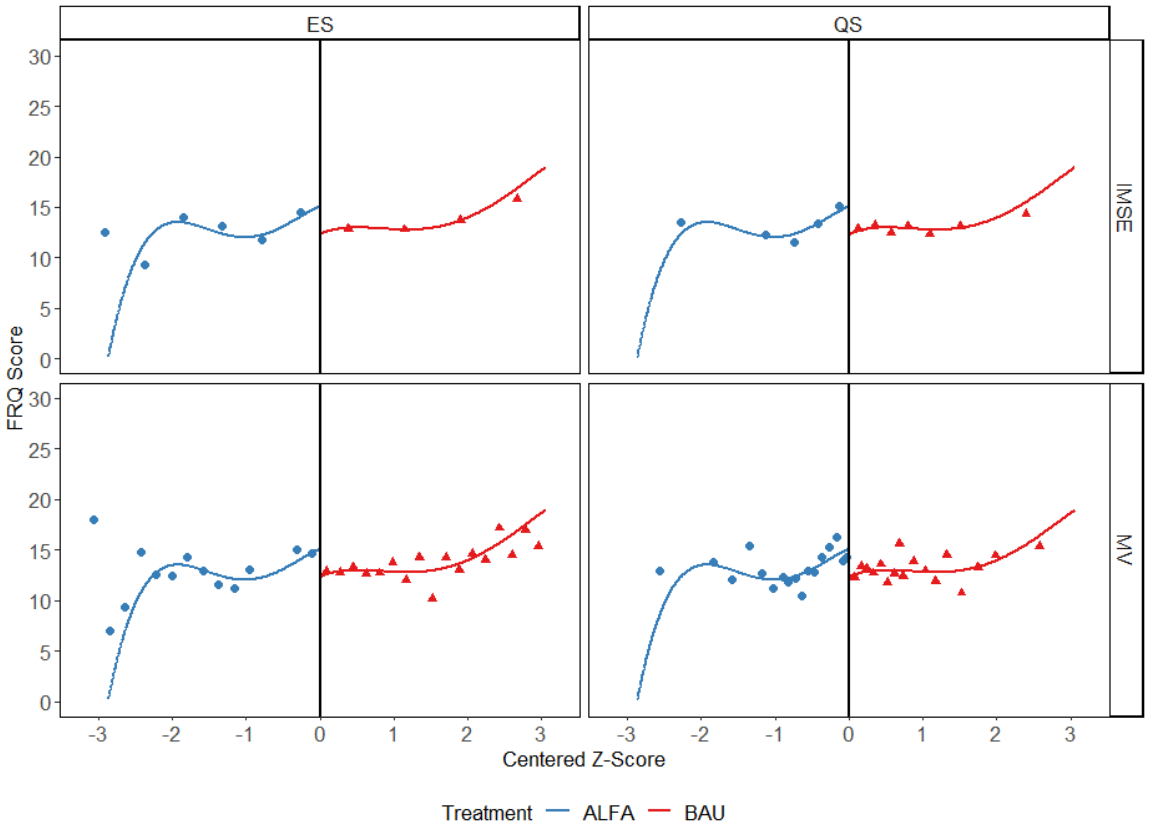


Figure 9. Smoothed Panel Plot (p=4) – FRQ RD Estimation Results

RD Estimation Results

Table 12 displays the ITT and Fuzzy design estimates obtained using ARMM FRQ as the outcome variable under the continuity-based framework. We present results using two different bandwidth types (MSE-optimal and CER-optimal), two weighting kernels, two bandwidth selections (Common, meaning the same bandwidth on each sit of the cutoff; Separate, meaning a different bandwidth on each side of the cutoff), and two polynomials ("1" is a linear specification, "2" is quadratic). We again prioritize models that use a triangular kernel (to give highest weight to units closest to the cutoff), separate bandwidths, and polynomials of Order 1. Models for both designs include the following additional covariates: student gender and dummy variables representing individual race categories with "did not respond" counting as its own category, ELL and SPED status as well as two dummies representing membership in Cohorts 1 and 2, respectively. We focus on the MSE-optimal point estimates yielded with the triangular kernel, though p-values from corresponding CER-optimal bandwidths are provided as well (see Appendix A for details). For both the ITT and Fuzzy designs, all estimates were positive in favor of ALFA Lab, though none were statistically significant. Some of the favored triangular-weighted linear specification models with separate bandwidths yielded estimates that were approaching significance.

Table 12. Continuity-Based (CB) Estimates – FRQ

Design	BW Type	Kernel	BW Select	p	n (TX / CT)	Bandwidth	Est (se)	CI	Robust Est (se)	Robust p-value	Robust CI	ES
ITT	MSE	Uniform	Common	1	149 / 89	[-0.54 - 0.54]	2.54 (1.45)	[-0.29, 5.37]	2.67 (1.68)	0.112	[-0.62, 5.97]	0.430
			Separate	1	170 / 97	[-0.62 - 0.59]	2.43 (1.39)	[-0.30, 5.16]	2.37 (1.65)	0.150	[-0.86, 5.59]	0.412
			Common	2	216 / 120	[-0.71 - 0.71]	2.38 (1.88)	[-1.30, 6.07]	2.44 (2.14)	0.254	[-1.76, 6.64]	0.403
			Separate	2	256 / 122	[-0.95 - 0.72]	1.61 (1.75)	[-1.83, 5.05]	1.24 (1.97)	0.530	[-2.63, 5.10]	0.273
		Triangular	Common	1	222 / 123	[-0.74 - 0.74]	2.51 (1.31)	[-0.07, 5.08]	2.49 (1.55)	0.109	[-0.55, 5.53]	0.424
			Separate	1	225 / 137	[-0.74 - 0.82]	2.51 (1.28)	[-0.01, 5.02]	2.50 (1.52)	0.100	[-0.48, 5.49]	0.425
			Common	2	231 / 133	[-0.78 - 0.78]	2.34 (1.87)	[-1.32, 6.01]	2.45 (2.13)	0.251	[-1.73, 6.63]	0.397
			Separate	2	249 / 133	[-0.92 - 0.78]	2.16 (1.82)	[-1.41, 5.73]	1.95 (2.04)	0.339	[-2.05, 5.96]	0.366
	CER	Uniform	Common	1	110 / 67	[-0.39 - 0.39]	1.96 (1.67)	[-1.30, 5.23]	2.06 (1.79)	0.249	[-1.44, 5.56]	0.332
			Separate	1	124 / 72	[-0.44 - 0.42]	2.59 (1.58)	[-0.51, 5.69]	2.61 (1.70)	0.126	[-0.73, 5.95]	0.439
			Common	2	135 / 78	[-0.49 - 0.49]	1.38 (2.31)	[-3.15, 5.92]	1.45 (2.40)	0.545	[-3.26, 6.16]	0.234
			Separate	2	177 / 79	[-0.66 - 0.50]	1.96 (2.09)	[-2.14, 6.06]	1.85 (2.16)	0.393	[-2.39, 6.09]	0.332
		Triangular	Common	1	148 / 88	[-0.54 - 0.54]	2.34 (1.51)	[-0.63, 5.31]	2.34 (1.65)	0.157	[-0.90, 5.58]	0.397
			Separate	1	148 / 97	[-0.54 - 0.59]	2.53 (1.48)	[-0.37, 5.43]	2.54 (1.61)	0.115	[-0.62, 5.70]	0.429
			Common	2	148 / 88	[-0.54 - 0.54]	2.62 (2.22)	[-1.73, 6.96]	2.69 (2.32)	0.246	[-1.85, 7.23]	0.443
			Separate	2	173 / 89	[-0.63 - 0.54]	2.44 (2.14)	[-1.75, 6.62]	2.40 (2.22)	0.280	[-1.96, 6.75]	0.413
Fuzzy*	MSE	Uniform	Common	1	165 / 104	[-0.60 - 0.60]	3.45 (1.70)	[0.11, 6.79]	4.04 (1.91)	0.034	[0.30, 7.77]	0.584
			Separate	1	170 / 104	[-0.62 - 0.59]	3.20 (1.70)	[-0.13, 6.52]	3.47 (1.98)	0.080	[-0.42, 7.36]	0.541
			Common	2	187 / 112	[-0.68 - 0.68]	2.49 (2.62)	[-2.64, 7.62]	2.15 (3.02)	0.477	[-3.78, 8.08]	0.422
			Separate	2	263 / 104	[-0.99 - 0.61]	2.46 (2.51)	[-2.45, 7.38]	2.08 (2.79)	0.457	[-3.39, 7.54]	0.417
		Triangular	Common	1	268 / 161	[-1.03 - 1.03]	3.06 (1.45)	[0.21, 5.90]	3.55 (1.67)	0.033	[0.29, 6.82]	0.517
			Separate	1	242 / 127	[-0.86 - 0.75]	3.03 (1.62)	[-0.15, 6.21]	3.07 (1.91)	0.107	[-0.66, 6.80]	0.513
			Common	2	238 / 137	[-0.82 - 0.82]	3.17 (2.46)	[-1.66, 8.00]	3.37 (2.84)	0.235	[-2.19, 8.93]	0.537
			Separate	2	259 / 127	[-0.97 - 0.75]	2.66 (2.45)	[-2.14, 7.46]	2.44 (2.73)	0.372	[-2.91, 7.78]	0.451
	CER	Uniform	Common	1	120 / 72	[-0.43 - 0.43]	2.92 (2.05)	[-1.10, 6.94]	3.22 (2.14)	0.133	[-0.98, 7.42]	0.495
			Separate	1	124 / 72	[-0.44 - 0.43]	3.31 (2.06)	[-0.72, 7.34]	3.47 (2.19)	0.113	[-0.82, 7.77]	0.561
			Common	2	127 / 75	[-0.47 - 0.47]	2.60 (3.55)	[-4.35, 9.55]	2.46 (3.69)	0.506	[-4.79, 9.70]	0.440
			Separate	2	211 / 71	[-0.68 - 0.42]	2.33 (2.95)	[-3.45, 8.11]	2.22 (3.05)	0.467	[-3.76, 8.19]	0.395
		Triangular	Common	1	225 / 126	[-0.74 - 0.74]	3.10 (1.64)	[-0.12, 6.32]	3.39 (1.78)	0.057	[-0.10, 6.89]	0.525
			Separate	1	170 / 89	[-0.62 - 0.54]	2.87 (1.88)	[-0.82, 6.55]	2.89 (2.04)	0.156	[-1.10, 6.88]	0.485
			Common	2	156 / 90	[-0.56 - 0.56]	3.68 (3.16)	[-2.51, 9.88]	3.82 (3.32)	0.251	[-2.70, 10.33]	0.624
			Separate	2	179 / 86	[-0.67 - 0.52]	3.55 (3.04)	[-2.40, 9.50]	3.50 (3.15)	0.266	[-2.66, 9.67]	0.602

*See Exhibit B5.3 on p. B-14 of Appendix B for first-stage regression output related to the participation indicator.

Figure 10 plots the raw data within the bandwidth, superimposing the local linear regression curves estimated using a triangular kernel.

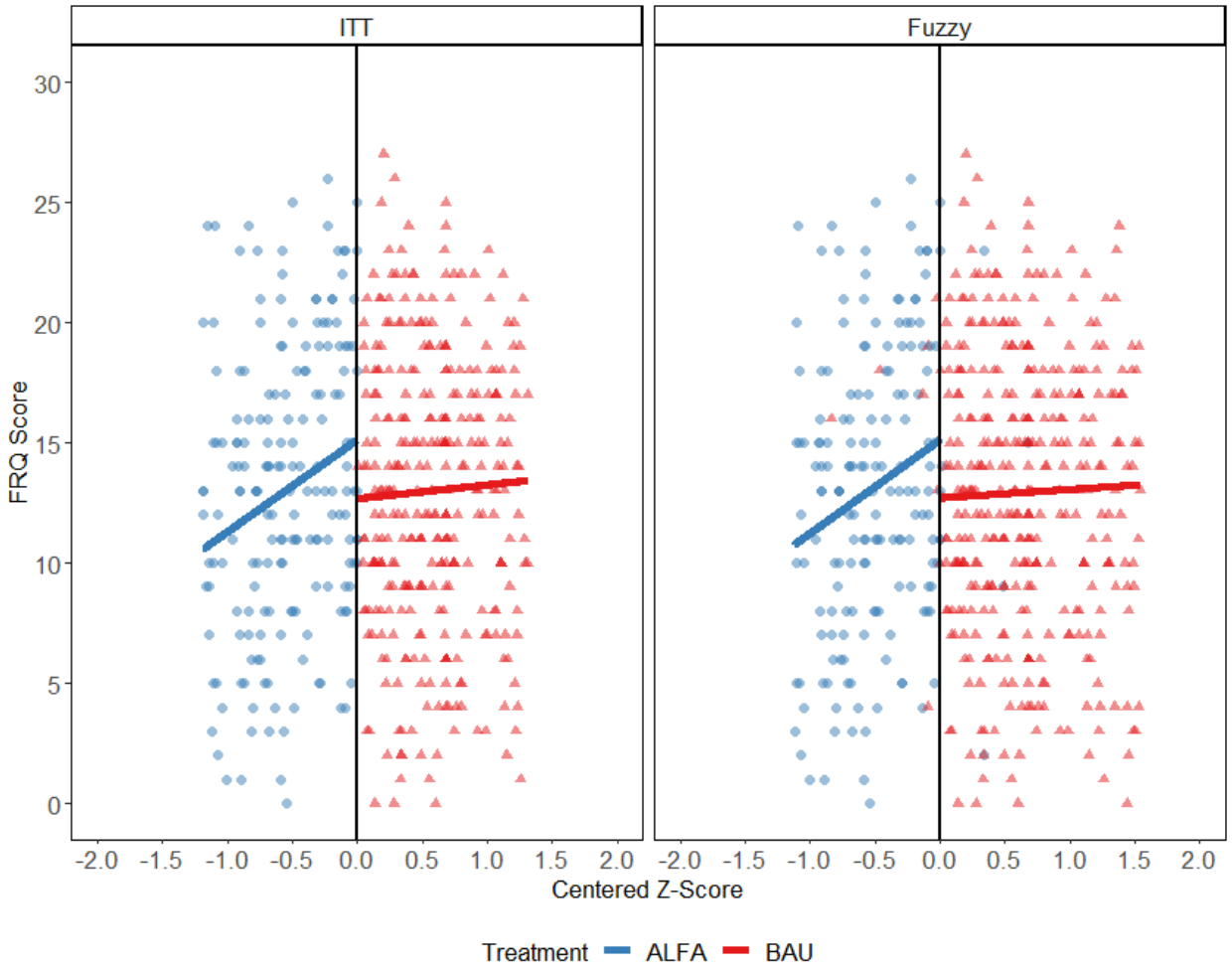


Figure 10. Estimated ITT and Fuzzy RD Effect Plots – FRQ

Research Question 4 – Exploratory (Implementation Fidelity)

Summary of Findings

Across the three cohorts, participating high schools made considerable progress implementing the ALFA Lab program according to data collected by the ICF research team. The majority of teachers from all cohorts reported in the teacher survey that they were implementing ALFA Lab components—such as Daily Launch, Main Station, Wordology, Collaboration Station, and Media Madness—every class meeting or nearly every class meeting (Table D9, Appendix D). In Cohorts 2 and 3, survey respondents noted that sometimes a station would take longer than the amount of time allotted and would therefore be split

across multiple days. Overall, survey respondents reported the ALFA Lab components were Helpful to Very Helpful (Table D21, Appendix D).

Teachers/instructional assistants who responded to the teacher survey reported they were Satisfied to Strongly Satisfied with the teacher's handbook and unit materials, the instructional/support materials, and the summer training (Table D23, Appendix D). Further, based on classroom observations from Cohort 1 and Cohort 3 conducted by ICF researchers, all observed ALFA Lab classes demonstrated evidence of implementing most (66%–89%) of the ALFA Lab components included in the CFAI (Table F1, Appendix F) and similar PLATO scores across participating schools, except for one school in Cohort 3 that had slightly lower scores (Tables E1–E4, Appendix E).

Implementation Fidelity

Fidelity of implementation (FOI) was calculated to better understand the fidelity and quality of program implementation and whether the program was implemented in treatment classrooms as intended by the ALFA Lab program. Across cohorts, the ICF research team gathered data for a FOI score from a multitude of sources, including teacher surveys, classroom observations, and interviews with ALFA Lab teachers. A composite FOI score was calculated (Table 14, p. 36) involving key components surrounding instruction, responsive instruction, student collaboration, and themes that the ALFA Lab is based around, along with a moderator score (Table 13, p. 36) comprising general use of ALFA Lab materials, student-teacher ratio, classroom management, school/district support, and adequate time. Moderators are factors believed to impact key components. FOI was not ascertained for cohort 2 due to disruptions from the COVID-19 pandemic.

In Cohort 1, moderator scores were comparable across both schools. School 1 (NY) had a moderator score of 3.03 with low ratings regarding General Use of Materials (2.0) and School/District Support (2.0). School 3 had a moderator score of 3.1 with scores above 3.0 on all moderators except School/District Support (0.0). Specific to Cohort 3, School 4 (AL) had the lowest rating (2.0) across schools on Classroom Management, followed by School 2 (CT) (2.5). Three out of four Cohort 3 schools had low ratings on School/District Support: School 3 (CA) (0.0), School 1 (NY) (2.0), and School 2 (CT) (2.0). School 4 (AL) had the highest rating on General Use of Materials (3.3), Student/Teacher Ratio (4.0), and School/District Support (4.0); thus, having the highest moderator score among Cohort 3 schools. Across both cohorts, among all moderators, School/District Support had generally low scores, although this does not impact the overall Moderator Score given their relatively smaller weight of 0.1.

Table 13. Moderator Scores Across Cohorts

Cohort	Site	General Use of Materials (.25)	Classroom Management (.25)	Student/ Teacher Ratio (.20)	Adequate Time (.20)	School/ District Support (.10)	Moderator Score
Cohort 1	School 1 (NY)	2.000	3.556	3.545	3.667	2.000	3.031
	School 3 (CA)	3.250	3.273	3.500	3.833	0.000	3.097
Cohort 3	School 1 (NY)	3.200	3.333	3.000	3.667	2.000	3.167
	School 2 (CT)	3.200	2.500	3.556	4.000	2.000	3.136
	School 3 (CA)	3.250	3.636	3.500	4.000	0.000	3.222
	School 4 (AL)	3.333	2.000	4.000	3.667	4.000	3.267

Table 14 demonstrates the calculated FOI score across each site in both cohorts, incorporating weights for each component, including the moderator scores calculated in Table 13. In Cohort 1, School 1 (NY) and School 3 (CA) had scores above 3.0 for all key components, having similar FOI scores indicating high fidelity. In Cohort 3, School 4 (AL) had the lowest rating for three of the four key components: Instruction (2.0), Responsive Instruction (2.4), and Student Collaboration (3.2). School 2 (CT) had low ratings on the themes and projects the ALFA Lab is Based Around (1.9) and Instruction (2.8). Both School 2 (CT) and School 4 (AL) had overall FOI Scores of 2.98 and 2.59, respectively, indicating moderate fidelity. School 1 (NY) and School 3 (CA) had scores above 3.0 for all key components in Cohort 3 as well, having similar FOI scores indicating high fidelity.

Table 14. FOI Scores Across Cohorts

	Site	Instr.* (.25)	Responsive Instr. (.25)	Student Collaboration (.20)	Lab Based Around (.15)	Moderator Score (.15)	FOI Score	FOI Label
Cohort 1	School 1 (NY)	3.619	3.833	3.200	3.167	3.031	3.433	High
	School 3 (CA)	3.833	4.000	4.000	3.333	3.097	3.723	High
Cohort 3	School 1 (NY)	3.238	3.840	3.733	3.000	3.167	3.441	High
	School 2 (CT)	2.800	3.556	3.200	1.875	3.136	2.981	Moderate
	School 3 (CA)	3.167	4.000	3.429	3.000	3.222	3.411	High
	School 4 (AL)	2.000	2.353	3.200	2.500	3.267	2.593	Moderate

* “Instr.” is short for instruction.

Implementation Context for ALFA Lab

Four high schools participated in implementing the ALFA Lab program across the three cohorts. Table 15 outlines the high schools participating in each cohort and respective academic year. The profiles of these schools are described in Appendix E.

Table 15. Participating High Schools by Cohort

	Year 1 (2019–2020)	Year 2 (2020–2021)	Year 3 (2021–2022)
School 3 (CA)	X	X	X
School 1 (NY)	X	X	X
School 4 (AL)			X
School 2 (CT)			X

Literacy Context

Across all three of the cohorts, the participating high schools described varying literacy contexts present in their schools. In Cohort 1, when administrators and teachers (ELA and ALFA Lab teachers) were asked to evaluate the literacy of students in their schools, there was a consensus across participants that students were disengaged from reading in general and did not engage in it outside of school requirements. In cohort 2, the COVID-19 pandemic resulted in participating schools transitioning to remote learning and exacerbated existing challenges with student engagement and participation. Lastly, in Cohort 3, administrators and teachers (both ELA and ALFA teachers) continued to note that students seemed to be generally disengaged with reading and writing, as they were prior to the COVID-19 pandemic.

When evaluating the impact of COVID-19 on student literacy levels, Cohort 3 site visit participants agreed that the pandemic exacerbated gaps in the literacy skills of students. Participants observed that some students were behind grade levels, which they attributed to insufficient opportunities to practice reading.

Other Literacy Programs

For many of the participating schools, the ALFA Lab program was not the only literacy program being implemented. Other programs included Reading XP, Edmentum, Strategic Reading, Read 190, Language!, and Saturday Academy. In Cohort 1, even though each of the schools implemented different reading intervention programs besides ALFA Lab, teachers noted similar literacy strategies used across the programs, such as a variety of text annotation strategies to improve reading comprehension. In Cohort 2, one participating school described that due to the COVID-19 pandemic, they lost access to a tutoring program that pulled in local college students as tutors. Additionally, in all three cohorts, outside of external literacy programs, ELA teachers across the schools described incorporating various strategies to foster literacy skills, including annotation, vocabulary building, and strategic reading.

Staffing

There was great variation among the teaching staff charged with implementing the ALFA Lab program at the participating schools, in terms of years of experience teaching and experience implementing reading

intervention programs. In Cohort 1, ALFA Lab and ELA teachers had a range of first-year teachers to those with over 20 years of experience. At two of the three schools, more experienced teachers were assigned to lead the ALFA Lab classes, while at the third a first-year teacher was assigned. Variations in staff experience may have influenced the implementation of the ALFA Lab program. In Cohort 2, unlike the first year, not all of the ALFA Lab classes had instructional assistants to help implement the program, which was partially the result of challenges due to the COVID-19 pandemic. In Cohort 3, variations in ALFA Lab teacher experience and the lack of an instructional assistant continued to be potential factors in implementation differences across the schools.

School Structures

Implementation of the ALFA Lab program may have been influenced by differing structures across schools. ALFA Lab classes were developed to be used in a variety of bell schedules, including a traditional bell schedule with seven 45–55-minute periods each day or an extended bell schedule with three to four 70–90-minute extended blocks. Table 16 describes the variation in implementation of the ALFA Lab program across the participating schools for each cohort.

Table 16. ALFA Lab Program Structure across Cohorts

Cohort	Program Start Time	Length of Class Period	Average Number of ALFA Lab Sections	Average Number of Students per Section
Year 1 (2019–20)	One was delayed	2 of 3 were 70–90 mins.	5	18
Year 2 (2020–21)	All began with school year	70–90 mins and 37 mins.	4	19
Year 3 (2021–22)	All began with school year	3 of 4 were 45–55 mins.	3	22

Note. Table reports aggregated teacher survey responses across participating schools for each cohort.

In Cohort 2, due to the COVID-19 pandemic, the two participating schools both made significant adjustments to their instructional schedule to protect student health. One of the participating schools offered virtual synchronous classes in shortened 37-minute blocks, which were followed by an "extended hours" block for students to meet in a teacher's "Zoom room," one-on-one, or in small groups. The point of the extended hours was for students to ask questions; receive extra help; and/or to complete activities, readings, and assignments with the teacher available to assist if they virtually raised their hand. Alternatively, the other participating school utilized a hybrid model splitting students into two cohorts, with ninth graders (and 12th graders) attending in-person learning on Mondays and Thursdays and the other cohort (10th and 11th graders) attending in-person learning on Tuesdays and Fridays. All students learned remotely on Wednesdays, and on the two days that their cohort was not in person. Of course, ALFA Lab classes and control classes followed the ninth-grade schedule.

In Cohort 3, all schools resumed in-person learning (either immediately or eventually) and normal (or new normal) implementation of the ALFA Lab program. An example of a new normal practice in some schools was to use fillable e-form versions of the curriculum's worksheets for station activities. These forms were first introduced during the pandemic. Students opened them on their laptops, filled them in during the appropriate points of the activity, and then submitted them electronically to the teacher.

Perception of ALFA Lab Implementation

This section provides an overview of stakeholder perceptions of program implementation, organized in terms of prominent program successes and key challenges across the three cohorts. Findings reported in the section include perceptions shared in focus groups with ALFA Lab teachers/instructional assistants and school administrations as well as survey responses from ALFA Lab teachers/instructional assistants from the three cohorts.

Program Implementation Successes

ALFA Lab teachers and administrators identified key areas in which ALFA Lab implementation was successful across the implementation of the three cohorts—organized program structure and strong support from JHU.

Organized Program Structure

Across the interviews and focus groups completed with each participating cohort, participants shared implementation successes facilitated by the ALFA Lab program structure. The first cohort described the program structure as organized, well-defined, and using a clear implementation framework. The second cohort emphasized how the program accurately and successfully matched the literacy rates of students in the program at the beginning of the course and grew in difficulty with the students through the implementation. ALFA Lab teacher participants said that using grade-appropriate content boosted students' self-esteem and improved their literacy capabilities. In Cohort 3, as schools transitioned back to in-person learning following the COVID-19 pandemic, an administrator shared that they believed the smaller class sizes in the ALFA Lab program helped students adapt back into in-person learning.

Additionally, the group learning facilitated through the stations continued to be noted as a success in program implementation across cohorts. In Cohort 1, participants indicated that station teaching (with the support of an instructional assistant) provides structure in the classroom for teachers, a variety of activities, and the opportunity for more differentiated instruction for students. Even amidst the COVID-19 pandemic, Cohort 2 participants noted that the use of group work, through station learning, had benefited students academically and personally as it not only boosted students' self-esteem, but it also set them up for success in their first year in high school. Lastly, Cohort 3 teacher participants echoed the administrator in describing how they believed the program structure allowed students to adapt quickly back to in-person learning and promoted collaboration between students through the station group work.

Strong JHU Support

Across the cohorts several participants described the strong partnerships they had developed with JHU staff during program implementation, which provided necessary support and resources throughout the program. In the first cohort, participants noted that the materials and curriculum JHU provided were of very high quality, organized, and effective in their delivery means. Participants noted the staff, specifically their JHU liaisons, were supportive and willing to answer any questions they had about the program. They also expressed that JHU staff were very supportive in modifying the program as needed. In Cohort 2, the strong relationships with JHU continued as participants shared that JHU satisfactorily supported the schools not only during the second year of program implementation, but also in the transition to online learning due to the COVID-19 pandemic. ALFA Lab teachers reported JHU staff

assisted in the transition to virtual learning by providing online resources for ALFA Lab teachers and converting materials into online resources.

Key Challenges

ALFA Lab teachers and administrators identified key challenges faced during ALFA Lab implementation in each of the three cohorts, including the fixed program structure, inadequate time for activities, and challenges related to the COVID-19 pandemic. There were a couple of other unique challenges cited by teachers/administrators in each of the cohorts. This section describes each of the challenges in greater detail.

Fixed program structure

Although the program structure was lauded as an implementation strength by staff across cohorts, some perceived the prescribed program structure as being rigid. Student or even teacher absences could derail the curriculum pacing, particularly in an extended-block bell schedule, as teachers had to cover a specific portion in each class. ALFA Lab staff across the three cohorts noted the need to allow for flexibility within the program structure. An unintended consequence of the program structure's fixed nature was that as students became familiar with the program, as one staff member noted, they did not feel the need to engage with the program materials at a deeper level; that is, students did not engage in critical thinking. Furthermore, the fixed program structure was a challenge when transitioning to virtual learning as a result of the pandemic.

Inadequate Time for Program Activities

A prevalent theme reported by staff across all three cohorts included insufficient time to complete the activities within the stations as a challenge. One reason for this was the prescribed curriculum pacing, which some staff noted as being fixed in timelines and without flexibility. Because of the fixed pacing guidelines, there was no additional time to offer supplemental activities to broaden student understanding and engagement of the topic. Moreover, the time scripted within the ALFA Lab program did not translate to the time ALFA Lab classes actually had on the ground. One reason for this was the time it took to set up resources and stations at the beginning of class and, in some cases, students needed additional teaching of pretopics before the start of the lesson. Staff in Cohort 1 reported using KWL charts,[6] providing additional guiding questions and additional "write-a-letter" guidance as tools for preteaching. Staff observed that the pacing for ALFA Lab projects seemed rushed as a result.

ELA Instruction by PLATO

During observations of classrooms in Cohorts 1 and 3, the ICF evaluation team noted some key challenges in using the PLATO rubric where ALFA Lab classes did not, overall, score higher than comparison classes for the instructional elements that were scored—Strategy Use & Instruction, Feedback, Classroom Discourse, and Text-Based Instruction (See Tables E1-E4, Appendix E). It was expected that using the ALFA Lab curriculum would allow an ALFA Lab teacher to score higher than the comparison teachers on PLATO instructional elements. Because the comparison ELA classes did not

[6] A KWL chart breaks concepts into three categories: What I *Know* (K), What I *Want* to Know (W), and What I *Learned* (L).

contain struggling readers, it might have been appropriate for more advanced instructional strategies than ALFA Lab teachers. Moreover, PLATO scores were gathered in two 20-minute observations during single-day site visits, and hence were based on a single snapshot of classroom contexts and practices that were contingent on the objectives of the lessons taught that day and may not reflect the overall instructional practices used in each class.

COVID-19 Related Challenges

The COVID-19 pandemic posed a major challenge for ALFA Lab implementation for all schools across cohorts, particularly for cohorts 2 and 3. Additionally, School 3 in Cohort 1 was adversely affected by the pandemic because their implementation of ALFA Lab included some sections in the 2nd semester of 2019-2020 when school closed abruptly in mid-March. COVID-19-related challenges in later cohorts primarily centered on translating ALFA Lab materials and processes into a virtual setting and student disengagement as a function of transitioning to online learning and, later, dealing with the continuing impact of the pandemic after the return to in-person instruction.

Implementing ALFA Lab in a virtual setting. ALFA Lab staff from both sites in Cohort 2 reported on the challenges of transitioning ALFA Lab programming to online synchronous formats, particularly with station learning, where students work in multiple groups and move through ALFA stations. They said this was difficult to manage and implement remotely. One example noted by an administrator is that while online platforms can have separate breakout rooms for station learning, it becomes challenging for teachers to monitor station learning in separate online spaces. Staff from both sites in Cohort 2 noted that they completed the Main Station together on their respective online platforms, before splitting up into breakout rooms or moving to the next activities. ALFA Lab activities also had to be moved to online formats, such as through digital worksheets, interactive slides, and online discussions through chat messaging tools and verbal discussion on the online platform. As much as Cohort 2 staff endeavored to maintain the prescribed ALFA Lab program structure, they found that the station learning structure did not allow for much flexibility—especially for teachers in an online setting—and was not conducive to collaboration between students in a virtual environment.

Lack of student engagement and participation. Cohort 2 staff expressed concern about the lack of student engagement and participation because of the pandemic. Staff attributed this to the shift to online learning. With online learning, students are expected to be able to work independently as they turn on their computers, join online classes, and complete their learning activities remotely. In the case of ALFA Lab classes, staff noted that they faced challenges getting students to turn on their cameras and participate in group discussions. Because students would not turn on their cameras, staff found it challenging to monitor students' learning. Staff suspected that students faced significant distractions during the school day especially since students may not be monitored by their parents during class time and hence may have engaged in distractions such as video games and social media. Staff also noted that during the pandemic, some parents relied on their older children to help out around the house during class time, including housework, caring for siblings, or going on errands. This low engagement and participation manifested in students' written work as they sometimes did not turn in their assignments.

Moreover, Cohort 2 staff also expressed concern about their students' social-emotional well-being during the COVID-19 pandemic, when students experienced socialization and anxiety-related issues. While non-ALFA Lab staff could pause instruction to check in on their students' well-being, ALFA Lab staff reported that this flexibility was not afforded by the program structure. They noted that they

would have appreciated having the flexibility to pause academic instruction to provide opportunities and activities for their students' social-emotional health.

Continuing effects of the COVID-19 pandemic. Staff from three of the four Cohort 3 schools noted that they continued to grapple with the effects of the pandemic, which posed a challenge to ALFA Lab implementation. School 1 (NY) started the school year using a hybrid instructional mode and then returned to in-person instruction a couple of months into the fall semester. Moreover, because COVID-19 protocols were still in place, minor adjustments needed to be made to reduce the risk of disease transmission, such as instead of students moving between stations, the materials for the next station were brought to them. As a result, staff had to spend a considerable amount of time preparing the classroom to accommodate these restrictions. Furthermore, staff from two Cohort 3 schools shared that student disengagement persisted as a challenge especially since they observed students experiencing issues transitioning back to in-person instruction.

Other Challenges

The challenges described in this subsection were unique to specific schools or to a specific cohort and it is worth noting those experiences when considering perceptions of program implementation.

Quality of materials. An ALFA Lab teacher from a Cohort 3 school noted that the materials used in the program appeared juvenile in their presentation, particularly in the images and graphics used. This was perceived as an implementation challenge because the use of juvenile graphics reminded students that they were academically behind, which may have influenced behavioral issues among students who felt the material was condescending.

Station rotation order. ALFA Lab staff from two of the four Cohort 3 schools shared that the station rotation order was a challenge: Students not being taught the necessary vocabulary prior to their rotation impacted learning at subsequent stations. This was specific to the Wordology Station where the vocabulary knowledge needed to complete activities was contingent on reading selections from the Main and Collaboration stations. Not having the required vocabulary accessible for each station made it challenging for ALFA Lab teachers, if they did not have a lab assistant, because the teacher could not assist students in the satellite stations at the same time they were leading Main Station. This was especially a concern for the later units, which were deemed to have more difficult vocabulary. Staff from one ALFA Lab school also found that similar questions and issues arose across the four stations at various points, which they felt could have been addressed more efficiently with the whole group. Although this was not reported as a challenge prior to the pandemic, staff theorized that the difficulty with the station rotation order may have been a side effect of the COVID-19 academic decline.

Lack of summative assessments. In Cohort 3, several participating ALFA Lab teachers shared challenges related to the assessment of student growth and achievement. One ALFA Lab teacher said that while there are formative assessments, there are no summative assessments at the end of each unit. Another ALFA Lab teacher shared similar concerns regarding limited data collection throughout the semester. The teacher noted that without assessments throughout, there was no way for the teacher to assess student progress or understanding.

Research Question 5 – Exploratory (Cost Analysis)

Summary of Findings

The average cost of a one-semester ALFA Lab is $29,731, with the cost per student at $1,351. The total costs of operating an ALFA Lab program varied largely across the four participating schools, ranging from $22,684 to $90,776. The primary factors driving the difference were the number of sections and number of participating students in each school. The more sections a school has with more participating students, the higher the total cost with similar per student cost around $1,200.

Below is a discussion of the costs for each ingredient and the overall cost of the program reported in Exhibit 5. All cost assumptions are based on information from Cohort 3 to better reflect the post-COVID reality.

Personnel

The primary personnel involved in the program include teachers, lab assistants, and school principals. A total of four teachers and two lab assistants were reported as the key personnel in the program, with an average of one teacher and one lab assistant per lab. They spent about 15 hours per week on the program, focusing on teaching and preparing lessons. In addition, the teachers and lab assistants needed to participate in trainings during the summer and the academic year. The training lasted about 10 hours in summer and 7 hours during the academic year (e.g., observations, discussions, professional development for different units in the curriculum). For program administration, a principal reported spending about 10 to 20 hours per year.

Teacher/Lab assistant. Based on the median wage reported by the U.S. Department of Labor's *Occupational Outlook Handbook* (U.S. Bureau of Labor Statistics, 2022), as of 2021, high school teachers had a median salary of $61,820. The number of working hours for teachers in the K–12 academic year is 1,440 (36 weeks, 5 days a week, 8 hours a day), which leads to about $43 for the hourly rate of a teacher. Therefore, the cost of a teacher/lab assistants working 15 hours a week for 13 weeks, plus training hours (15 hours), is about $9,030.

Principal. According to the median wage reported by the U.S. Department of Labor's *Occupational Outlook Handbook*, the median salary for a high school principal was $98,420. The number of working hours for principals is 2,080, which comes out to about $47 for the hourly rate of a principal. The semiannual cost for a principal to administer the ALFA Lab program (10 hours) is about $470.

Training staff. According to the ALFA Lab program, two training staff provided virtual training to teachers and lab assistants during the summer prior to the start of the ALFA Lab in 2021–2022. On average, the summer training lasted about 10 hours, ranging from 6 to 12 hours across schools. During the academic year, the training staff provided virtual trainings for 7 hours. The hourly rate of the trainer was $53. So, the cost of training staff is $901.

Facilities

The facilities used are regular classrooms.

***Classrooms*.** The ALFA Lab programs reported using a regular classroom for about 7 hours per week for teaching lessons, which adds up to 91 hours in 13 weeks, about 6.3% of the instruction year (1,440 hours). The regular classroom was assumed to be 700 square feet, according CostOut (Hollands et al., 2020), at a cost-per-year of $16,235. The classroom was reported to be used about 6.3% of the time during the year. Therefore, the unit cost of using one classroom for one school semester for an ALFA Lab costs about $1,023.

Equipment

Itemized equipment used in ALFA Lab programs includes computers/laptops/electronic tablets and printers.

***Computers/Electronic tablets*.** According to the teacher survey, all students were required to use a computer, laptop, or electronic tablet when participating in the ALFA Lab. A new computer or electronic tablet costs between $400 and $600. Assuming the equipment will only be used in the ALFA Lab program with an expected life of 3 years, a computer/electronic tablet costs about $84 per person, per semester. Estimating about 10% of the use is attributed to the program,[7] the total cost of computers/electronic tablets for one ALFA Lab (including two teachers and 22 students) is $192 per semester.

Printer. We assume 10% of one school printer was used for ALFA Lab. A printer costs on average $300. With the expected life of 3 years, the cost of using a printer for ALFA Lab (10%) is about $5 per semester.

Materials

***ALFA Lab's lesson materials*.** The ALFA Lab provides three units of lesson materials: Feisty Felines, Heroes, and Galaxy. Each lab teacher/assistant will need one manual for each unit and each class section will need a set of the consumable and nonconsumable materials for the unit. Costs of lesson materials are listed in Exhibit 14 below based on a report from the program developer.

Other Inputs

Additional inputs may include student field trips and school supplies.

Student field trip. Some schools reported student field trips for the ALFA Lab program. The cost of one student field trip is hard to estimate due to the variation of location, activities, transportation, tickets/admission, and other associated costs. According to an article in *Education Week*, a school district budgeted $360,000 for its 15,500 K–12 students participating in at least one field trip (Mongeau, 2015), which was about $23 per student. We use this number for the estimate purpose, but the actual cost may vary a great deal for this item.

Training staff travel. In Cohort 3, due to COVID-19, all training was provided virtually and training staff did not need to travel. However, with the pandemic situation winding down, in-person training is

[7] Given that most of the schools in the United States provided laptops to every student during COVID-19 for remote learning, we assume only 10% of the laptop use time should be attributed to the program.

anticipated for the summer. Therefore, we estimate the travel costs based on Cohort 1 information to provide a comprehensive look at program costs. In Cohort 1, one training staff provided two in-person training sessions (one in the summer and one during the academic year) and each lasted 2 days. According to the U.S. Department of Transportation's Bureau of Transportation Statistics (Airline Origin & Destination Survey, 2021), the 2021 national average cost of a round-trip air ticket was around $307. As reported by FederalPay.org (Fiscal Year [FY] 2021 Per Diem Rates), the average rate of lodging per night was $96 and the rate for meals per day was $55. The daily car rental rate of an economy sedan was maximized at $60, according to the U.S. General Services Administration (Vehicle Short Term Rental, 2020). Therefore, the expense of a 2-day trip for training is about $1,458 per educator.

Student supplies. Teachers also reported other school supplies needed during the ALFA Lab section, including paper copies and binders.

Overall, the costs of ALFA Lab were identified as $29,731 per semester to serve 22 students, which is about $1,351 per student. Over 89% of the costs came from personnel and materials (Table 17).

Table 17. Overall Costs of ALFA Lab

Category	Ingredients	Unit Cost	Unit needs	Total Cost	Percent Cost of Category
Personnel	Teachers/Lab Assistants	$9,030	2	$18,060	68%
	Principals	$470	1	$470	
	Training Staff	$901	2	$1,802	
Facilities	Classroom	$1,023	1	$1,023	3%
Equipment	Computers/Electronic Tablets	$8	24	$192	1%
	Printer	$5	1	$5	
Materials	*Feisty Felines Teacher's Manual*	$20	2	$40	21%
	Feisty Felines Nonconsumable Resources (Unit Library, Posters and Games)	$454	2	$908	
	Set of Feisty Felines Reading Selections and Station Activity Pages	$9	22	$198	
	Heroes Teacher's Manual	$19	2	$38	
	Heroes Nonconsumable Resources	$1345	2	$2,690	
	Set of Heroes Reading Selections and Station Activity Pages	$8	22	$176	
	Galaxy Teacher's Manual	$18	2	$36	
	Galaxy Nonconsumable Resources	$970	2	$1,940	
	Set of Galaxy Reading Selections and Station Activity Pages	$8	22	$176	
Other Inputs	Student Field Trip	$23	22	$506	7%
	Training Staff Travel	$1,458	1	$1,458	

Category	Ingredients	Unit Cost	Unit needs	Total Cost	Percent Cost of Category
	Student Supplies: Binder	$3	3	$9	
	Student Supplies: Paper Copies	$0.04	100	$4	
Total Cost of ALFA Lab (One Semester)				$29,731	100%
Cost Per Student (Assuming 22 Students)				$1,351	

The total costs of operating an ALFA Lab program varied widely across the four participating schools, ranging from $22,684 to $90,776 (Table 18). The primary factors driving the difference were the number of sections and number of participating students in each school. The more sections a school has, with more participating students, the higher the total cost.

Table 18. ALFA Lab Costs by School

Category	Ingredient	Unit $	School 1 – NY		School 2 – CT		School 3 – CA		School 4 – AL	
			Unit Needs	Total Cost	Unit Needs	Total Cost	Unit Needs	Total Cost	Unit Needs	Total Cost
# ALFA Lab Sections			**4**		**1**		**3**		**1**	
# Students			**70**		**24**		**66**		**23**	
Personnel	Teachers/Lab Assistants	$9,030	8	$72,240	2	$18,060	6	$54,180	2	$18,060
	Principals	$470	4	$1,880	1	$470	3	$1,410	1	$470
	Training Staff	$901	2	$1,802	2	$1,802	2	$1,802	2	$1,802
Facilities	Classroom	$1,023	4	$4,092	1	$1,023	3	$3,069	1	$1,023
Equipment	Computers/Electronic Tablets	$8	72	$576	26	$208	68	$544	25	$200
	Printer	$5	4	$20	1	$5	3	$15	1	$5
Materials	*Feisty Felines Teacher's Manual*	$20	2	$40	2	$40	2	$40	2	$40
	Feisty Felines Nionconsumable Resources	$454	2	$908	2	$908	2	$908	2	$908
	Set of Feisty Felines Reading Selections & Station Activity Pages	$9	70	$198	24	$216	66	$594	23	$207
	Heroes Teacher's Manual	$19	2	$38	2	$38	2	$38	2	$38
	Heroes Nonconsumable Resources	$1,345	2	$2,690	2	$2,690	2	$2,690	2	$2,690
	Set of Heroes Reading Selections & Station Activity Pages	$8	70	$560	24	$192	66	$528	23	$184

Category	Ingredient	Unit $	School 1 – NY		School 2 – CT		School 3 – CA		School 4 – AL	
			Unit Needs	Total Cost	Unit Needs	Total Cost	Unit Needs	Total Cost	Unit Needs	Total Cost
# ALFA Lab Sections			**4**		**1**		**3**		**1**	
# Students			**70**		**24**		**66**		**23**	
	Galaxy Teacher's Manual	$18	2	$36	2	$36	2	$36	2	$36
	Galaxy Nonconsumable Resources	$970	2	$1,940	2	$1,940	2	$1,940	2	$1,940
	Set of Galaxy Reading Selections & Station Activity Pages	$8	70	$560	24	$192	66	$528	23	$184
Other Inputs	Student Field Trip	$23	70	$1,610	0	-	66	$1518	0	-
	Training Staff Travel	$1,458	1	$1,458	1	$1,458	1	$1,458	1	$1,458
	Student Supplies: Binder	$3	24	$72	3	$9	9	$27	3	$9
	Student Supplies: Paper Copies	$0.04	400	$16	100	$4	300	$12	100	$4
Total Costs (One Semester)			**$90,736**		**$29,291**		**$71,337**		**$29,258**	
Costs Per ALFA Lab Unit			**$22,684**		**$29,291**		**$23,779**		**$29,258**	
Costs Per Student			**$1,296**		**$1,220**		**$1,080**		**$1,272**	

Research Question 6 – Exploratory (Profiles)

Summary of Findings

The emergent three-profile solution for reading motivation demonstrates unique profile-specific patterns that are theoretically useful given that these patterns tap into varying degrees of the higher GRM construct. The three reading motivation profiles exhibited are labelled as high reading motivation (High RM) profile, low reading motivation (Low RM) profile, and an ambivalent/neutral (Ambivalent RM) profile.

One-way Analyses of Variance (ANOVAs) were conducted to examine the relationship between reading motivation profile membership and reading outcomes, specifically reading frequency and reading achievement. Results indicated that students ascribed to the High RM profile exhibited significantly higher ARMM FRQ scores than those in the Low RM and ambivalent/neutral profiles. Similarly, significant differences were found between the profiles on STAR scores, where students with High RM outperformed those with Low RM and ambivalent/neutral profiles. Taken together, results suggest that a high reading motivation profile is associated with better reading outcomes since students who are more motivated tend to read more and tend to have better reading skills than their peers who are not as motivated.

Reading Motivation Profiles

The profiles were thus named because they displayed patterns of high, ambivalent/neutral, or low reading motivation (Figure 11). Also noted that the three negative constructs (i.e., antisocial, avoidance, and perceived difficulty) were reversed to indicate a lack of the negative construct (i.e., lack of avoidance, lack of perceived difficulty, and lack of antisocial goals for reading). Thus, high scores on these constructs indicate a lack of the negative construct and low scores indicate presence of the negative construct.

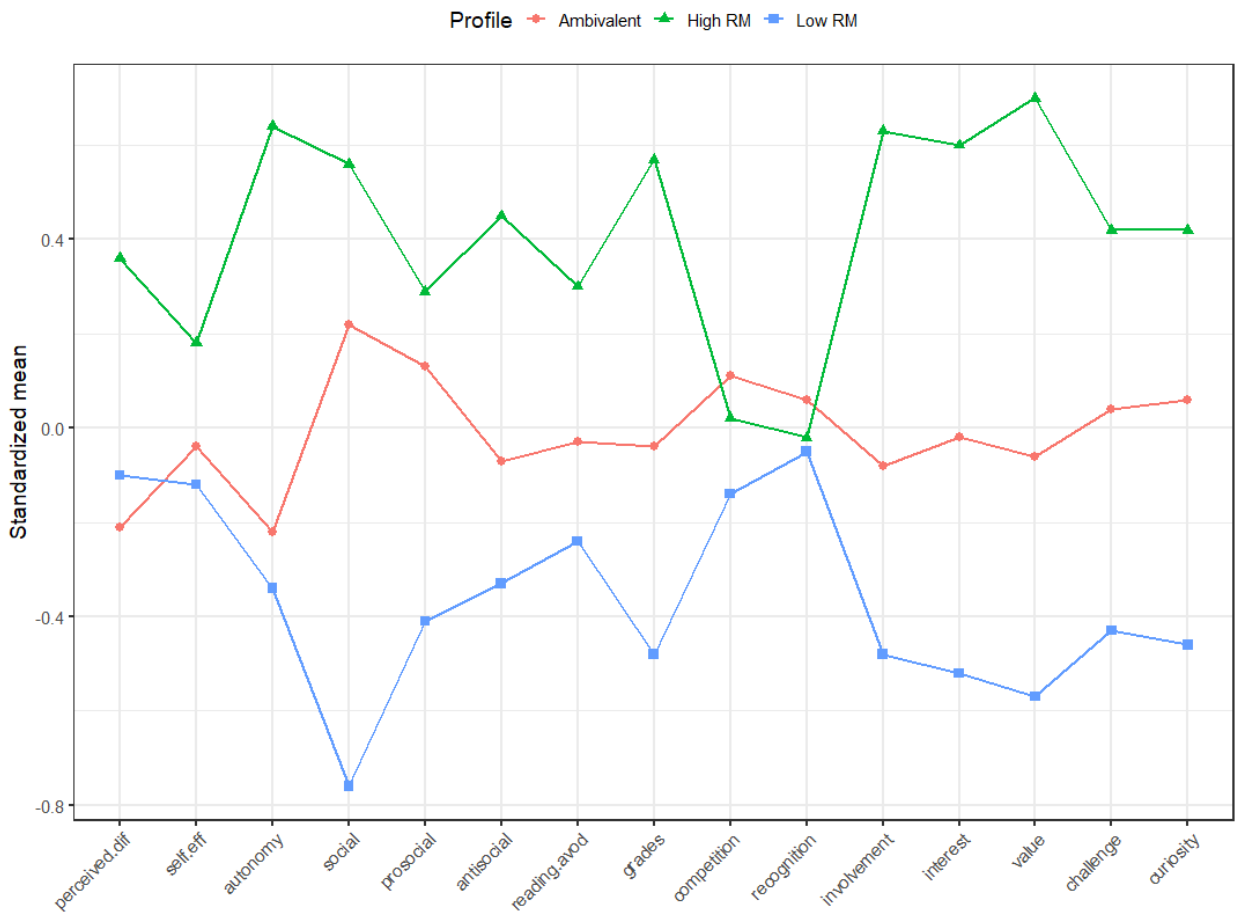


Figure 11. Emergent Three-Profile Reading Motivation Solution

Students ascribed to the High RM profile displayed higher-than-average scores on almost all reading motivation constructs—intrinsic motivation (curiosity, interest, involvement, challenge, and value), social motivation (social, prosocial, and lack of antisocial goals for reading), self-efficacy, autonomy, lack of perceived difficulty of texts, and lack of reading avoidance—whereas students attributed to the Low RM profile displayed lower-than-average scores on these constructs. Students in the ambivalent/neutral profile demonstrated average scores on all reading motivation constructs except for slight advantages in social motivation and prosocial reading goals, and slight deficits in autonomy and lack of perceived difficulty of texts. Note that all three profiles displayed similarly average levels on two extrinsic reading motivation constructs, competition and recognition, but not with grades. This might be because these two constructs contain some social aspects of reading motivation where there is a likelihood to derive external incentives from others (e.g., recognition). The issue of how these constructs is disparate from other aspects of social reading motivation is outside the scope of this research question and can benefit from deeper examination of the theoretical underpinnings of reading motivation.

Relationship between Reading Motivation Profiles and Reading Outcomes

Since students from both treatment and comparison groups were included for analysis, a logistic regression was performed as a preliminary step to ascertain whether profile membership predicted the likelihood of being included in the treatment group. This was done because treatment assignment was based on cut-off scores on a reading measure and there is a recursive relationship between reading skill and reading motivation. Results indicated that the likelihood of a specific profile being included in the treatment is not significant, $\chi^2(2) = 3.52$, p =.172, with 1.3% (Nagelkerke R^2) of the variance explained.

Graphical Analysis

Figure 12. Mean ARMM FRQ Scores Across Profiles (Error Bars Represent Standard Errors)

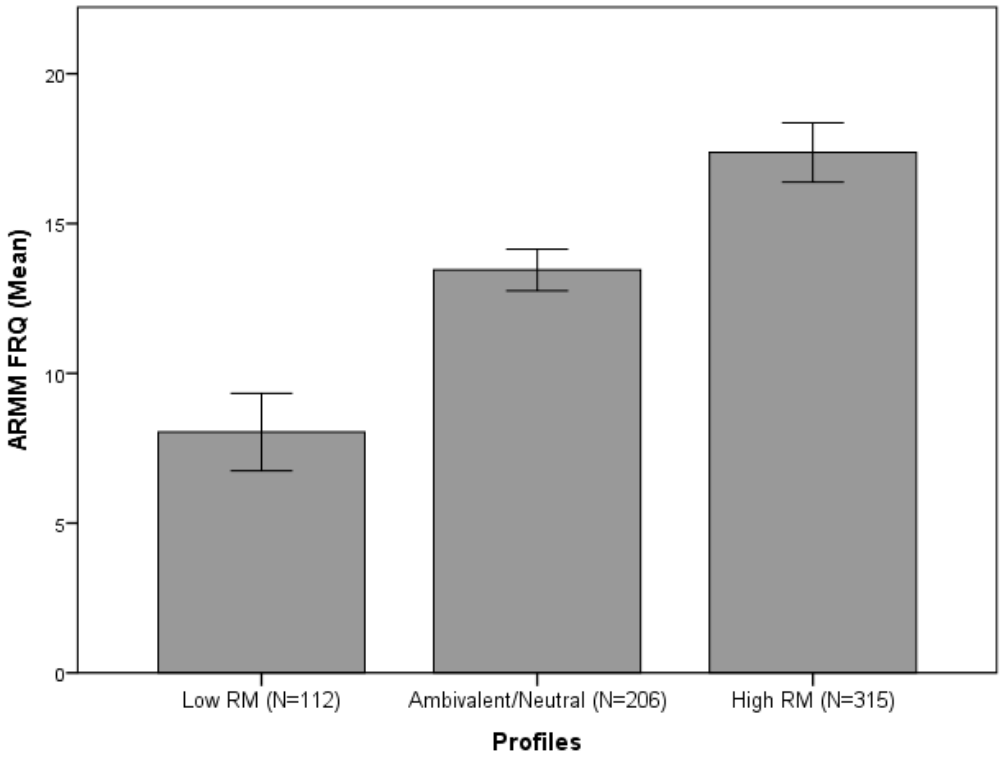


Figure 12 depicts the mean ARMM FRQ scores (i.e., reading frequency) for the three reading motivation profiles. Significant differences were observed, $F(2,354) = 63.08$, $p < .001$, $\eta_p^2 = 0.26$. Bonferroni post hoc comparisons (Table 19) revealed significant pairwise differences in the mean ARMM FRQ scores favoring students ascribed to the High RM profile relative to students with the other two profiles. Students with High RM (M = 17.37, SD = 5.09) read significantly more frequently than students with Low RM (M = 8.04, SD = 4.69) and ambivalent/neutral profiles (M = 13.45, SD = 4.96). Students with Ambivalent RM profiles also read significantly more than students with Low RM.

Table 19. Bonferroni Pairwise Comparisons on Reading Outcomes

Profile Comparisons		ARMM FRQ			STAR Scores		
		Mean Difference	SE	*p*	Mean Difference	SE	*p*
High RM	Ambivalent/ Neutral	3.93	0.61	<.001	138.12	35.32	<.001
	Low RM	9.34	0.84	<.001	214.62	48.98	<.001
Ambivalent/ Neutral	Low RM	5.41	0.77	<.001	76.57	44.96	.269

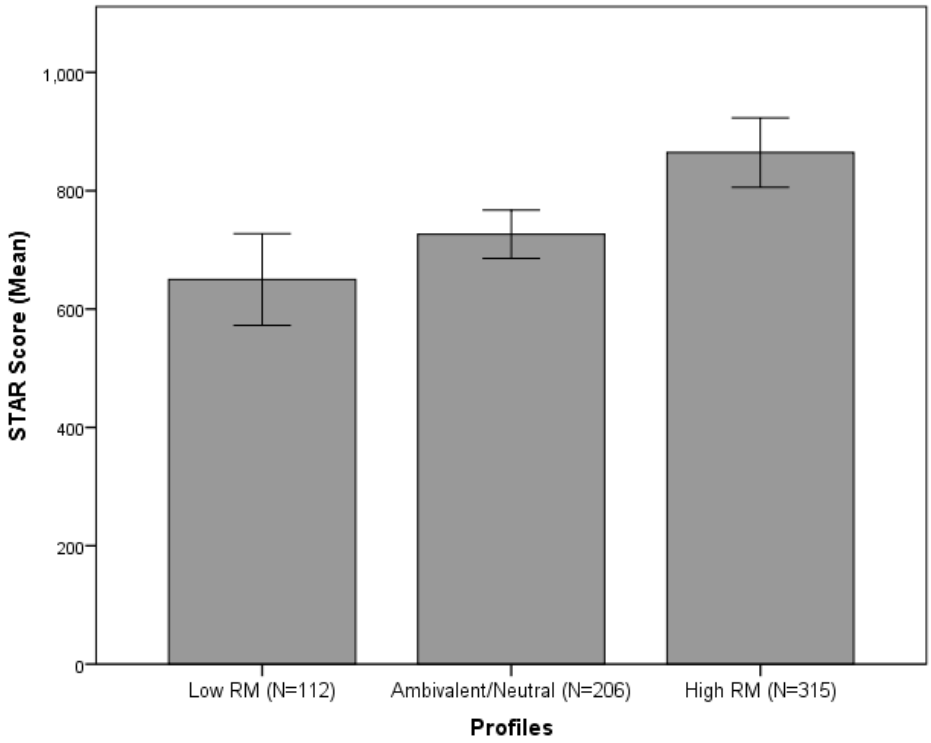


Figure 13. Mean STAR Scores Across Profiles (Error Bars Represent Standard Errors)

Figure 13 represents the mean STAR scores (i.e., reading achievement) for the three reading motivation profiles. Significant differences were observed, $F(2,294) = 11.81$, $p < .001$, $\eta_p^2 = 0.07$ between the profiles. Bonferroni post hoc comparisons (Table 19) revealed significant pairwise comparisons on the mean STAR scores favoring students ascribed to the High RM profile relative to the other two profiles. Students with High RM (M = 864.58, SD = 276.82) outperformed students with Low RM (M = 649.89, SD = 258.07) and Ambivalent/Neutral (M = 726.46, SD = 264.48) reading motivation profiles. These findings provide some support for the predictive utility of the emergent three-profile reading motivation solution given that there is considerable research evidence that students who are motivated to read tend to read more and have improved reading skills relative to their unmotivated peers.

Discussion and Conclusion

The RD design study did not show that the ALFA Lab program significantly impacted reading motivation or achievement, however there was some evidence of slight positive program effects.

The implementation of the ALFA Lab program and this research study (including the research team's ability to collect data from participating schools) were greatly complicated by the COVID-19 pandemic—for all three cohorts. Although the FOI study, conducted for Cohorts 1 and 3, suggested a high level of implementation, the measures used to assess fidelity were designed prior to the pandemic and did not account for factors brought on by the pandemic, such as student and staff absences. For example, the CFAI—a major source for assessing implementation fidelity—primarily accounted for teacher actions, not student actions or their exposure to the program based on the number of days they were in school or their ability to connect virtually. Although attendance data were not collected as part of this study—and were largely deemed unreliable during the pandemic—the evaluation team knows qualitatively, based on discussions with teachers, that student and staff absences were considerable, which limited the potential for program exposure. Although absences were widespread and not limited to students and staff in the treatment group, there is extensive evidence pointing to disproportionate impacts of the pandemic on academic outcomes for students of color, ELLs, and students receiving SPED services (U.S. Department of Education, Office for Civil Rights, 2021). As shown in Table 2, there were slightly higher percentages of students of color, ELLs, and students receiving SPED services in the ALFA Lab treatment group, suggesting that more students participating in ALFA Lab faced disproportionate hurdles brought on by the pandemic compared to students in the BAU group. Accordingly, results from the impact study may not be indicative of the full potential of the program during more typical implementation conditions. Further, although the program did not show a statistically significant impact, researchers heard qualitatively that the ALFA Lab helped to build reading confidence for struggling readers.

Although most of the challenges regarding ALFA Lab implementation related to conditions brought on by the COVID-19 pandemic, ALFA Lab teachers also shared some challenges regarding implementation that were not entirely related to COVID that program developers may consider for any future revisions made to the program. For example, teachers shared that the fixed nature of the program structure and limited time available to implement the program made it difficult to complete all of the program's activities. However, these issues were exacerbated by student and staff absences (which occurred more frequently than normal during the pandemic) and by COVID-inspired efforts to protect students by limiting their movement (instead asking teachers to move from classroom to classroom throughout the day). That is, once the pandemic began, participating schools began asking teachers to have station materials and equipment move rather than student groups when it was time to rotate stations within a classroom and/or to have the whole lab (teachers, assistants, and materials) move from classroom to classroom so that students in a section had limited contact with other students outside their group over the course of the day. Unfortunately, requiring the adults and materials to move so frequently made bell-to-bell instruction a rarity. Finally, teachers from one school perceived the materials to be too juvenile in nature. (In the opinion of staff at this school, the program had not resolved satisfactorily the issue of creating and finding materials that adolescents find interesting and appropriate but that are at the right instructional level for struggling readers.)

The cost analysis demonstrated that the per-student cost of participating in the program was similar across schools at $1,200. Over 70% of the cost came from staffing salaries and/or facilities/equipment, which are provided by the school district and not a responsibility of the schools. Therefore, for the program start-up, including only materials and professional development, it costs approximately $360 per student, which is within the range of other reading intervention programs with similar classroom set up. For example, the start-up cost per student for READ 180® was about $716 as of 2017, excluding teacher salaries and facilities (U.S. Department of Education, 2016). The Exemplary Center for Reading Instruction program, Reading Rockets, with negligible provision in student materials and staff development representing the largest expense, cost about $117 per student (American Federation of Teachers, n.d.).

Despite the pandemic disruptions to program implementation and to this research study as well as the lack of statistically significant main effects on outcomes overall, the study did reveal important findings for subgroups. For example, the moderation analysis found significant, positive effects of ALFA for males, non-ELL students, non-SPED students, and Hispanic students in boosting reading frequency. Further the sub-study regarding student reading motivation profiles, provided an important contribution to research on reading motivation. The profile analysis demonstrated how a high reading motivation profile is associated with better reading outcomes since students who are more motivated tend to read more and tend to have better reading skills than their peers who are not as motivated. The reading motivation profiles sub-study also indicated that students who are ambivalent about reading—because it is difficult for them and their self-efficacy as readers is low—can still be motivated to read frequently when given the opportunity to pursue social and prosocial goals while reading (as they do in the ALFA Lab with readings that also appeal to their intrinsic motivation by engaging their interests, curiosity, and values with an appropriate but not overwhelming level of challenge).

Finally, the study demonstrated that RD is a promising design for estimating impacts of education programs implemented at multiple sites, each with their own unique contexts and complexities. Despite each site's desire to use unique, local measures as the mechanism for assignment in each cohort, through standardization of measures and the use of background demographic covariates to improve precision, overall program impact estimates were obtained. This was accomplished through focused collaboration and clear communication between the implementation team, evaluation team, and school-based site staff. By choosing to work *with* each site to accommodate needs during the assignment process, a rigorous design was instituted while minimizing the burden on schools.

Overall, the ALFA Lab reading intervention is a low-cost intervention with promise and may benefit from enhancements, as suggested by the implementation study, to improve its efficacy. Further study during more normal implementation conditions may be warranted to assess its full potential to enhance reading achievement. Future studies of the program should incorporate student attendance to account for the degree to which students are exposed to the program.

References

American Federation of Teachers. (n.d.) Exemplary Center for Reading Instruction (ECRI) intervention program. *Reading Rockets.* https://www.readingrockets.org/article/exemplary-center-reading-instruction-ecri-intervention-program

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis & Management, 22*(4), 551–75.

Cattaneo, M., Idrobo, N., & Titiunik, R. (2020). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press. https://doi.org/10.1017/9781108684606

Cattaneo, M. D., Jansson, M., & Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association, 115*(531), 1449–1455. https://doi.org/10.1080/01621459.2019.1635480

Cattaneo, M. D., Jansson, M., & Ma, X. (2021). Local regression distribution estimators. *Journal of Econometrics,* online. https://doi.org/10.1016/j.jeconom.2021.01.006

Cattaneo, M. D., Jansson, M., & Ma, X. (2022). lpdensity: Local polynomial density estimation and inference. *Journal of Statistical Software, 101*(2). https://doi.org/10.18637/jss.v101.i02.

Cattaneo, M. D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics, 14*, 821–851.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., & Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics, 78*(4), 1229–1248.

Davis, M. H., Wang, W., Kingston, N. M., Hock, M., Tonks, S. M., & Tiemann, G. (2020). A computer adaptive measure of reading motivation. *Journal of Research in Reading, 43*(4), 434–453. https://doi.org/10.1111/1467-9817.12318

Dong, N., Kelcey, B., & Spybrook, J. K. (2020). Design considerations in multisite randomized trials probing moderated treatment effects. *Journal of Educational and Behavioral Statistics, 46*(5), 527–559.

Enders, C. K. (2010). *Applied missing data analysis.* Guilford Press.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*(3)*,* 233–239.

Gill, B., Lockwood, J. R., Martorell, F., Setodji, C. M., & Booker, K. (2007). *State and local implementation of the No Child Left Behind Act.* U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470.

Haegerich, T. M., & Massetti, G. M. (2013). Commentary on subgroup analysis in intervention research: Opportunities for the public health approach to violence prevention. *Prevention Science 14*(2), 193–198.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis. A regression-based approach* (3rd ed.). Guilford Press.

Hopkins, D. & Craig, W. (2014, February 6). Theories of action – Teacher practice and student achievement. *Teaching Times.* https://www.teachingtimes.com/theories-of-action/

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309–334. https://doi.org/10.1037/a0020761

Levin, H. M., & McEwan, P. (2001). *Cost-effectiveness Analysis: Methods and applications*. Monterey Park, CA: Sage.

Levin, H., McEwan, P., Belfield, C., Bowden, A. & Shand, R. (2018). *Economic Evaluation in Education: Cost-Effectiveness and Benefit-Cost Analysis*. Thousand Oaks, CA: SAGE Publications.

Mongeau, L. (2015, May 28). With 'local control' funds, Calif. district buys field trips. *EducationWeek.* https://www.edweek.org/ew/articles/2015/05/28/with-local-control-funds-calif-district-buys.html

Roberts, G., Scammacca, N., & Roberts, G. J. (2018). Causal mediation in educational intervention studies. *Behavioral Disorders, 43*(4), 457–465. https://doi.org/10.1177/0198742917749560

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons.

STAR Reading (n.d.). *Understanding reliability and validity*. Renaissance Learning Inc.

Supplee, L. H., Kelly, B. C., MacKinnon, D. M., & Barofsky, M. Y. (2013). Introduction to the special issue: Subgroup analysis in prevention and intervention research. *Prevention Science 14*(2), 107–110.

Tingley, D. Yamamoto, T., Hirose, K., Keele, L., Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59(5),* 1–38. https://doi.org/10.18637/jss.v059.i05

U.S. Bureau of Labor Statistics. (2022). *Occupational outlook handbook* (high school teachers). U.S. Department of Labor. https://www.bls.gov/ooh/education-training-and-library/high-school-teachers.htm.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2016). *WWC intervention report: Adolescent literacy – READ 180*. https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_read180_112916.pdf

U.S. Department of Education, Institute of Education Sciences. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. National Center for Education Evaluation and regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf

U.S. Department of Education, Office for Civil Rights. (2021). *Education in a pandemic: The disparate impacts of COVID-19 on America's students.* https://www2.ed.gov/about/offices/list/ocr/docs/20210608-impacts-of-covid19.pdf

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45(3),* 1–67. https://doi.org/10.18637/jss.v045.i03

Wardenaar, K. J. (2021). Latent profile analysis in R: A tutorial and comparison to Mplus. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/wzftr

Wong, V. C., Steiner, P. M. & Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics, 38*(2), 107–141. https://doi.org/10.3102/1076998611432172

Appendix A: Bin Identification & Impact Estimation

Bin Identification

Smoothed plots of RD data are based on mean values of the outcome for units within a bin. The mimicking variance (MV) method chooses the number of bins to ensure binned mean variability is nearly equal to the variability of the raw data. This leads to a larger number of bins relative to the integrated mean-squared error (IMSE) method, providing a better sense of variability. The IMSE method chooses the number of bins that minimizes the asymptotic mean squared error that results from approximating the data by binned means. Choosing a larger number of bins means less bias, but each bin then comprises fewer units which in turn increases variability. The IMSE method facilitates assessing the shape of the underlying regression function and identifying discontinuities away from the cutoff.

The bins can be chosen so that: (a) they are of equal length (evenly-spaced) or (b) they have the same number of observations (quantile-spaced) but with varying lengths. The former tends to exhibit greater variability (since each bin can contain a different number of units) while the latter tends to have more comparable variability across bins (since each bin has the same number of units represented). Quantile methods also provide a natural visual representation of the density of observations across the assignment variable continuum.

Point Estimation & Inference

Mean-squared error (MSE) optimal point estimates are obtained, given a kernel weight and choice of polynomial form, by optimizing the bias-variance trade-off that stems from the bandwidth choice (that is, a bandwidth is chosen to minimize the mean-squared error of the point estimate). The triangular kernel is the MSE-optimal choice for point estimation (Cattaneo, Idrobo & Titiunik, 2020). MSE-optimal bandwidths around the cut-score are identified to minimize the mean-squared error of the point estimate (i.e., optimizing the trade-off between using more data leading to an increase in bias and using less data leading to an increase in variance). Thus, MSE-optimal point estimates are reported along with robust bias-corrected confidence intervals (and associated p-values) that account for the increased variability resulting from estimation of the bias present when approximating the unknown regression function when obtaining a point estimate using the MSE-optimal bandwidth. An alternative approach to inference is to use coverage error (CER-optimal) bandwidths along with MSE-optimal point estimates. CER-optimal bandwidths are identified to minimize the coverage error of the robust bias-corrected confidence intervals. Note that CER-optimal point estimates exhibit increased variability relative to MSE-optimal point estimates and should be ignored.

Appendix B: *WWC Standards 5.0* Evidence

The *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0* (*WWC Standards 5.0*) details standards that must be met to attain a rating of "Meets Standards Without Reservations." This appendix presents information necessary for assessing the validity of the RD impact estimates presented above relative to the *WWC Standards 5.0*.

B.1 Integrity of the Forcing Variable

For an RD design to yield unbiased estimates of intervention effects, there must be no systematic manipulation of the forcing variable that would influence how units are assigned to intervention or business-as-usual (BAU) groups. The integrity of the forcing variable (WWC Standard 1) must be established (a) institutionally, (b) statistically, and (c) graphically.

B.1.1 Description of the Forcing Variable Scoring & Assignment Process

Exhibit B1.1. Assignment Process Description By Cohort & Site

Cohort	School	Description
1	CA	Researchers used the standardized Degrees of Reading Power assessment from the spring of Grade 8 to identify a cut-score of 7.5 during the summer prior to Grade 9. The cut-score was chosen to identify the number of students the school had the capacity to serve.
	NY	Researchers used the standardized New York State English Language Arts Power assessment from the spring of Grade 7 to identify a cut-score of 581 during the summer prior to Grade 9. The cut-score was chosen to identify the number of students the school had the capacity to serve.
2	CA*	Researchers used the standardized California Assessment of Student Performance and Progress (CAASPP) standardized Grade 7 English Language Arts (ELA) assessment and Grade 8 ELA, social studies, and science course points (e.g., A+ = 4.3). Both the ELA scale score and the ELA reading subscale categorical performance indicator were standardized and averaged to yield the first subcomposite value. The second subcomposite value is the average of the three course point values, each standardized). These two subcomposite values were averaged to yield the overall assignment score. The cut-score (-0.167) was chosen to identify the number of students the school had the capacity to serve.
	NY	Researchers used the standardized New York State English Language Arts Power assessment from the spring of Grade 7 to identify a cut-score of 588 during the summer prior to Grade 9. The cut-score was chosen to identify the number of students the school had the capacity to serve.
3	AL	Researchers used the standardized Edmentum Reading assessment from the spring of Grade 8 to identify a cut-score of 1,160 during the summer prior to Grade

Cohort	School	Description
		9. The cut-score was chosen to identify the number of students the school had the capacity to serve.
	CA	Researchers used the standardized STAR Reading assessment from the fall of Grade 8 to identify a cut-score of 481 during the summer prior to Grade 9. The cut-score was chosen to identify the number of students the school had the capacity to serve during the fall block of the cohort year. Researchers used the standardized STAR Reading assessment from the spring of Grade 8 to identify a cut-score of 588 during the fall of Grade 9. The cut-score was chosen to identify the number of students the school had the capacity to serve during the spring block of the cohort year.
	CT	Researchers used the standardized STAR Reading assessment from the fall of Grade 9 to identify a cut-score of 6.4 during the first week of Grade 9. The cut-score was chosen to identify the number of students the school had the capacity to serve during the fall block of the cohort year.
	NY	Researchers used the standardized iReady Reading assessment from the spring of Grade 8 to identify a cut-score of 600 during the summer prior to Grade 9. The cut-score was chosen to identify the number of students the school had the capacity to serve.

*The resulting assignment variable for CA Cohort 2 was comprised of two subcomposites, described below.

- A subcomposite based on two indicators drawn from the Grade 7 CAASPP test reports:
  - The scale score from the 2019 ELA test
  - A categorical variable representing each student's performance in the reading area of the ELA test (1 = below standard, 2 = near standard, and 3 = above standard)
- A subcomposite based on three indicators drawn from student transcripts:
  - Grade 8 course points (A+ = 4.3, A = 4, A- = 3.7, etc.) for ELA
  - Grade 8 course points for social studies
  - Grade 8 course points for science

The first composite score for each student was calculated by creating a z-score of the ELA scale score and a z-score of the reading category ratings. The average of these two z-scores served as the first assignment subcomposite value, formulaically represented as:

$$Subcomp_{1ij} = \left(\frac{\left(ELA_i - \overline{ELA_j}\right)}{ELAsd_j} + \frac{\left(RD_i - \overline{RD_j}\right)}{RDsd_j} \right) / 2$$

Here ELA_i is the CAASPP English scale score for each student *i* in site *j* and $\overline{ELA_j}$ is the average CAASPP English scale score at site *j*. RD_i is the CAASPP reading subtest performance category for each student and $\overline{RD_j}$ is the average CAASPP reading subtest performance category at site *j*. Note that each student could possess one of three unique z-scores based on this categorical variable. The second subcomposite value is formulated as:

$$Subcomp_{2ij} = \left(\frac{\left(Eng8_i - \overline{Eng8_j}\right)}{Eng8sd_j} + \frac{\left(SS8_i - \overline{SS8_j}\right)}{SS8sd_j} + \frac{\left(Sci8_i - \overline{Sci8_j}\right)}{Sci8sd_j} \right) / 3$$

Here $Eng8_i$ is the 2019–2020 first semester English grade point for each student *i* in site *j* (and $\overline{Eng8_j}$ is the average English grade point at site *j*. Correspondingly, $SS8_i$ is the social studies grade point for each student and $Sci8_i$ is the science grade point for each student (and their accompanying site-based average terms). The overall composite assignment variable value for each student in a site is the average of the two subcomposite scores:

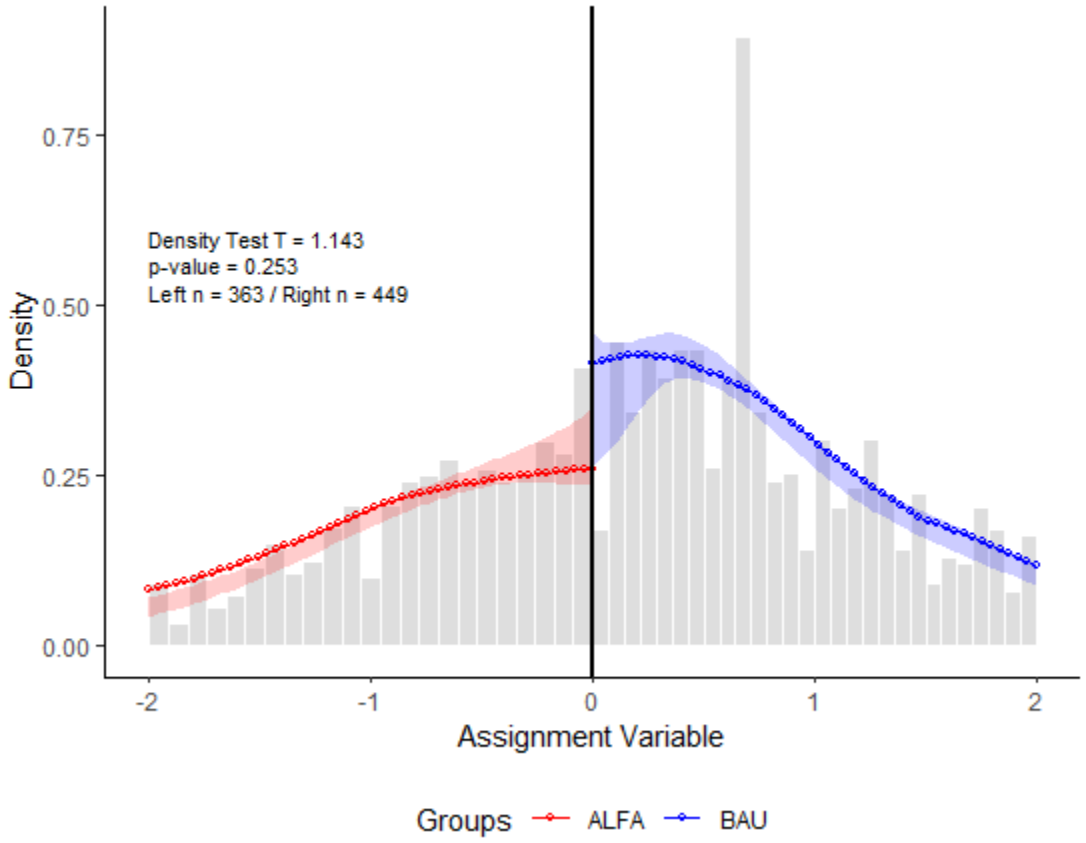
$$Comp_{ij} = \left(Subcomp_{1ij} + Subcomp_{2ij}\right)/2$$

B.1.2 Institutional Integrity of the Forcing Variable

WWC standards require that a statistical test (substantiated in the research methods literature) be used to establish the smoothness of the forcing variable density around the cutoff. We made use of the local polynomial density estimators proposed by Cattaneo, Jansson, and Ma (2019) and the graphical procedures and confidence bands defined by Cattaneo, Jansson, and Ma (2021, 2022).

Figure B1.2.1 below presents a density plot of the standardized, centered assignment variable used in the RD analysis of the STAR and ARMM outcomes. The sample used to conduct the inferential manipulation test below comprised all students who were eligible for assignment (i.e., those with assignment values; those who had not opted out of the study or withdrawn from school prior to assignment). The test statistic (T), the associated p-value, and the number of observations identified in the left and right bandwidths, respectively, are included in the graph.

Figure B1.2.1. Density Plot & Inferential Assessment of Smoothness



Note the high density around 0.6. The running score value of 0.681 is associated with the Degrees of Reading Power assignment scores used in the Cohort 1 CA site. Fifty students attained a score of 9 (the highest score value on a restricted range variable (2.5 to 9).

B.2 Sample Attrition & Baseline Equivalence

An RD must have acceptable levels of overall and differential attrition, or must establish baseline equivalence on key covariates, to meet WWC standards.

B.2.1 Sample Attrition

Exhibit B2.1 presents the predicted attrition rates, by experimental group and overall, for each outcome and regression discontinuity (RD) design. The data modeled for each outcome included those falling within the bandwidth identified in the final RD model using triangular kernel weights and a linear specification. A binary indicator representing attrition (or not) served as the dependent variable in a weighted, generalized linear model including the standardized, centered running variable, an indicator denoting whether the running variable value was below (ALFA Lab) or above (BAU) the cut-score, and an interaction term between the running variable and the intervention indicator. Attrition indicators represent instances when students (a) did not have a valid outcome measure or (b) they left school after the study began (e.g., dropped out, transferred to a different school, etc.). We believe that the high levels of attrition in this study stem from immediate and longer-term impacts related to the COVID-19 pandemic. The ability to obtain outcome measures from students in both ALFA Lab and BAU groups, across all three cohorts, was severely compromised by remote-learning scenarios and increased teacher responsibility for learning recovery (*WWC Standards 5.0*, p. 41). Thus, students with missing outcome data should be excluded from the denominator in the attrition calculation. Exhibit B2.2 presents the predicted attrition rates when students with missing outcome data are excluded from the calculations.

Exhibit B2.1. Predicted Attrition Rates by Outcome

Outcome	Design	ALFA Rate	BAU Rate	Differential	Overall	Cautious Rating	Optimistic Rating
STAR	ITT	44.1	34.7	9.4	39.4	High	High
	Fuzzy	44.4	34.9	9.5	39.6	High	High
ARMM GRM	ITT	57.0	46.2	10.8	51.6	High	High
	Fuzzy	56.6	46.4	10.2	51.5	High	High
ARMM FRQ	ITT	56.4	47.8	8.6	52.1	High	High
	Fuzzy	56.3	48.0	8.3	52.1	High	High

Exhibit B2.2. Predicted Attrition Rates by Outcome – Missing Outcome Excluded

Outcome	Design	ALFA Rate	BAU Rate	Differential	Overall	Cautious Rating	Optimistic Rating
STAR	ITT	16.1	15.6	0.5	15.9	Low	Low
	Fuzzy	16.3	16.5	0.2	16.4	Low	Low
ARMM GRM	ITT	16.1	16.6	0.5	16.4	Low	Low

	Fuzzy	16.2	16.7	0.5	16.4	Low	Low
ARMM FRQ	ITT	16.1	16.7	0.6	16.4	Low	Low
	Fuzzy	16.1	16.6	0.5	16.4	Low	Low

Exhibit B2.3 presents predicted covariate proportions at the cut-point used in the intent-to-treat (ITT) and Fuzzy RD models, for each outcome.

Exhibit B2.3. Predicted Covariate Probabilities for Baseline Equivalence

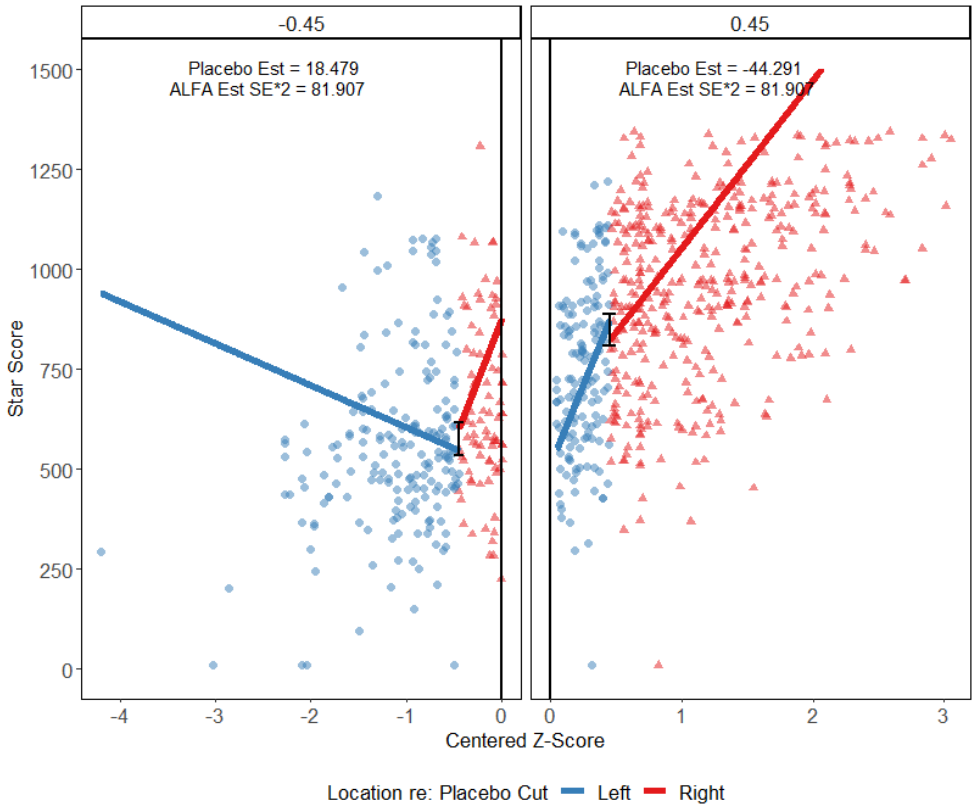
		ITT					Fuzzy				
Outcome	Covariate	ALFA Rate	BAU Rate	Differential	Cox *d*	Overall	ALFA Rate	BAU Rate	Differential	Cox *d*	Overall
STAR	Gender	53.2	55.6	2.4	-0.05	54.4	55.4	55.6	0.2	0.00	55.5
	ELL	13.0	9.6	3.4	0.21	11.3	9.6	9.6	0.0	0.00	9.6
	SPED	3.0	1.6	1.4	0.39	2.3	1.2	1.6	0.4	-0.18	1.4
	Native American	2.1	2.6	0.5	-0.13	2.4	2.4	2.6	0.2	-0.05	2.5
	Asian	1.3	1.2	0.1	0.05	1.3	0.9	1.2	0.3	-0.18	1.1
	African American	23.3	15.0	8.3	0.33	19.2	15.8	15.0	0.8	0.04	15.4
	Did Not Respond	2.6	0.7	1.9	0.81	1.7	1.0	0.7	0.3	0.22	0.9
	Hispanic	44.3	50.3	6.0	-0.15	47.3	49.9	50.3	0.4	-0.01	50.1
	Multi-Racial	1.8	2.4	0.6	-0.18	2.1	2.4	2.4	0.0	0.00	2.4
	White	24.7	27.8	3.1	-0.10	26.3	27.6	27.8	0.2	-0.01	27.7
ARMM GRM	Gender	56.8	54.6	2.2	0.05	53.4	57.7	54.7	3.0	0.07	53.4
	ELL	15.0	9.7	5.3	0.30	13.0	15.6	9.7	5.9	0.33	13.0
	SPED	3.6	1.7	1.9	0.47	3.4	3.6	1.8	1.8	0.43	3.4
	Native American	1.3	3.1	1.8	-0.54	2.2	1.7	3.1	1.4	-0.37	2.2
	Asian	1.6	1.0	0.6	0.29	1.2	1.4	1.0	0.4	0.21	1.2
	African American	19.6	15.3	4.3	0.18	23.5	19.1	15.2	3.9	0.17	23.5
	Did Not Respond	1.7	1.1	0.6	0.27	2.6	1.6	1.0	0.6	0.29	2.6
	Hispanic	49.3	50.6	1.3	-0.03	44.1	49.8	50.5	0.7	-0.02	44.1
	Multi-Racial	0.3	2.6	2.3	-1.32	1.8	0.6	2.6	2.0	-0.90	1.8
	White	26.2	26.3	0.1	0.00	24.7	25.8	26.4	0.6	-0.02	24.7
ARMM FRQ	Gender	59.7	54.2	5.5	0.14	53.4	59.8	54.3	5.5	0.14	53.4
	ELL	15.5	10.0	5.5	0.30	13.0	15.6	9.9	5.7	0.32	13.0
	SPED	5.3	1.8	3.5	0.68	3.4	5.5	1.8	3.7	0.70	3.4
	Native American	2.2	3.2	1.0	-0.23	2.2	2.2	3.2	1.0	-0.23	2.2
	Asian	1.0	0.9	0.1	0.06	1.2	0.9	0.9	0.0	0.00	1.2
	African American	19.9	14.8	5.1	0.22	23.5	19.9	14.8	5.1	0.22	23.5
	Did Not Respond	1.5	1.0	0.5	0.25	2.6	1.6	1.0	0.6	0.29	2.6
	Hispanic	48.4	50.8	2.4	-0.06	44.1	48.3	50.8	2.5	-0.06	44.1
	Multi-Racial	1.1	2.7	1.6	-0.55	1.8	1.2	2.7	1.5	-0.50	1.8
	White	25.9	26.6	0.7	-0.02	24.7	25.9	26.6	0.7	-0.02	24.7

*Cox *d* is the effect size for binary outcomes required by *WWC Standards 5.0* (see p. 177).

B.3 Continuity of the Outcome/Forcing Variable Relationship

WWC standards require establishing the continuity of the relationship between the outcome variable and the assignment variable. One indirect assessment is to see if discontinuities are detected at placebo cut-points along the continuum of the assignment variable. A placebo model was run using a placebo cut-point and only data falling below the true cut-point (i.e., zero) when the placebo cut was less than zero or data falling above the true cut-point when the placebo cut was greater than zero. The graphs presented below display raw data points via a scatterplot of the running variable on the x-axis and the outcome variable on the y-axis, along with the estimated effects generated from the placebo cut-point model. Data points are colored based on where they fall regarding the placebo cut (i.e., points to the left are blue, points to the right are red). The error bar is centered around the differences between the blue and red lines *at* the placebo cut-point, with the vertical distance of the error bar in either direction (above or below the middle) representing the standard error of the impact estimate from the target model (i.e., equaling two times the standard error). Annotated in each graph are the RD estimates associated with the placebo cut-score (Placebo Est) and twice the corresponding outcome impact estimate standard error (ALFA Est SE*2).

Figure B3.1.1. Scatterplot Plot & Error Bars – STAR Outcome Placebo Cut (-.45 / .45)



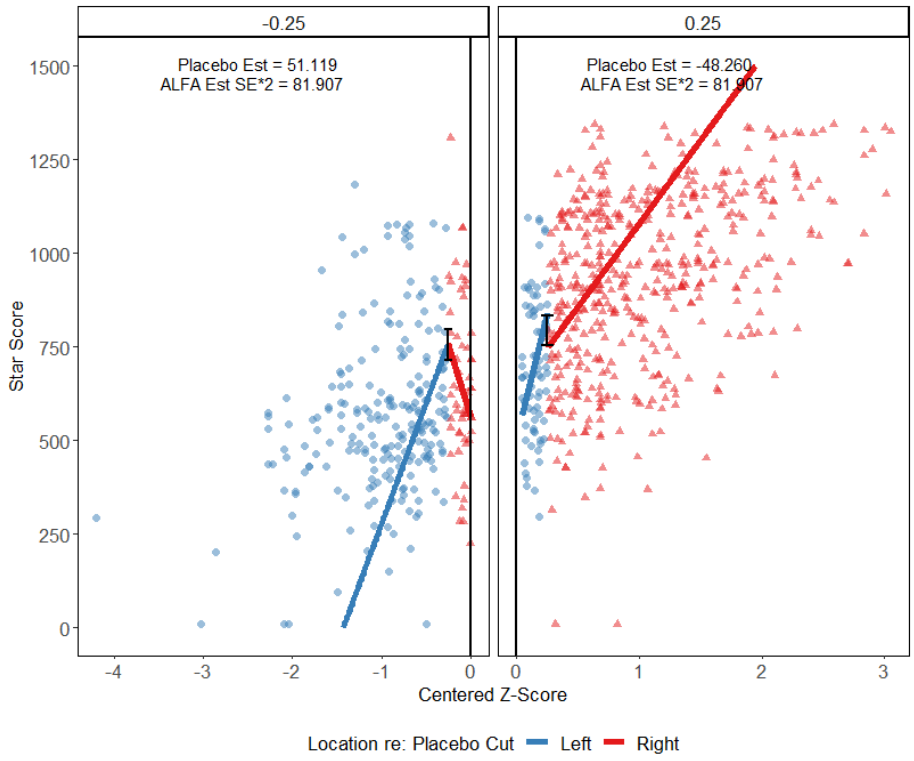


Figure B3.1.2. Scatterplot Plot & Error Bars – STAR Outcome Placebo Cut (-.25 / .25)

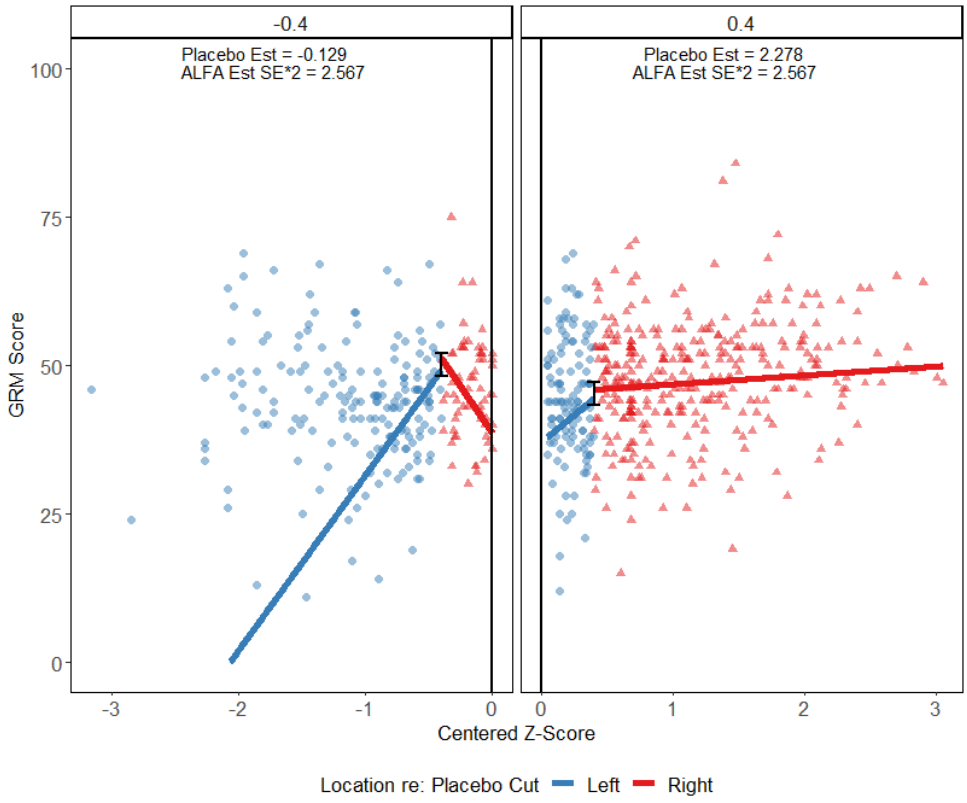


Figure B3.1.3. Scatterplot Plot & Error Bars – GRM Outcome Placebo Cut (-.4 / .4)

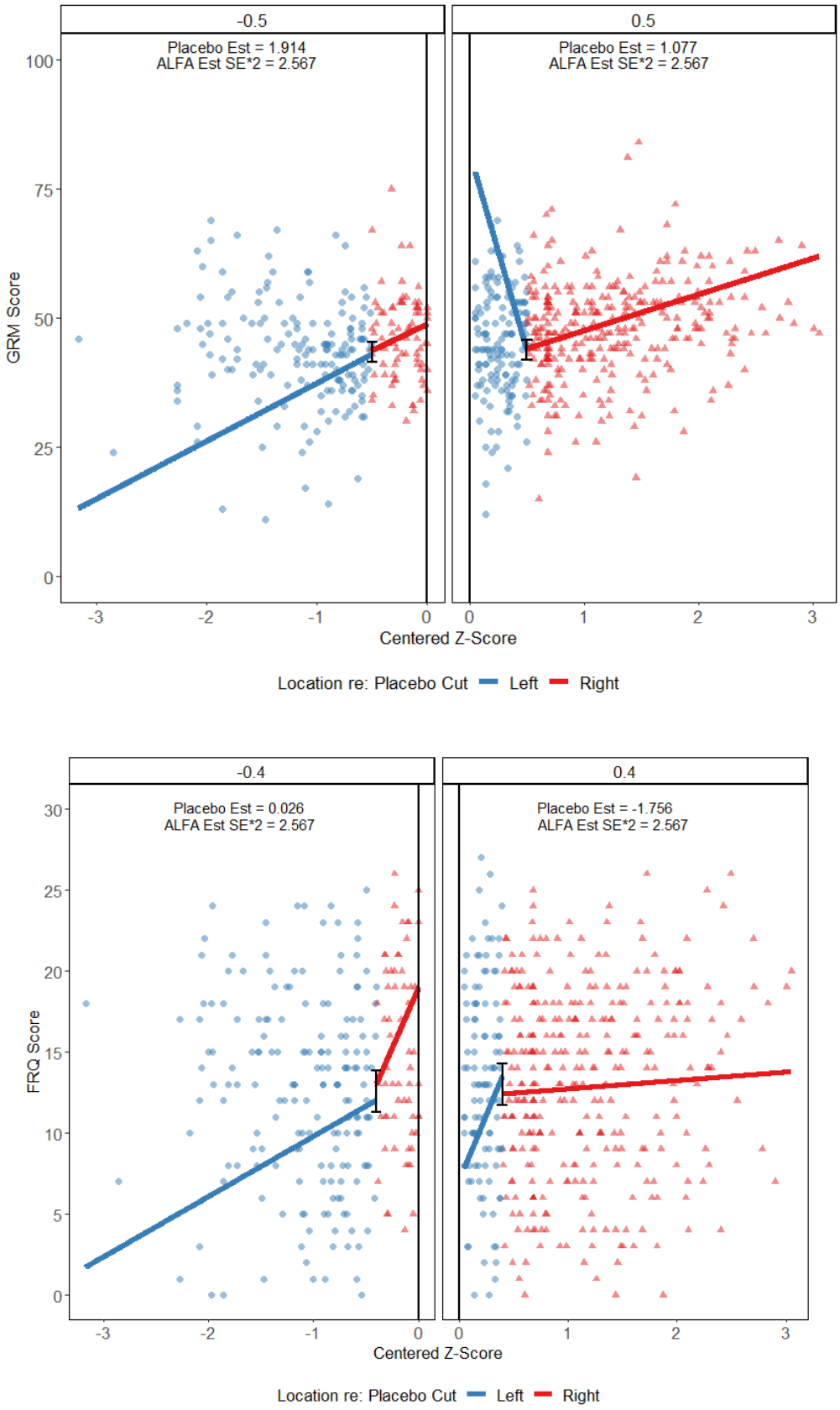


Figure B3.1.4. Scatterplot Plot & Error Bars – GRM Outcome Placebo Cut (-.5 / .5)

Figure B3.1.5. Scatterplot Plot & Error Bars – FRQ Outcome Placebo Cut (-.4 / .4)

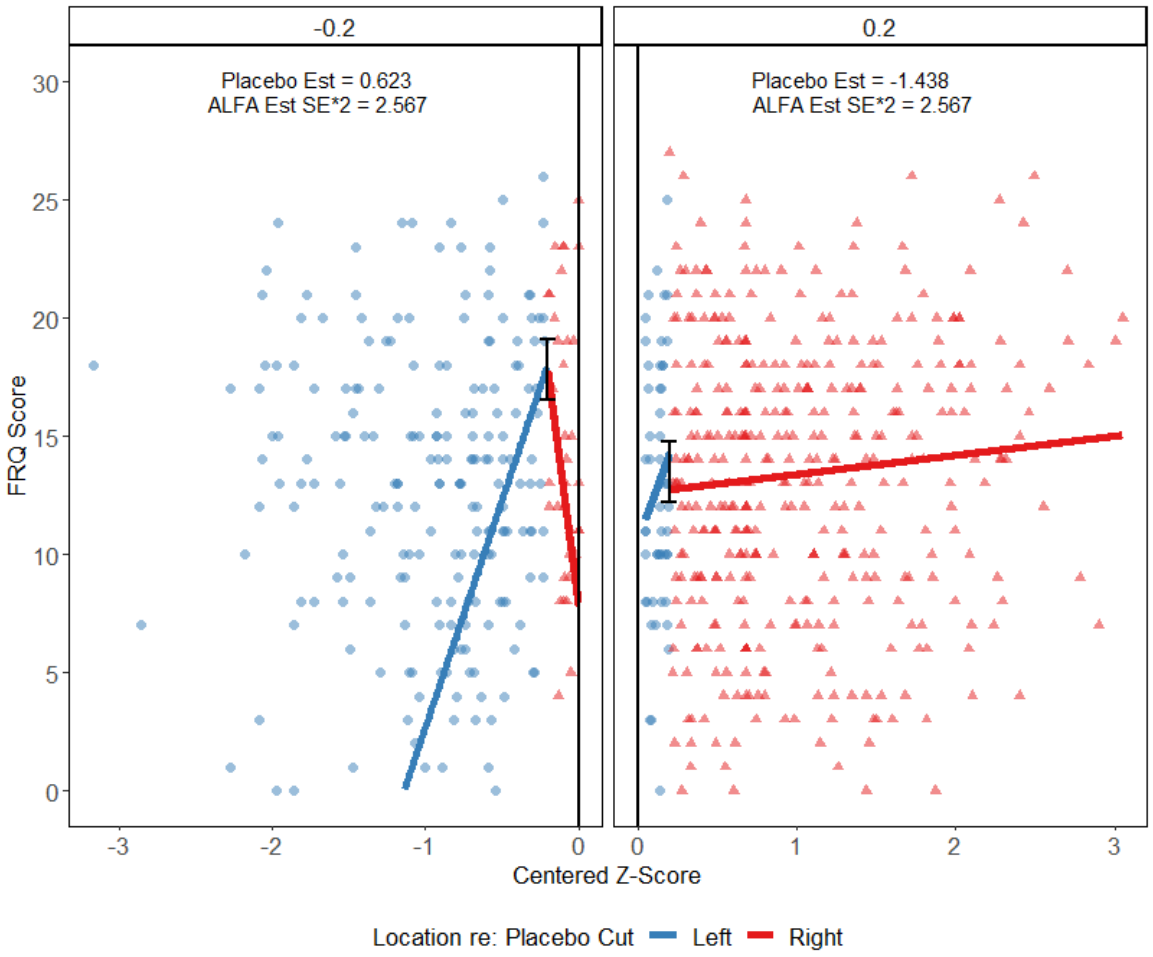


Figure B3.1.6. Scatterplot Plot & Error Bars – FRQ Outcome Placebo Cut (-.2 / .2)

Exhibit B3.1 on page B-11 displays the results of inferential assessment of impacts on the STAR outcome at various placebo cut-scores along the assignment score continuum. All estimates were obtained using the same bandwidth identification algorithm (i.e., separate bandwidths on both sides of the cut-off) and estimation model using a triangular kernel and linear specification that was used in obtaining the overall impact estimates. To avoid treatment effect contamination in estimating placebo effects, we followed guidance from Cattaneo and Titiunik (2022) and used only units that experienced ALFA Lab and were below the true cut (0) when the placebo cut was less than zero and units experiencing BAU and were above the true cut when the placebo cut was greater than zero. Note we present robust, bias-corrected estimates as we are primarily interested in inferential assessment of cut-scores along the assignment continuum. All eight impact estimates were not statistically significant.

Exhibit B3.1. Assessment of Discontinuities at Placebo Cut Scores – STAR Outcome

Placebo Cut	n (TX / BAU)	Est	p-value	CI	Bandwidth	ES
-1.50	38 / 17	185.05	0.190	[-91.75, 461.85]	[-0.67 - 0.77]	0.718
-0.85	39 / 45	-50.02	0.560	[-218.00, 117.96]	[-0.33 - 0.49]	-0.305
-0.50	32 / 89	99.82	0.287	[-84.07, 283.72]	[-0.29 - 0.83]	0.429
-0.20	41 / 93	15.89	0.889	[-206.77, 238.55]	[-0.20 - 0.88]	0.230
0.20	145 / 57	-56.99	0.409	[-192.38, 78.40]	[-0.69 - 0.15]	-0.181
0.50	176 / 93	-11.35	0.833	[-116.80, 94.09]	[-1.05 - 0.31]	-0.064
0.85	115 / 78	-119.27	0.157	[-284.59, 46.05]	[-0.96 - 0.35]	-0.407
1.50	44 / 45	121.55	0.106	[-25.61, 268.71]	[-0.57 - 0.35]	0.329

Exhibit B3.2 provides the same assessment of placebo discontinuities associated with the GRM outcome. All estimates were obtained using the same process as described above for the STAR placebo assessment. Three of the impact estimates were statistically significant, yielding an overall non-significance rate of 62.5 (5/8).

Exhibit B3.2. Assessment of Discontinuities at Placebo Cut Scores – GRM Outcome

Placebo Cut	n (TX / BAU)	Est	p-value	CI	Bandwidth	ES
-1.50	45 / 17	7.72	0.211	[-4.36, 19.81]	[-0.64 - 0.48]	0.613
-0.85	37 / 44	-8.22	0.029	[-15.63, -0.82]	[-0.40 - 0.64]	-0.761
-0.50	33 / 72	-0.93	0.790	[-7.75, 5.90]	[-0.28 - 0.88]	-0.210
-0.20	34 / 66	7.87	0.040	[0.38, 15.36]	[-0.20 - 0.82]	0.763
0.20	177 / 53	2.86	0.554	[-6.63, 12.35]	[-0.96 - 0.15]	0.376
0.50	127 / 28	-4.79	0.242	[-12.81, 3.24]	[-0.88 - 0.16]	-0.119
0.85	19 / 75	-16.20	0.000	[-24.09, -8.30]	[-0.18 - 0.44]	-1.412
1.50	26 / 37	4.97	0.408	[-6.80, 16.73]	[-0.39 - 0.36]	0.329

Exhibit B3.3 provides the same assessment of placebo discontinuities associated with the FRQ outcome. All estimates were obtained using the same process as described above for the STAR placebo assessment. All of the impact estimates were not statistically significant.

Exhibit B3.3. Assessment of Discontinuities at Placebo Cut Scores – FRQ Outcome

Placebo Cut	n (TX / BAU)	Est	p-value	CI	Bandwidth	ES
-1.50	45 / 9	1.97	0.391	[-2.53, 6.46]	[-0.67 - 0.43]	0.356
-0.85	35 / 33	-1.15	0.694	[-6.90, 4.60]	[-0.31 - 0.38]	-0.164
-0.50	35 / 74	-1.09	0.633	[-5.59, 3.40]	[-0.30 - 0.77]	-0.136
-0.20	34 / 62	-0.52	0.852	[-5.98, 4.94]	[-0.20 - 0.75]	-0.105
0.20	210 / 50	1.12	0.651	[-3.73, 5.97]	[-1.36 - 0.15]	0.247
0.50	136 / 27	-2.81	0.125	[-6.40, 0.78]	[-0.96 - 0.16]	-0.429
0.85	70 / 76	-2.61	0.296	[-7.51, 2.28]	[-0.67 - 0.46]	-0.372
1.50	42 / 29	-5.30	0.135	[-12.24, 1.64]	[-0.61 - 0.25]	-0.696

B.4 Functional Form & Bandwidth

To meet standards without reservations, the WWC requires: (i) statistical analysis that controls for the assignment variable (criterion A), (ii) use of a local regression within a justified bandwidth (criterion B), (iii) evidence that impact estimates are robust to (a) varying specifications of the polynomial relationship between the assignment variable and the outcome, and/or (b) different bandwidths (criterion D), a graphical analysis illustrating the relationship between the outcome and the forcing variable (criterion E), and a modeling approach that allows the relationship between the assignment variable and the outcome to differ on different sides of the cut-point (criterion F). Exhibit B4.1 provides hyperlinks to evidence for each *WWC™ Standards Handbook, Version 4* RD criterion.

Exhibit B4.1. Functional Form & Bandwidth Evidence

Criterion	Evidence
A	See formula in Regression Discontinuity Design
B	See description on page 7 in Regression Discontinuity Design
D	See Tables 8, 10, and 12 for the STAR, GRM, and FRQ outcomes, respectively
E	See Figures 4, 7, and 10 for the STAR, GRM, and FRQ outcomes, respectively
F	See formula in Regression Discontinuity Design

B.5 Fuzzy Design Standards

Exhibit B5.1. First-Stage Regression Estimates of Participation Indicator – STAR Outcome

	Kernel	BW Selection	p	Est	se	t	p-value	CI
MSE-Optimal	Uniform	Common	1	-0.77	0.058	-13.387	0.000	[-0.89, -0.66]
		Separate	1	-0.79	0.067	-11.796	0.000	[-0.92, -0.66]
		Common	1	-0.75	0.086	-8.748	0.000	[-0.92, -0.58]
		Separate	1	-0.75	0.085	-8.798	0.000	[-0.91, -0.58]
	Triangular	Common	1	-0.76	0.061	-12.481	0.000	[-0.88, -0.64]
		Separate	1	-0.76	0.069	-11.146	0.000	[-0.90, -0.63]
		Common	1	-0.74	0.087	-8.567	0.000	[-0.91, -0.57]
		Separate	1	-0.74	0.082	-9.073	0.000	[-0.90, -0.58]
CER-Optimal	Uniform	Common	1	-0.74	0.065	-11.357	0.000	[-0.87, -0.61]
		Separate	1	-0.72	0.070	-10.218	0.000	[-0.86, -0.58]
		Common	1	-0.73	0.104	-6.991	0.000	[-0.93, -0.52]
		Separate	1	-0.71	0.109	-6.576	0.000	[-0.93, -0.50]
	Triangular	Common	1	-0.76	0.065	-11.617	0.000	[-0.88, -0.63]
		Separate	1	-0.75	0.074	-10.097	0.000	[-0.89, -0.60]
		Common	1	-0.72	0.100	-7.171	0.000	[-0.92, -0.52]
		Separate	1	-0.72	0.098	-7.366	0.000	[-0.91, -0.53]

Exhibit B5.2. First-Stage Regression Estimates of Participation Indicator – GRM Outcome

BW Type	Kernel	BW Selection	p	Est	se	t	p-value	CI
MSE-Optimal	Uniform	Common	1	-0.81	0.071	-11.365	0.000	[-0.95, -0.67]
		Separate	1	-0.75	0.078	-9.598	0.000	[-0.90, -0.60]
		Common	1	-0.67	0.131	-5.132	0.000	[-0.93, -0.42]
		Separate	1	-0.71	0.107	-6.664	0.000	[-0.92, -0.50]
	Triangular	Common	1	-0.79	0.076	-10.436	0.000	[-0.94, -0.65]
		Separate	1	-0.76	0.077	-9.771	0.000	[-0.91, -0.60]
		Common	1	-0.74	0.106	-6.994	0.000	[-0.95, -0.54]
		Separate	1	-0.75	0.100	-7.449	0.000	[-0.94, -0.55]
CER-Optimal	Uniform	Common	1	-0.81	0.080	-10.074	0.000	[-0.96, -0.65]
		Separate	1	-0.76	0.078	-9.792	0.000	[-0.92, -0.61]
		Common	1	-0.66	0.134	-4.947	0.000	[-0.92, -0.40]
		Separate	1	-0.74	0.112	-6.589	0.000	[-0.96, -0.52]
	Triangular	Common	1	-0.78	0.077	-10.055	0.000	[-0.93, -0.63]
		Separate	1	-0.77	0.078	-9.785	0.000	[-0.92, -0.61]
		Common	1	-0.70	0.114	-6.141	0.000	[-0.93, -0.48]
		Separate	1	-0.72	0.108	-6.694	0.000	[-0.93, -0.51]

Exhibit B5.3. First-Stage Regression Estimates of Participation Indicator – FRQ Outcome

BW Type	Kernel	BW Selection	p	Est	se	t	p-value	CI
MSE-Optimal	Uniform	Common	1	-0.81	0.065	-12.523	0.000	[-0.93, -0.68]
		Separate	1	-0.79	0.069	-11.570	0.000	[-0.93, -0.66]
		Common	1	-0.73	0.115	-6.340	0.000	[-0.95, -0.50]
		Separate	1	-0.73	0.099	-7.319	0.000	[-0.92, -0.53]
	Triangular	Common	1	-0.80	0.061	-13.270	0.000	[-0.92, -0.69]
		Separate	1	-0.77	0.071	-10.779	0.000	[-0.91, -0.63]
		Common	1	-0.73	0.110	-6.670	0.000	[-0.95, -0.52]
		Separate	1	-0.73	0.103	-7.045	0.000	[-0.93, -0.53]
CER-Optimal	Uniform	Common	1	-0.78	0.076	-10.234	0.000	[-0.93, -0.63]
		Separate	1	-0.77	0.077	-9.975	0.000	[-0.92, -0.62]
		Common	1	-0.66	0.128	-5.184	0.000	[-0.92, -0.41]
		Separate	1	-0.71	0.113	-6.315	0.000	[-0.93, -0.49]
	Triangular	Common	1	-0.80	0.065	-12.340	0.000	[-0.92, -0.67]
		Separate	1	-0.77	0.074	-10.353	0.000	[-0.92, -0.62]
		Common	1	-0.69	0.117	-5.931	0.000	[-0.92, -0.46]
		Separate	1	-0.70	0.113	-6.244	0.000	[-0.92, -0.48]

Exhibit B5.4 summarizes the evidence that the study meets each of the eight criteria identified by the WWC standard for fuzzy RD designs.

Exhibit B5.4. Fuzzy Design Evidence

Criterion	Evidence
A	For all Fuzzy models, an indicator denoting participation was based on class assignment data collected from each site. Despite assignment scores, some students were assigned to courses representing the opposite of what should have occurred based on the cut-point. Because the rdrobust package is parameterized for impact estimates above the cutoff, the participation indicator was coded such that participants in ALFA Lab were = "0" and BAU were = "1," even if they should have been assigned to the opposite group based on the cut-point.
B	See description under formula in the Regression Discontinuity Design section denoting a single participation indicator, *T*, coded as described in Criterion A.
C	The rdrobust package automatically creates a binary indicator denoting group assignment once the assignment variable is appropriately scaled and centered around zero.
D	The same vector of covariates a used in the estimation of the first stage estimation as well as the Fuzzy RD impact estimate. Both estimates are obtained through a single call to the rdrobust package.
E	Participation is defined in the same manner for students above and below the cut-point. There is no reason to believe that the position of a student relative to the cut-point would impact outcomes in any way other than by predicting receipt of ALFA Lab (after conditioning on the assignment variable; See Figures 4, 7, and 10 for graphical depictions of assignment score—outcome variable modeled relationships for STAR, GRM and FRQ outcomes, respectively)
F	See Exhibits B5.1, B5.2, and B5.3 for first-stage estimates related to the STAR, GRM and FRQ outcomes, respectively
G	See description under Table 1 in Regression Discontinuity Design
H	Not applicable

Appendix C: Mediation & Moderation Analyses

Researchers and policymakers are not only interested in determining "what works" but also would like to understand "for whom and under what circumstances" programs work (Dong et al., 2020) and "how or why" programs work (Kraemer et al., 2008; Roberts et al., 2018). Interest in this nuance is fueled by a need to identify interventions that might be more impactful for some subgroups of individuals and less so for others, ultimately allowing policymakers to efficiently allocate resources to accomplish the greatest good (Haegerich & Massetti, 2013; Supplee et al., 2013). Similarly, understanding how interventions make an impact is vital to advancing education science and refining theories of action that represent the link between taking a particular action (cause) and expecting a certain effect (Hopkins & Craig, 2014). This can be accomplished by determining if outcomes vary by level of program implementation (Bloom et al., 2003) or by explicitly modeling construct measurements purported to function as part of the theory of action. The What Works Clearinghouse (WWC) provides limited guidance on how to approach these analyses within a regression discontinuity (RD) context. Further, there are few examples of these analyses in the research literature. The following sections present preliminary mediation and moderation analyses designed to better understand how Accelerating Literacy for Adolescents (ALFA) Lab might be impacting students and who might be impacted more or less by exposure to ALFA Lab, respectively.

As documented in Appendix B.2, missing outcome data (particularly General Reading Motivation [GRM] and Reading Frequency [FRQ] scores) were also a problem when conducting mediation analysis within the same bandwidth identified in the main impact analysis using STAR scores. We utilized multiple imputation to generate "complete" data to be used in our mediation analyses. Results in C.1.1 suggest this approach was successful in maintaining the integrity of data distributions. Analyses exploring the mediating effect of GRM and FRQ on STAR scores did not yield any significant effects (see Exhibit C1.3.1). However, the proportion of the total effect mediated by these two measures was 0.30 and 0.22, respectively. Likely, the lack of significance is due to the relatively small sample size available (Fritz & MacKinnon, 2007). Analyses exploring whether ALFA Lab impacted subgroups of students differently regarding the STAR assessment yielded no significant effects, though some results did trend in favor of ALFA Lab (see the effects for male and Hispanic students in Exhibit C2.1.1). Results for the GRM scores also did not yield any significant results (Exhibit C2.1.2). Finally, analyses modeling FRQ scores as the outcome found significant effects among males, non-English Language Learner (ELL), Non-special education (SPED), and Hispanic students (see Exhibit C2.1.3).

C.1 Mediation of ALFA Lab Impact Estimate

C.1.1 Missing Data & Multiple Imputation

Unfortunately, as documented in Appendix B, attrition (defined as missing Adaptive Reading Motivation Measure [ARMM] GRM and FRQ outcomes) was a problem in this study. To accurately explore whether

either ARMM score acted as a mediator of the ALFA Lab impact on STAR reading achievement scores, we need to retain all records located within the bandwidth identified for the linear, triangular kernel weighted ITT impact model. ARMM GRM was missing at a rate of approximately 28% and the ARMM FRQ was missing at a rate of approximately 30%.[8] To overcome the sample size loss stemming from listwise deletion, multiple imputation procedures were used to impute missing ARMM scores across multiple datasets (Enders, 2010; van Buuren, 2012).

We made use of the *mice* R package (van Buuren & Groothuis-Oudshoorn, 2011) to conduct predictive mean matching multiple imputation of the ARMM GRM and FRQ scores. We used the STAR scores; the standardized assignment variable; cohort-by-site indicators; the demographic indicators for gender, ELL, SPED; and the indicators for all student race categories and gender.[9] A total of 30 imputations were generated and used in the estimation of the mediation models presented below. Figure C.1.1.1 shows the distribution of original (blue line) and imputed (red lines) values for the variables imputed. These graphs suggest the multiple imputation procedure yielded imputed values like those that were non-missing.

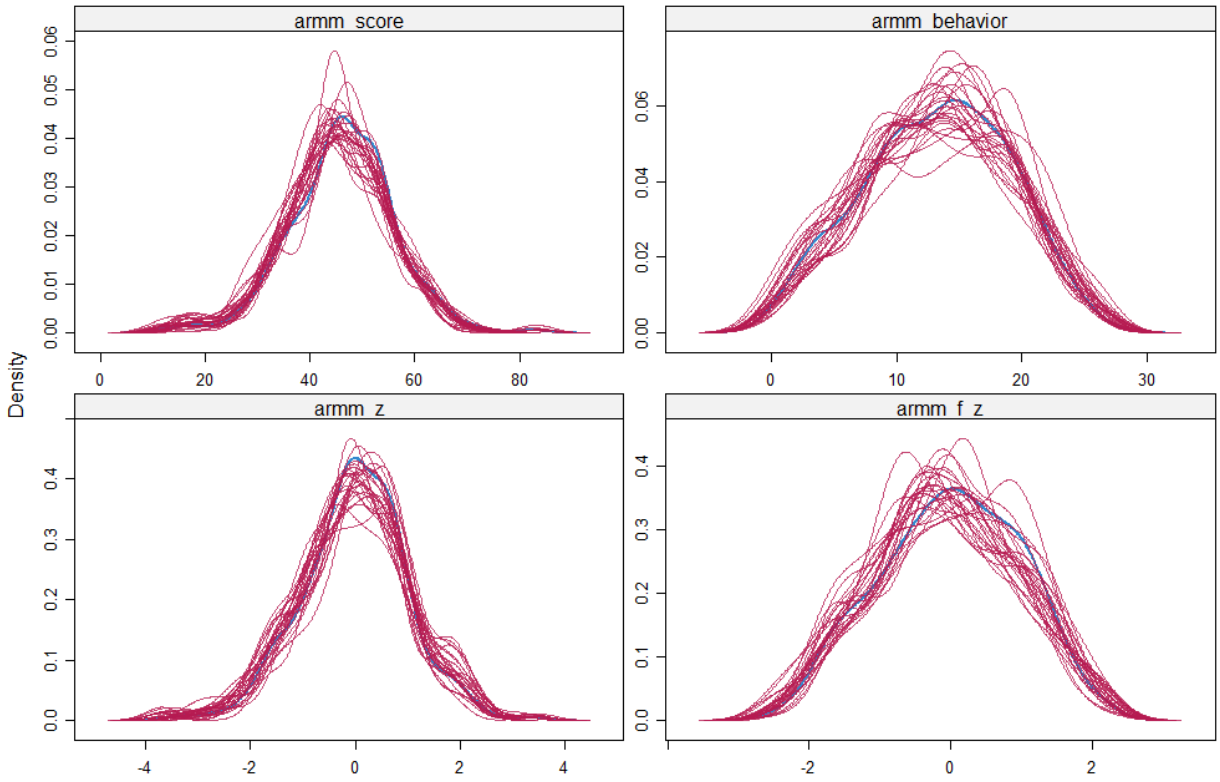


Figure C1.1.1. Density Plot of Imputed & Original Values

[8] These rates were calculated using the data analyzed to estimate the ALFA Lab intent-to-treat (ITT) impact estimate on STAR reading achievement scores, which had no missing STAR scores or covariate data points.
[9] The indicators for the racial category Hispanic and the indicator NY Site Cohort 1 were identified as collinear during the imputation process.

C.1.2 Exploring Mediation in RD Studies

Best practice in the analysis of data from RD designs suggests the use of a local polynomial regression within an empirically justified bandwidth *WWC™ Standards Handbook, Version 5.0*). This practice was followed in the current study. However, methods for conducting mediation within a local regression framework have not yet been developed. To explore mediation for the present study we proceed in an ad-hoc fashion as follows. First, we obtain the mean-squared error (MSE)-optimal bandwidth for estimating the total effect of ALFA Lab on the outcome using a triangular kernel and a linear model with different slopes allowed on either side of the cut-point. Next, using this bandwidth, we fit weighted OLS regressions (using weights determined by the triangular kernel) as described in equations (C1) and (C2).

$$M = i_M + a_1 X + a_2 W + a_3 XW + e_M \quad \text{(C1)}$$

$$Y = i_Y + c_1' X + c_2' W + c_3' XW + bM + e_Y \quad \text{(C2)}$$

In these equations M represents the mediator of interest, which is either the ARMM GRM or the ARMM FRQ score, Y represents the outcome of interest (STAR scores), X represents the assignment variable, and W represents the treatment indicator. As is crucial in an RD design, both equations control for the assignment variable and allow for the relationship between the assignment variable and the outcome to differ on either side of the cut-point.

Subject to the usual sequential ignorability assumptions associated with mediation analyses (Imai et al., 2010) the value of a_2 can be interpreted as the causal effect of ALFA Lab on the mediator (at the cut-point), the value of c_3' can be interpreted as the direct effect of ALFA Lab (its impact through paths other than the hypothesized mediator) at the cut-point and $a_2 b$ can be interpreted as the indirect effect of ALFA Lab (through the hypothesized mediator path). The total effect is represented by $c_2' + a_2 b$.

Each of the estimated effects are presented in a summary table, along with several effect size measures. The proportion of the total effect associated with the indirect effect is the first effect size, where $ACME$ is the average causal mediation effect, calculated simply as

$$ACME_{prop} = a_2 b \,/(a_2 b + c_2') \quad \text{(C3)}$$

The ratio of the indirect effect to the total effect is calculated as

$$ACME_{ratio} = a_2 b \,/c_2' \quad \text{(C4)}$$

G.1.3 Mediation Results

We utilized the *mediation* R package (Tingley et al., 2014) in conjunction with the results of the models described by equations (C1) and (C2) to perform inference about direct and indirect effects. For each of the 30 imputed data sets, we estimated the models represented in equations (C1) and (C2), submitted those results to the *mediations* function to obtain summarized mediation analysis results, and used Rubin's rules (Rubin, 1987) to summarize across the 30 sets of mediation results. Exhibit C1.3.1 shows the results from the mediation analysis exploring the mediation effect of the ARMM GRM and the FRQ scores. Looking at the top portion of the table, the ACME is the indirect effect of ALFA Lab on STAR

scores that go through GRM. The average direct effect (ADE) is the direct effect of ALFA Lab on STAR scores, controlling for GRM. The total effect is the sum of the direct and indirect effects on STAR scores. The proportion mediated represents the proportion of the effect of the assignment score-by-treatment interaction on STAR scores that go through the GRM (i.e., the ACME estimate divided by the total effect estimate).

Exhibit C1.3.1. Mediation Analysis Results

Mediator	Effect	Estimate	Lower CI	Upper CI	p-value
General Reading Motivation (GRM)	ACME	7.96	-9.20	24.95	0.36
	ADE	18.27	-55.08	91.28	0.62
	Total Effect	26.23	-47.36	99.49	0.48
	Proportion	0.30			
	Ratio	0.44			
Reading Frequency (FRQ)	ACME	5.81	-7.76	19.30	0.40
	ADE	20.43	-53.45	93.98	0.59
	Total Effect	26.24	-47.72	99.87	0.49
	Proportion	0.22			
	Ratio	0.28			

The estimated total effect of 26.23 is approximately equal to the 26.19 ITT STAR impact estimate. Here again, the total effect is not statistically significant (i.e., the confidence interval contains zero). Both the indirect effect (the portion of the total effect operating through GRM) and the direct effect were non-significant, with about 30% of the total effect associated with the indirect effect of the GRM based on the proportion effect size, while the ratio of the indirect effect to the direct effect was less than 1. Figure C.1.3.1 displays these results graphically in red.

Comparable results were seen when exploring the mediation effect associated with FRQ. The estimated total effect was 26.24, the indirect effect (the portion of the total effect operating through FRQ) was 5.81, and the direct effect of ALFA Lab was 20.43. All effects were non-significant, and the proportion of the total effect associated with the indirect effect of FRQ was 22% and the ratio was 0.28. Figure C.1.3.1 displays these results graphically in blue.

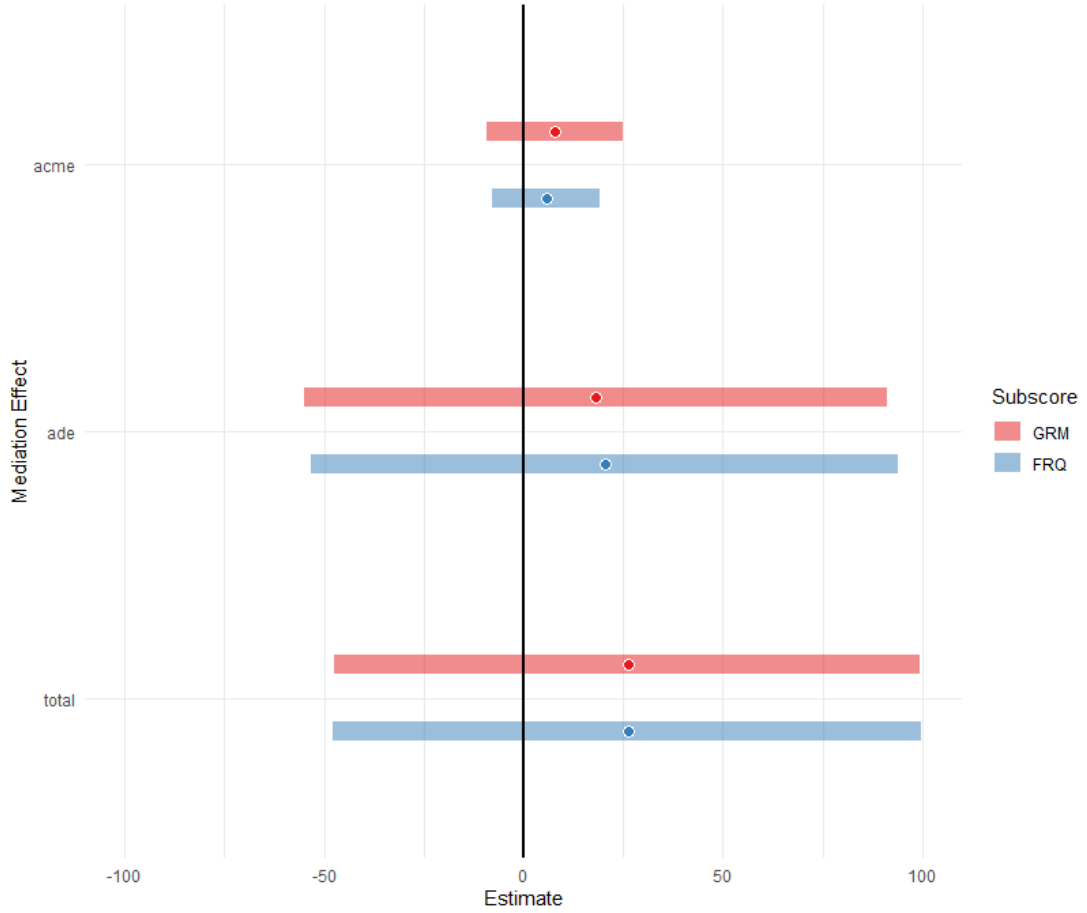


Figure C1.3.1. General Reading Motivation and Reading Frequency Mediator Effect Plot

C.2 Moderation of ALFA Lab Impact Estimates

C.2.1 Exploring Moderation in RD Studies

As was the case for mediation, local linear regression techniques for moderation analyses in RD have not been developed. Therefore, we once again utilize data within the MSE-optimal bandwidth for estimating the total effect of ALFA Lab on the outcome using a triangular kernel and a linear model with different slopes allowed on either side of the cut-point. Once this bandwidth is obtained, we fit a standard moderation model to data within the bandwidth, using triangular kernel weights and controlling for the assignment variable and its interaction with the treatment indicator.

Specifically, we estimate the model defined by equation (C7)

$$Y = i_Y + b_1X + b_2W + b_3Z + b_4XW + b_5WZ + e_Y \quad \text{(C7)}$$

where Z represents the moderating variable of interest and all other variable definitions were defined previously. Since all of our moderators are dummy (0-1) coded, b_2 can be interpreted as the average

causal effect of ALFA Lab at the cut-point for students whose value for Z is 0 and b_5 can be interpreted as the difference between the causal effect of ALFA Lab at the cut-point for students whose value for Z is 0 and the causal effect of ALFA Lab at the cut-point for students whose value for Z is 1. In other words, b_5 estimates the extent to which Z moderates the ALFA Lab effect.

Exhibit C2.1.1 displays the results of the moderator analyses conducted using STAR reading achievement scores as the outcome. Note that all other covariates were included in each analytic model when focusing on a moderator (i.e., when gender was the focal moderator; ELL, SPED, and race indicators; and cohort-by-site indicators were included in the model). When estimates are positive, this indicates that the effect within that moderator was in favor of ALFA Lab (i.e., the average outcome value for that moderator subgroup was larger for the ALFA Lab participants than for the business-as-usual [BAU] group). Conversely, when negative, the mean value in the BAU group was higher. None of the estimates were statistically significant, though the effect within the male subgroup yielded the lowest p-value (and a large effect size). Figures C2.1.1 through C2.1.4 display the RD moderator effect plots for gender, ELL status, SPED status, and race, respectively.

Exhibit C2.1.1. Star Moderation Analysis Results

Outcome	Moderator	Value	n (TX / CT)	Est	p-value	ES
STAR	**Gender**	Female	68 / 66	-7.55	0.86	-0.04
		Male	84 / 80	69.79	0.11	0.28
	ELL	Non-ELL	130 / 132	24.75	0.51	0.11
		ELL	22 / 14	41.52	0.55	0.22
	SPED	Non-SPED	149 / 144	26.37	0.48	0.12
		SPED	3 / 2	127.42	0.57	1.72
	Race	American Indian	4 / 3	-167.25	0.25	-0.59
		Asian	2 / 1	-591.16	0.02	NA
		African American	32 / 25	43.05	0.43	0.20
		DNR	4 / 2	-49.56	0.71	-0.18
		Hispanic	71 / 71	59.98	0.17	0.29
		Multi	3 / 4	-36.68	0.82	-0.11
		White	36 / 40	-2.61	0.96	-0.01

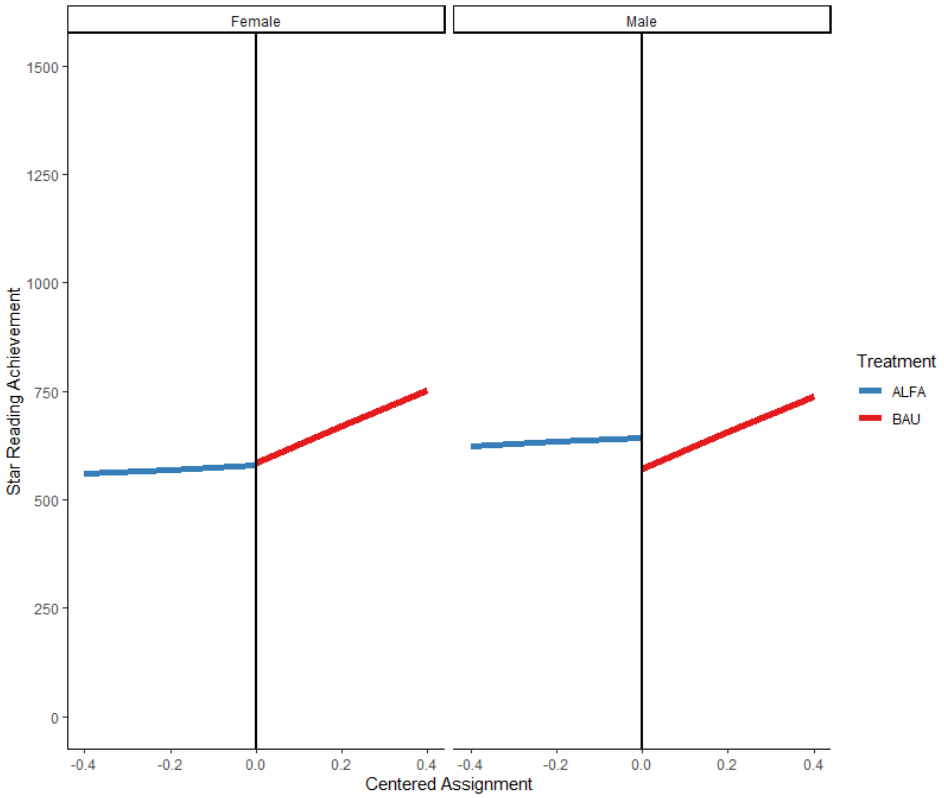


Figure C2.1.1. Student Gender Moderator Effect Plot – STAR Reading Achievement

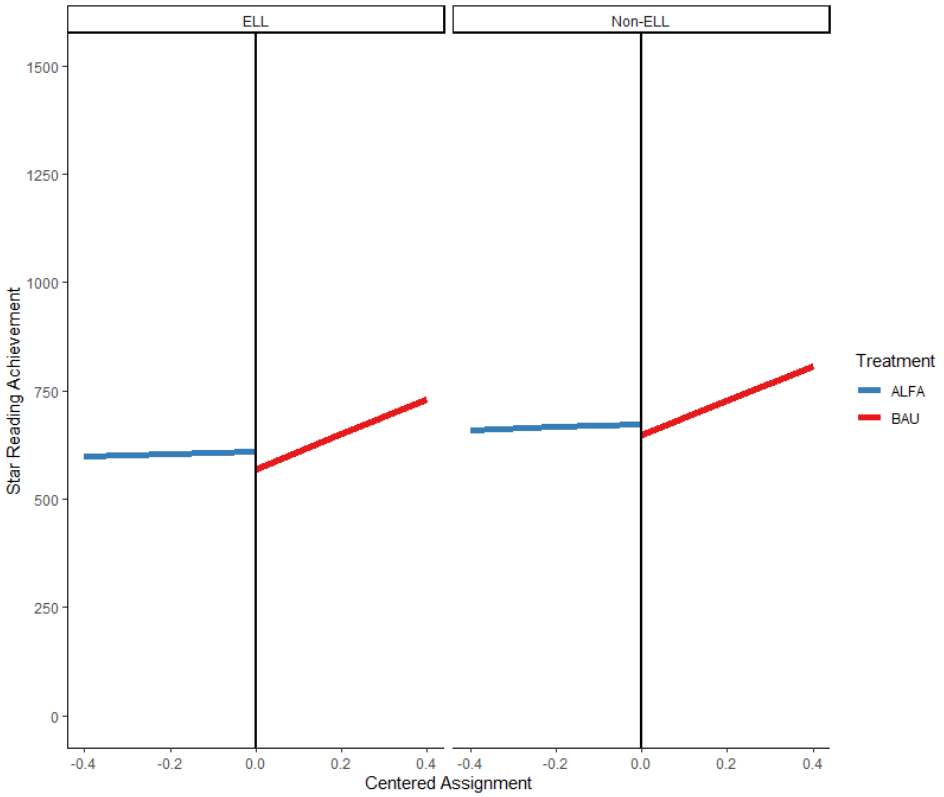


Figure C2.1.2. Student ELL Status Moderator Effect Plot – STAR Reading Achievement

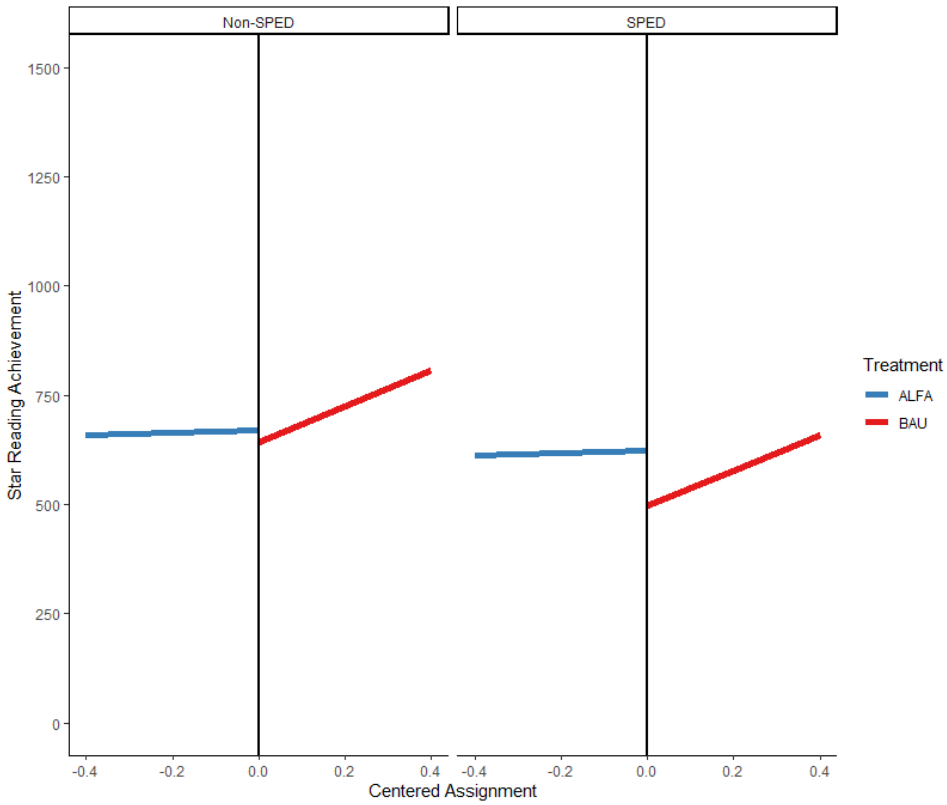


Figure C2.1.3. Student SPED Status Moderator Effect Plot – STAR Reading Achievement

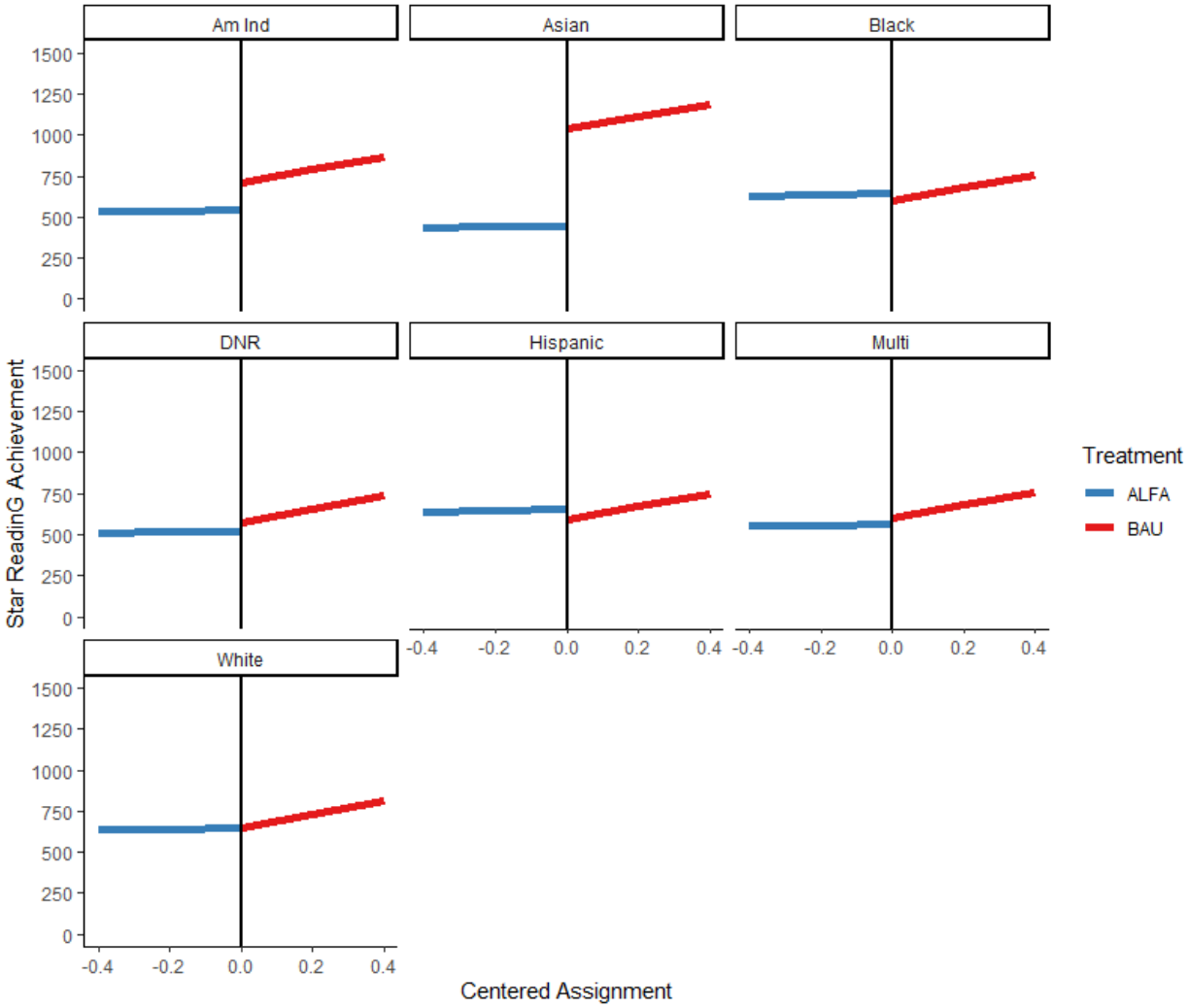


Figure C2.1.4. Student Race/Ethnicity Moderator Effect Plot – STAR Reading Achievement

Exhibit C2.1.2 displays the results of the moderator analyses conducted using GRM scores as the outcome. When estimates are positive, this indicates that the effect within that moderator was in favor of ALFA Lab (i.e., the average outcome value for that moderator subgroup was larger for the ALFA Lab participants than for the BAU group). Conversely, when negative, the mean value in the BAU group was higher. None of the estimates were statistically significant. Figures C2.1.5 through C2.1.8 display the GRM RD moderator effect plots for gender, ELL status, SPED status, and race, respectively.

Exhibit C2.1.2. General Reading Motivation Moderation Analysis Results

Outcome	Moderator	Value	n (TX / CT)	Est	p-value	ES
GRM	**Gender**	Female	32 / 97	0.67	0.77	0.07
		Male	41 / 127	1.40	0.53	0.14
	ELL	Non-Ell	62 / 203	0.75	0.70	0.08
		ELL	11 / 21	4.42	0.25	0.52
	SPED	Non-SPED	70 / 220	1.03	0.59	0.11
		SPED	3 / 4	2.47	0.76	0.60
	Race	Amer Indian	1 / 7	5.15	0.67	NA
		Asian	1 / 2	-13.94	0.14	NA
		African American	15 / 32	1.46	0.60	0.14
		DNR	1 / 2	4.32	0.61	NA
		Hispanic	34 / 115	0.68	0.77	0.08
		Multi*	-	-	-	-
		White	21 / 59	1.82	0.50	0.15

*No multiracial students in the treatment group within bandwidth.

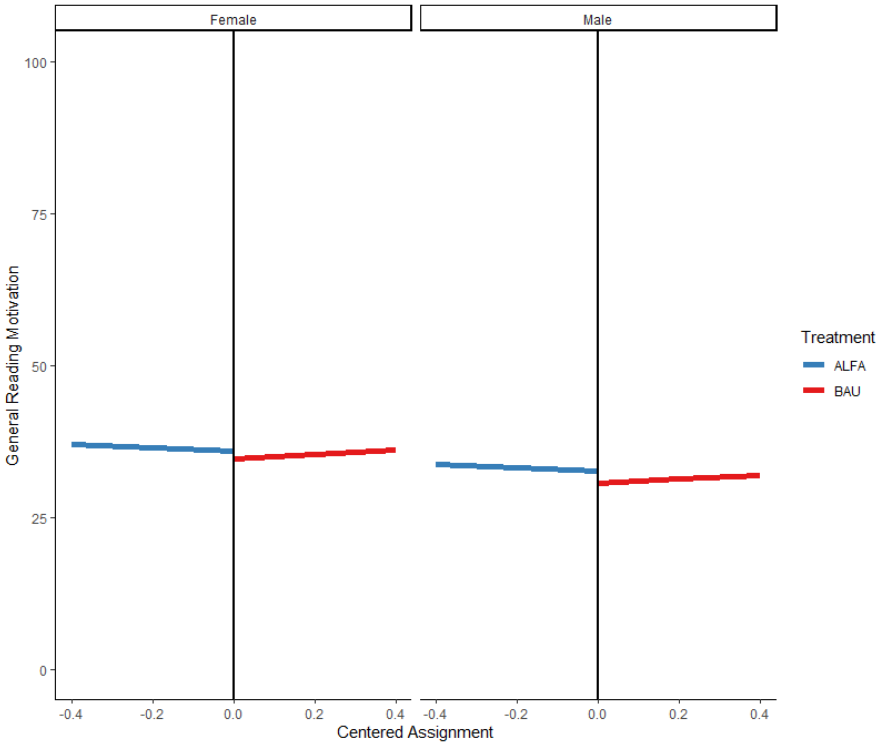


Figure C2.1.5. Student Gender Moderator Effect Plot – GRM

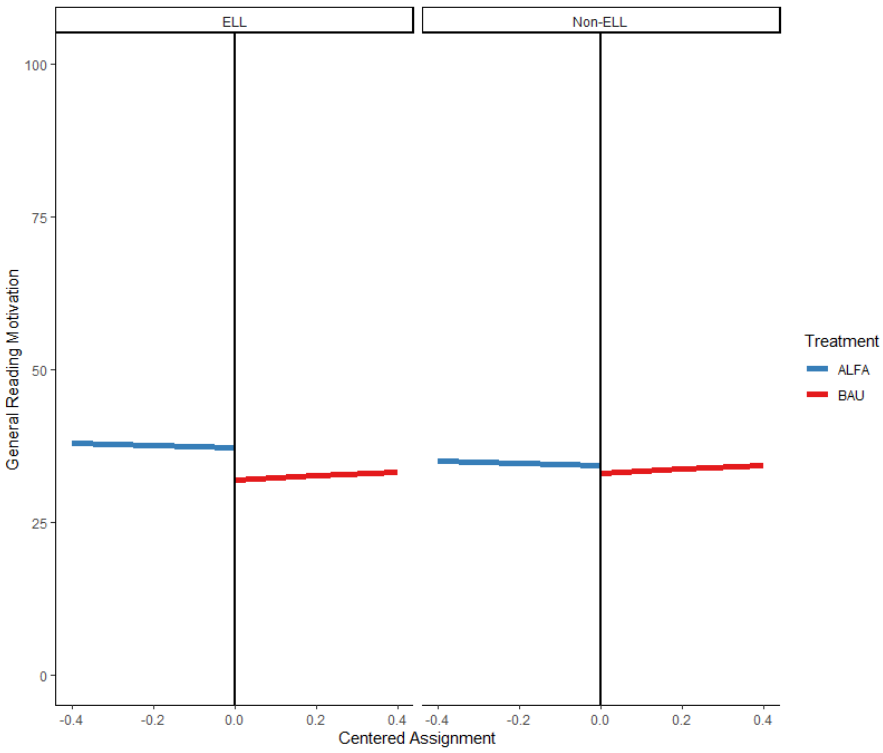


Figure C2.1.6. Student ELL Status Moderator Effect Plot – GRM

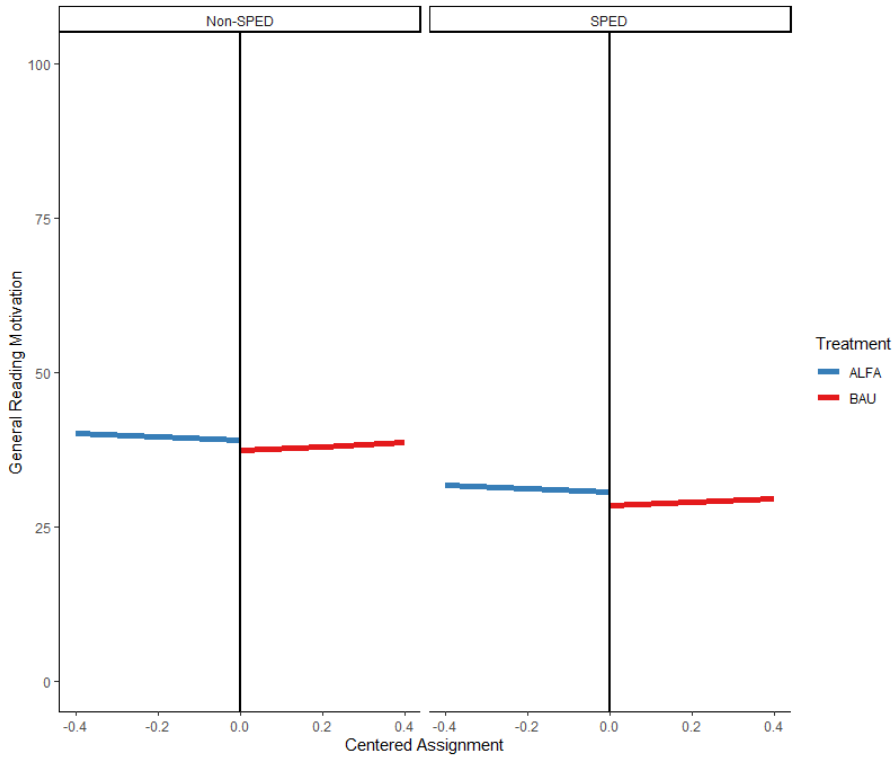


Figure C2.1.7. Student SPED Status Moderator Effect Plot – GRM

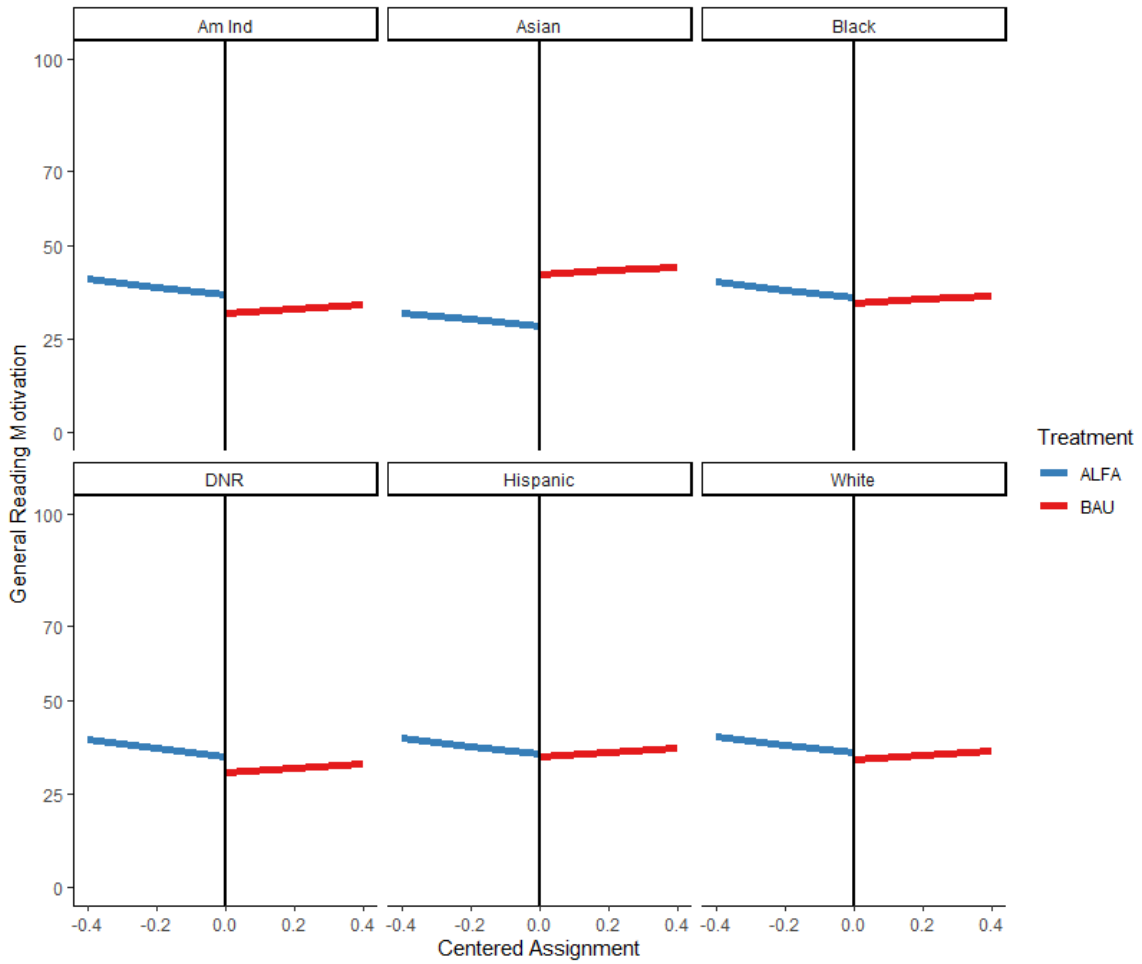


Figure C2.1.8. Student Race/Ethnicity Moderator Effect Plot – GRM

Exhibit C2.1.3 displays the results of the moderator analyses conducted using Reading Frequency scores as the outcome. The effect associated with Males was statistically significant, meaning that ALFA Lab males had higher FRQ scores than BAU males. Similar results were seen with the Non-ELL, Non-SPED, and Hispanic subgroups: in each, the ALFA Lab group had higher mean FRQ scores than the BAU group. Figures C2.1.9 through C2.1.12 display the FRQ RD moderator effect plots for gender, ELL status, SPED status, and race, respectively.

Exhibit C2.1.3. Reading Frequency Moderation Analysis Results

Outcome	Moderator	Value	n (TX / CT)	Est	p-value	ES
FRQ	**Gender**	Female	54 / 98	1.77	0.17	0.30
		Male	83 / 127	3.21	0.01	0.55
	ELL	Non-Ell	115 / 204	2.41	0.03	0.41
		ELL	22 / 21	3.73	0.08	0.69
	SPED	Non-SPED	130 / 220	2.49	0.02	0.42
		SPED	7 / 5	3.51	0.45	0.77
	Race	American Indian	4 / 7	7.04	0.13	0.98
		Asian	1 / 2	-8.98	0.13	NA
		African American	25 / 31	1.46	0.37	0.26
		DNR	2 / 2	1.02	0.84	0.10
		Hispanic	69 / 116	3.81	0.00	0.66
		Multi	2 / 7	-3.12	0.66	-0.42
		White	34 / 60	1.62	0.29	0.27

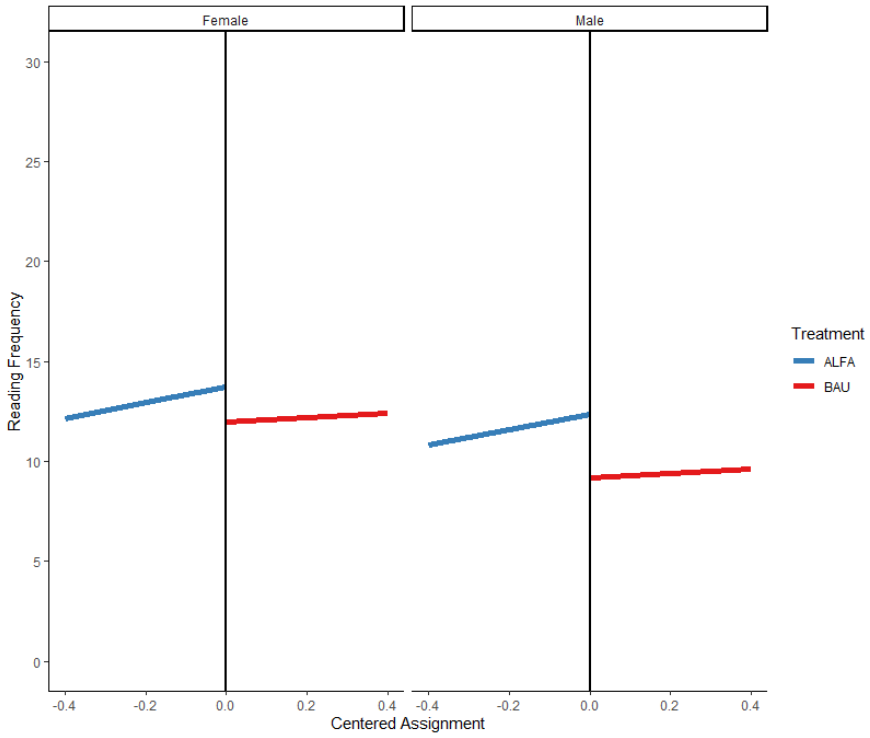


Figure C2.1.9. Student Gender Moderator Effect Plot – FRQ

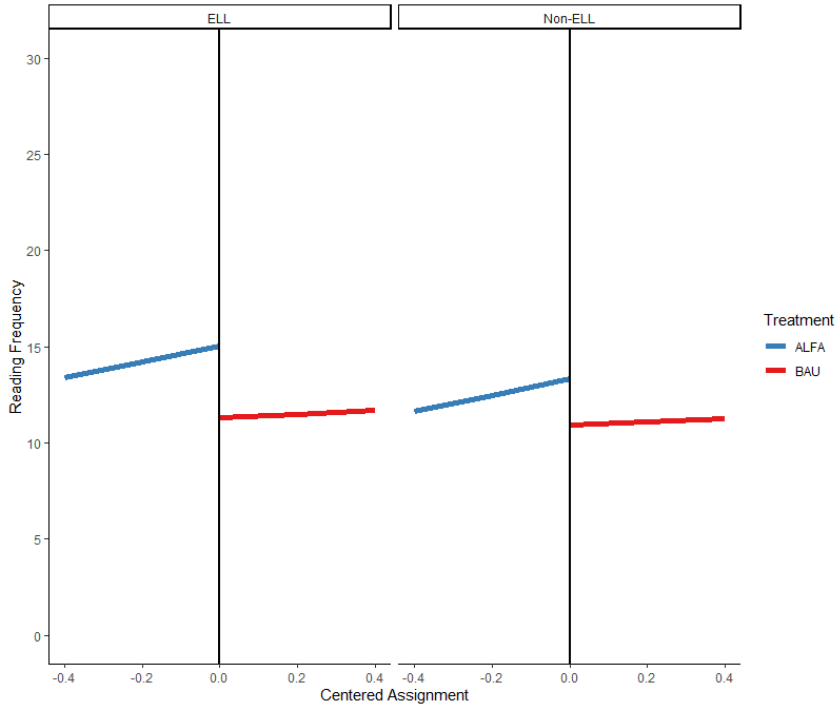


Figure C2.1.10. Student ELL Status Moderator Effect Plot – FRQ

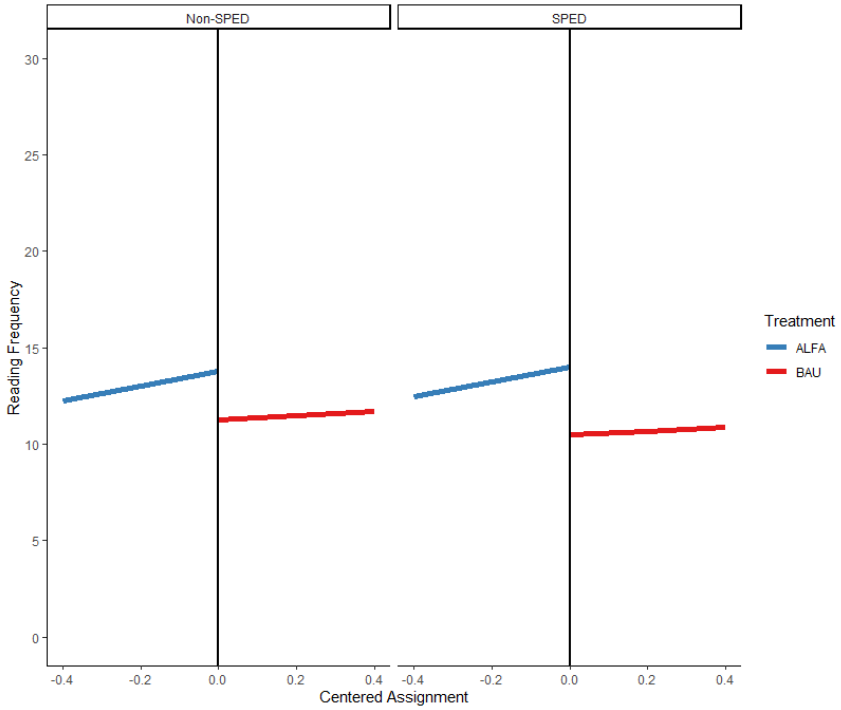


Figure C2.1.11. Student SPED Status Moderator Effect Plot – FRQ

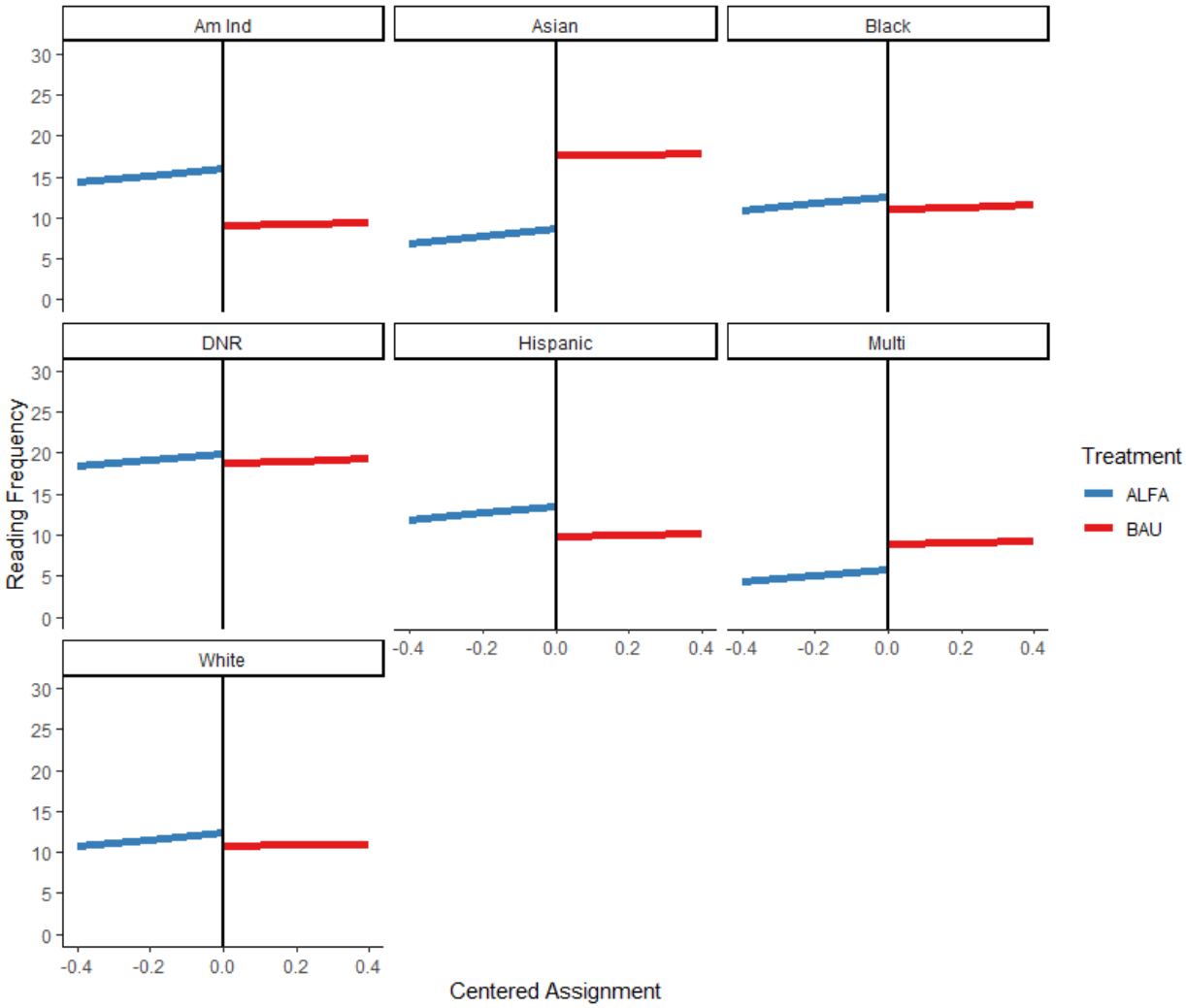


Figure C2.1.12. Student Race/Ethnicity Moderator Effect Plot – FRQ

C.3 Mediation/Moderation Limitations

Making use of the bandwidths identified during the estimation of the ITT impact estimates presents a limitation for the mediation/moderation analyses. Using those bandwidths ignores the trivariate relationship that exists between the assignment variable, outcome score, and mediator/moderator in question. Purportedly, including the variability of a third variable would identify a different data-driven bandwidth that would balance the bias-variance tradeoff associated with the location of bandwidths along the assignment score continuum.

Appendix D: ALFA Teacher Survey Data

We present responses to surveys of ALFA teachers from three cohorts. Due to the disruption of the COVID-19 pandemic, we administered a shortened survey for Cohort 2 teachers, omitting items not applicable to instruction in virtual settings.

Background Questions

Table D1. Respondents by School

Cohort	School	Respondents
Cohort 1	School 1 (NY)	3
	School 3 (CA)	1
Cohort 2	School 1 (NY)	2
	School 3 (CA)	2
Cohort 3	School 1 (NY)	2
	School 3 (CA)	1
	School 2 (CT)	2
	School 4 (AL)	1

Table D2. Length of ALFA Lab Classes

Cohort	Length of ALFA Lab Class(es)	
Cohort 1 (n=4)	The ALFA Lab class is approximately 45–55 minutes long, more or less (like a regular class period)	25%
	The ALFA Lab class is approximately 70–90 minutes long, more or less (like an extended or block-scheduled class period)	75%
	Other	0%
Cohort 2 (n=4)	The ALFA Lab class is approximately 45–55 minutes long, more or less (like a regular class period)	0%
	The ALFA Lab class is approximately 70–90 minutes long, more or less (like an extended or block-scheduled class period)	50%
	Other	50%
Cohort 3 (n=6)	The ALFA Lab class is approximately 45–55 minutes long, more or less (like a regular class period)	66.7%
	The ALFA Lab class is approximately 70–90 minutes long, more or less (like an extended or block-scheduled class period)	33.3%
	Other	0%

Table D3. ALFA Lab Start at the Beginning of ALFA Lab

Cohort	Did the ALFA Lab start at the beginning of the school year?	
Cohort 1 (n=4)	Yes, it started along with other classes at the start of the school year.	100%
	No, there were delays.	0%
Cohort 2 (n=4)	Yes, it started along with other classes at the start of the school year.	100%
	No, there were delays.	0%
Cohort 3 (n=6)	Yes, it started along with other classes at the start of the school year.	100%
	No, there were delays.	0%

Table D4. More than One ALFA Lab Sections Per School

Cohort	Is there more than one ALFA Lab class section at your school?	
Cohort 1 (n=4)	Yes	100%
	No	0%
Cohort 2 (n=4)	Yes	100%
	No	0%
Cohort 3 (n=6)	Yes	50%
	No	50%

Table D5. Average Number of ALFA Lab Sections

Cohort	How many ALFA Lab class sections are there at your school?
Cohort 1 (n=4)	5
Cohort 2 (n=4)	4
Cohort 3 (n=3)	3

Note. This question was presented to only those respondents who noted that there was more than one ALFA Lab section at their school.

Table D6. Average Number of Students Served in ALFA Lab Class Section

Cohort	On average, how many students are served in each ALFA Lab class section?
Cohort 1 (n=4)	22
Cohort 2 (n=4)	19
Cohort 3 (n=6)	22

Note. This question was presented in a different format for Cohort 2 respondents.

Table D7. Variation of Class Size Over Time

Cohort	Do student numbers vary over time?	
Cohort 1 (n=4)	Yes	50%
	No	50%
Cohort 2 (n=4)	Yes	100%
	No	0%
Cohort 3 (n=6)	Yes	50%
	No	50%

Table D8. Variation of Class Size Over Time – Explanation

Cohort	Please explain your answer in the field below.
Cohort 1 (n=2)	We have had a couple of students check out of school or we have switched their class period. At one point we had about 24 per class, but now they are at 20, 22, and 23.
	More students were added due to their 7th grade scores.
Cohort 2 (n=4)	So[me] are misplaced initially and are removed at a later time!
	Right now, my sections are 15, 18, and 19 students. Sometimes I have up to 24 students (last year), but this year has been 15–20 each section.
	Some students are removed or added due to data!
	It depends on the semester as well as the period. This year they've mostly been under 20 but at least 15. Last year they were mostly 20–24.
Cohort 3 (n=3)	New additions or students might be taken out for another course.
	Students are changed due to scheduling issues.
	Usually 24 are placed into my class, but some move or switch schools.

Note. This question was presented to only those respondents who noted that ALFA Lab class sizes varied over time.

Table D9. Implementation Frequency of ALFA Components

Frequency in which following activities were implemented:						
Cohort	**Activity**	**Never**	**About once or twice per week**	**About once per week**	**Nearly every ALFA Lab class meeting**	**Every ALFA Lab class meeting**
Cohort 1 (n=4)	The Daily Launch	0%	0%	0%	0%	100%
	The Stations (Main Station, Wordology, Collaboration Station, Media Madness)	0%	0%	0%	0%	100%
Cohort 2 (n=4)	The Daily Launch	0%	25%	0%	50%	25%
	The Stations (Main Station, Wordology, Collaboration Station, Media Madness)	0%	0%	0%	75%	25%
Cohort 3 (n=6)	The Daily Launch	0%	0%	0%	33.3%	66.7%
	The Stations (Main Station, Wordology, Collaboration Station, Media Madness)	0%	0%	0%	33.3%	66.7%

Table D10. Reason(s) for Not Implementing ALFA Lab Activities Every Class Meeting

Cohort	Please explain your answer(s) in the field below:
Cohort 2 (n=4)	Because of the time constraints (37 minute periods), on the first day of rotation, I do Main Station (vocab and reading strategies). Then I give them time (if there even is time, since vocab takes longer online) to start working on their written work. If I feel there is time and/or it is relevant, I'll show a video or discuss with them the daily launch. I also go over all the pages they'll be completing for that two-day rotation. On the 2nd day of rotation, they have all class period to complete the written assignments. Unfortunately, because some of them are refusing to do work (not just in my class, but all their classes), it doesn't really work to put them into groups. We stay online together in order for them to ask me for help if they need it. I also do not have an instructional assistant this year, which is why I can't teach Main Station in small groups either. Our district doesn't allow the option to NOT be live with our students, so I can't even rotate the days I see the students to create smaller groups.
	Because of distance learning, the set up has been very different. The students do the work for each station, but we don't have enough time to implement the exact way it should be in live class.
	Daily Launch is the initial activity, but the other stations are often done every other day.
	We tried to implement every component of the lab as specified by the manual.
Cohort 3 (n=2)	Daily Launch is daily, but Wordology, Collaboration and Media Madness are every other day mostly! There are times when all stations are used.
	There were times when a station would take more time than allotted, so we would spread the session out into 2 days.

Note. This question was presented to only those respondents who did not select “every ALFA Lab class meeting” in describing the frequency with which ALFA Lab activities were implemented.

Table D11. Time Devoted to ALFA Lab Components (Standard Bell Schedule)

	Approximate time devoted to the following activities:					
Cohort	**Activity**	**Fewer than 5 minutes**	**About 5 minutes**	**About 10 minutes**	**About 20 minutes**	**About 30 minutes**
Cohort 1 (n=1)	The Daily Launch	0%	100%	0%	0%	0%
	EACH station	0%	100%	0%	0%	0%
Cohort 2 (n=2)	The Daily Launch	0%	0%	100%	0%	0%
	EACH station	0%	0%	0%	100%	0%
Cohort 3 (n=4)	The Daily Launch	0%	0%	100%	0%	0%
	EACH station	0%	0%	0%	100%	0%

Note. This question was presented to only those respondents who indicated that their ALFA Lab class lasted 45–55 minutes.

Table D12. Time Devoted to ALFA Lab Components (Extended Bell Schedule)

	Approximate time devoted to the following activities:					
Cohort	**Activity**	**Fewer than 10 minutes**	**About 10 minutes**	**About 20 minutes**	**About 30 minutes**	**About 45 minutes**
Cohort 1 (n=3)	The Daily Launch	33%	0%	67%	0%	0%
	EACH station	0%	0%	33%	67%	0%
Cohort 3 (n=2)	The Daily Launch	0%	50%	50%	0%	0%
	EACH station	0%	0%	50%	50%	0%

Note. This question was presented to only those respondents who indicated that their ALFA Lab class lasted 70–90 minutes.

Table D13. Time Devoted to ALFA Lab Components – Explanation for Deviations from Standard Time Allocations

Cohort	Please explain your answer(s) in the field below.
Cohort 1 (n=3)	We are told that Launches should take about 10 minutes. Some administrators say '5' minutes for a 'Do Now'! However, with this population of students that are reading far below their grade level, everything is slow. Their initiative to start. This means even writing their name on their work. Reading instructions has not become an academic habit, so we have had a lesson on: 'why' you 'must' read instructions, and then, how to understand key words and phrase for what to do for the activity. I have learned as a teacher, I cannot assume anything regarding the children's knowledge base. This experience is very shocking, but it raises the importance of intervention on the high school level. Some teachers and administrators don't believe it is necessary.
	Daily launch can vary from 3–8 minute. Then we use the rest of our 53-minute period to rotate to two stations daily.
	Sometimes 2 rotations; sometimes 4 rotations.
Cohort 2 (n=2)	There were times where students needed more time to complete the assignments, so they were afforded extra time.

Note. This question was presented to only those respondents who did not select standard time allocations for ALFA Lab components.

Table D14. Types of ALFA Lab Stations Implemented in Last Month

	Types of ALFA Lab Stations implemented in the last month:					
Cohort	**Main Station**	**Collaboration Station**	**Wordology**	**Media Madness**	**None**	**Other**
Cohort 1 (n=4)	100%	100%	100%	100%	0%	20% (Lending Library)
Cohort 2 (n=4)	75%	75%	75%	75%	25%	25%
Cohort 3 (n=6)	83.3%	83.3%	83.3%	83.3%	16.7%	0%

Table D15. Types of ALFA Lab Stations Implemented in Last Month—Explanation

Cohort	Please explain your answer(s) in the field below.
Cohort 1 (n=1)	Students need to read much more. The habit, desire and practice of reading is so important. We have added an independent reading station because it is so necessary for kids that have their heads stuck in a phone texting or watching a video or glued to social media. Plus, the VALUE of reading just to ENJOY reading has been lost to most of our students.
Cohort 2 (n=1)	The course was completed last semester. This is a new semester. ALFA LABS are finished. Students are in English 9.
Cohort 3 (n=1)	Semester changes alfa lab finished and English 9 began.

Note. This question was presented to only those respondents who selected “None” or “Other” in referencing the types of ALFA Lab stations implemented in the last month.

ALFA Lab Program Materials

Table D16. Types of Materials in ALFA Lab

	What types of materials have you used for the ALFA Lab? How many of each type of material were required? Did materials need to be replaced?			
		Yes, used in the ALFA Lab	**# of copies required**	**Yes, need to be replaced**
Cohort 1 (n=4)	ALFA Lab's Student Worksheets	75%	70, 72	75%
	ALFA Lab Teacher's Handbook	75%	2, 3	25%
	Feisty Felines Teacher's Manual	75%	2, 3	25%
	The Set of Feisty Felines Reading Selections	75%	25, 30	25%
	Heroes Teacher's Manual	75%	2, 3	25%
	The Set of Heroes Reading Selections	75%	25,30	25%
	Galaxy Teacher's Manual	75%	2, 3	25%
	The Set of Galaxy Reading Selections	75%	25, 30	25%
	Computers (desktop)	50%	0, 3	25%
	Computers (laptop)	50%	0, 7	25%
	Electronic tablets (e.g., Kindles, iPads)	50%	24	25%
	Printers	50%	0, 1	25%
	Projectors	50%	1	0%
	Other (please describe in the space below)	25%	1	0%
Cohort 3 (n=6)	ALFA Lab's Student Worksheets	100%	23, 24, 30, 70, 72	66.7%
	ALFA Lab Teacher's Handbook	83.3%	2, 4	16.7%
	Feisty Felines Teacher's Manual	100%	2, 4	16.7%
	The Set of Feisty Felines Reading Selections	100%	24, 26, 30, 70	33.3%
	Heroes Teacher's Manual	100%	2, 4	16.7%
	The Set of Heroes Reading Selections	100%	24, 26, 30, 70	50%
	Galaxy Teacher's Manual	66.7%	2, 70	16.7%
	The Set of Galaxy Reading Selections	66.7%	26, 30, 70	33.3%
	Computers (desktop)	33.3%	2	0%
	Computers (laptop)	66.7%	23, 6	16.7%
	Electronic tablets (e.g., Kindles, iPads)	16.7%	0, 24	0%
	Printers	33.3%	1	0%
	Projectors	50%	0	0%
	Other (please describe in the space below)	33.3%	23, 24	16.7%

Note. This question was not presented to respondents in Cohort 2.

Table D17. Other Equipment Used

Cohort	If you selected Other in the previous question, please describe what those materials were.
Cohort 1 (n=1)	I used a small rolling [board] that I purchased last year. This board was very helpful for Main Station instruction. I think each station needs one.
Cohort 3 (n=2)	Google Chromebooks
	Student Notebooks

Note. This question was presented to only those respondents who selected "Other" in referencing the types of ALFA Lab materials used. This question was not presented to respondents in Cohort 2.

Table D18. Use and Location of Computers

Cohort	If you identified that desktop computers, laptops, or tablets were used in the ALFA Lab, where were they located?				
	In a computer lab	They reside in the ALFA Lab	They were brought into the ALFA classroom (e.g., on a portable cart)	In another location	N/A; computers were not used in the ALFA Lab
Cohort 1 (n=4)	0%	100%	0%	0%	0%
Cohort 3 (n=6)	0%	66.7%	0%	33.3%	0%

Note. This question was not presented to respondents in Cohort 2.

Table D19. Use of Internet Connection

Cohort	If you identified that desktop computers, laptops, or tablets were used in the ALFA Lab, was an internet connection needed?	
Cohort 1 (n=4)	Yes, an internet connection was needed.	100%
	No, an internet connection was not needed.	0%
	N/A; computers were not used in the ALFA Lab.	0%
Cohort 3 (n=6)	Yes, an internet connection was needed.	100%
	No, an internet connection was not needed.	0%
	N/A; computers were not used in the ALFA Lab.	0%

Note. This question was not presented to respondents in Cohort 2.

Table D20. Use of Donated Materials

Cohort	What donated materials were used in the program? (Materials not purchased by the school.) Please describe them in the field below.
Cohort 1 (n=3)	All materials
	I had to purchase three new head phones. One was stolen, one was damaged and I bought an extra one as a replacement.
	Books, copies, binder, tabs, art supplies, games.
Cohort 3 (n=3)	A larger variety of word-based games
	All materials were granted.
	Earbuds for each student

Note. This question was not presented to respondents in Cohort 2.

Perceptions of the ALFA Lab

Table D21. Helpfulness of ALFA Lab Elements in Improving Student Reading

Cohort	How would you rate the following ALFA Lab elements in helping to improve student reading?					
		Not at all helpful	Slightly helpful	Helpful	Very Helpful	Mean Score (1–4)
Cohort 1	The overall ALFA Lab design (n=3)	0%	0%	100%	0%	3.0
	Daily Launch (n=3)	0%	0%	100%	0%	3.0
	Main Station (n=3)	0%	0%	67%	33%	3.3
	Collaboration Station (n=3)	0%	0%	67%	33%	3.3
	Wordology (n=3)	0%	0%	67%	33%	3.3
	Media Madness (n=3)	0%	0%	67%	33%	3.3
Cohort 3	The overall ALFA Lab design (n=5)	0%	20%	60%	20%	3.00
	Daily Launch (n=6)	0%	16.7%	50%	33.3%	3.17
	Main Station (n=6)	0%	0%	66.7%	33.3%	3.33
	Collaboration Station (n=6)	0%	16.7%	50%	33.3%	3.17
	Wordology (n=6)	0%	33.3%	33.3%	33.3%	3.00
	Media Madness (n=6)	0%	16.7%	50%	33.3%	3.17

Note. For these items, the mean was calculated based on the assigned values for each of the helpfulness response options: 1 – *Not at all helpful*, 2 – *Slightly helpful*, 3 – *Helpful*, 4 – *Very helpful*.

This set of items were not presented to respondents in Cohort 2.

Table D22. Helpfulness of ALFA Lab in Promoting Student Outcomes

Cohort	How helpful is the ALFA Lab is in promoting the following student outcomes?	Not at all helpful	Slightly helpful	Helpful	Very helpful	Mean Score (1–4)
Cohort 1 (n=3)	***Student Engagement Outcomes***					
	Active listening	0%	33%	67%	0%	2.7
	Participating in discussion	0%	33%	67%	0%	2.7
	Staying on task	0%	33%	67%	0%	2.7
	Student Learning Outcomes					
	Reading comprehension	0%	33%	67%	0%	2.7
	Reading fluency	0%	33%	67%	0%	2.7
	Vocabulary development	0%	0%	67%	33%	3.3
	Writing skills	0%	0%	100%	0%	3.0
	Student Disposition Outcomes					
	Reading motivation	0%	33%	67%	0%	2.7
	Reading self-efficacy and belief	0%	33%	67%	0%	2.7
Cohort 3 (n=6)	***Student Engagement Outcomes***					
	Active listening (n=5)	0%	25%	75%	0%	2.6
	Participating in discussion (n=4)	0%	0%	25%	75%	3.75
	Staying on task (n=4)	25%	25%	25%	25%	2.5
	Student Learning Outcomes					
	Reading comprehension (n=6)	0%	0%	66.7%	33.3%	3.33
	Reading fluency (n=6)	16.7%	16.7%	66.7%	0%	2.5
	Vocabulary development (n=6)	0.0%	16.7%	50%	33.3%	3.17
	Writing skills (n=6)	0%	66.7%	16.7%	16.7%	2.5
	Student Disposition Outcomes					
	Reading motivation (n=6)	16.7%	16.7%	50%	16.7%	2.67
	Reading self-efficacy and belief (n=6)	16.7%	16.7%	66.7%	0%	2.5

Note. For these items, the mean was calculated based on the assigned values for each of the helpfulness response options: 1 – *Not at all helpful*, 2 – *Slightly helpful*, 3 – *Helpful*, 4 – *Very helpful*. This set of items were not presented to respondents in Cohort 2.

TTable D23. Satisfaction with ALFA Lab

	How satisfied are you with the ALFA Lab intervention, supplies, and support?						
Cohort		**Strongly Dissatisfied**	**Dissatisfied**	**Neutral**	**Satisfied**	**Strongly Satisfied**	**Mean Score (1–5)**
Cohort 1 (n=3)	The overall ALFA Lab reading intervention	0%	0%	33%	67%	0%	3.7
	Teacher's handbook and unit materials	0%	0%	0%	100%	0%	4.0
	Instructional materials, support materials, and supplies	0%	0%	0%	67%	33%	4.3
	Summer training	0%	33%	33%	0%	33%	3.3
	Monthly coaching from JHU or TDS	0%	33%	67%	0%	0%	2.7
	Support from the school's own faculty (e.g., literacy leader/specialist/ department head)	0%	0%	33%	67%	0%	3.7
Cohort 3 (n=5)	The overall ALFA Lab reading intervention	0.0%	20%	0.0%	60%	20%	3.80
	Teacher's handbook and unit materials	0.0%	20%	0.0%	60%	20%	3.80
	Instructional materials, support materials, and supplies	0%	0%	0%	80%	20%	4.20
	Summer training	0%	0%	0%	60%	40%	4.40
	Monthly coaching from JHU or TDS	0%	0%	80%	0%	20%	3.40
	Support from the school's own faculty (e.g., literacy leader/specialist/ department head)	0%	0%	60%	20%	20%	3.60

Note. Some items do not total 100% because of rounding. The mean for these items was calculated based on the assigned values for each of the satisfaction response options: 1 – *Strongly dissatisfied*, 2 – *Dissatisfied*, 3 – *Neutral*, 4 – *Satisfied,* 5 – *Strongly satisfied*. This set of items were not presented to respondents in Cohort 2.

Table D24. Teacher Perceptions on ALFA Lab Themes/Units Liked by Students

	Please rate the degree to which your students liked the following ALFA Lab themes.						
Cohort		**Strongly Disliked**	**Disliked**	**Neutral**	**Liked**	**Strongly Liked**	**Mean Score (1–5)**
Cohort 1	Big Cats (n=3)	0%	0%	33%	67%	0%	3.7
	Heroes (n=3)	0%	0%	0%	67%	33%	4.3
	Space (n=3)	0%	0%	0%	33%	67%	4.7
Cohort 3	Big Cats (n=6)	16.7%	16.7%	16.7%	33.3%	16.7%	3.17
	Heroes (n=6)	0%	0%	16.7%	50%	33.3%	4.17
	Space (n=4)	0%	0%	25%	75%	0%	3.75

Note. For these items, the mean was calculated based on the assigned values for each of the likeability response options: 1 – *Strongly disliked*, 2 – *Disliked*, 3 – *Neutral*, 4 – *Liked,* 5 – *Strongly Liked*.

This set of items were not presented to respondents in Cohort 2.

Table D25. Respondent Comments on What Works Best in ALFA Lab

Cohort	**In the field below, please share any other feedback you have about what works BEST in the ALFA Lab.**
Cohort 1 (n=2)	Media Madness was interactive, but some videos didn't work! Wordology was great!
	Most of the students liked the Big Cats at first, some began asking to do something else. I think the program should include other endangered species, brown and polar bears, monarch butterflies, blue whales, etc. to expand student knowledge and interest. Although I feel many of the students that did complain lacked the discipline to stay focused on a topic for a period of time, perhaps the program can add more variety to this unit by including other animals. The best part of the program is the station routine and the quality of the reading materials. They are colorful, interesting and topics connect to adolescents. The rotation of the stations does push students to stay on task, but still students need more time for quality completion of work. The independent Media Station is good for this generation of students. It must be closely monitored and sometimes adjustments of the video link have to be made by the teacher.
Cohort 3 (n=4)	Feisty Felines is too long!
	Having students use earbuds to listen to the media madness videos decreased the amount of commotion and noise.
	I think Main Station is very beneficial, as many students rarely get the opportunity to work with their teachers in small groups. I think being able to have selections read to them and then discuss them in that setting is helpful.
	Support from [JHU] along with willingness to supply other items needed.

Note. This question was not presented to respondents in Cohort 2.

Table D26. Respondent Comments on What Works Least in ALFA Lab

Cohort	In the field below, please share any other feedback you have about what works LEAST in the ALFA Lab.
Cohort 1 (n=2)	ALFA needs a reading workbook to help students practice the reading skills more and the program needs a writing practice and grammar book to help students practice these skills more. The Wordology worksheets provide the most practice of specific skills the students need in order to really learn the new vocabulary. ALFA LAB should include some practice of basic skills on the laptops as well. ALFA LAB has a good foundation, but it needs more practice materials. Students with 3-6 year achievement gaps have been allowed to just hand in work in any condition. Holistic instruction practices have caused many of the students to think, just 'do' the work. But, the quality of the work is often horrible or they copy answers from the text, not realizing they are plagiarizing or students need additional instruction on how to answer questions (especially 'Why' questions). Often their answers either make no sense or they leave out key information in the answer.
	Main station didn't have much activities except for inference and discussion.
Cohort 3 (n=4)	Stations work! Materials worked! Consistency worked. Increased in rigor worked.
	There was too much paper.
	Collaboration Station, while I like the activities, isn't working very well this year. Students have been so used to being isolated after a year doing distance learning that it's like pulling teeth to get students to actually collaborate and work together. We've had to actually have them do the reading with their group first before giving them the paper or their group will work in silence.
	I feel like we were often rushed and could have benefited from either more time or less material to present. Also, it was often chaotic with 4 different stations going on at the same time with only two adults. My students needed more help than they got at times. I feel that maybe having all students working on the same station at the same time, or even two separate stations, while still being divided into 4 small groups would have helped. I feel that we could have had more time to explain and assist, as each group often had the same questions about the station. Directions and questions could have been then answered whole group.

Note. This question was not presented to respondents in Cohort 2.

Table D27. Respondents' Other Comments

Cohort	In the field below, please share any other comments you may have about the ALFA Lab.
Cohort 1 (n=2)	Please include grammar worksheets/activities.
	I do appreciate the opportunity to use his program and share in its develop[ment]. Our school definitely needs this program, and I hope our comments can only enhance what is a sound structure that does address the needs of ninth grade students that are reading and writing below grade level. The program needs additional support, so it can provide an even more comprehensive package and reach its fullest potential in eradicating an achievement gap for high school students.
Cohort 3 (n=3)	A good intervention platform.
	I found that there were inconsistencies between the manual and what was expected. Students also did not like only reading excerpts from certain stories.
	Overall, our test scores improved, so I would say it was beneficial.

Note. This question was not presented to respondents in Cohort 2.

Appendix E: PLATO Observation Scores

We present school-by-school results from classroom observation using Protocol for Language Arts Teaching Observation (PLATO). In Cohort 1, we observed both treatment and comparison classes in person. In Cohort 3, observations were conducted for treatment classes only using video recordings. We did not observe any classes for Cohort 2.

Table E1. School 1 (NY) PLATO Scores

	Cohort 1						Cohort 3		
Class Type	**ALFA**			**Comparison**			**ALFA**		
PLATO Element	**Seg. 1**	**Seg. 2**	**Overall**	**Seg. 1**	**Seg. 2**	**Overall**	**Seg. 1**	**Seg. 2**	**Overall**
Strategy Use & Instruction	3	2	3	1	1	1	1	2	2
Feedback	2	3	3	2	4	4	1	2	2
Classroom Discourse	2	3	3	3	3	2	1	3	3
Uptake of Student Responses	2	3	3	3	3	3	1	3	3
Opportunities for Student Talk	3	3	3	2	3	3	1	3	3
Text-Based Instruction	3	3	3	3	4	4	2	2	2
Use of Authentic Texts	3	3	3	3	4	4	2	2	2
Production of Texts	1	1	1	2	1	2	2	1	2
Total Score			**12 (out of 16)**			**11 (out of 16)**			**9 (out of 16)**

Note. In Cohort 1, ICF researchers observed two 20-minute teaching segments (seg.) for ALFA and comparison classes. In Cohort 3, ICF researchers virtually observed two to three 15-minute teaching segments for the ALFA Lab class. The white rows represent the four elements that teachers were scored on—Strategy Use & Instruction, Feedback, Classroom Discourse, and Text-Based Instruction—as well as the total score. Shaded rows are sub-elements used to calculate the scores for each element. Each element (and sub-element) could receive a score of 1–4. The total score is a sum of the overall score for each of the four elements. Overall scores represent the highest score achieved over each segment.

Table E2. School 2 (CT) PLATO Scores

	Cohort 3		
Class Type	**ALFA**		
PLATO Element	**Segment 1**	**Segment 2**	**Overall**
Strategy Use & Instruction	1	1	1
Feedback	2	2	2
Classroom Discourse	1	2	2
Uptake of Student Responses	1	2	2
Opportunities for Student Talk	1	2	2
Text-Based Instruction	1	2	2
Use of Authentic Texts	1	2	2
Production of Texts	1	2	2
Total Score			**7 (out of 16)**

Note. ICF researchers virtually observed two to three 15-minute teaching segments for the ALFA Lab class. The white rows represent the four elements that teachers were scored on—Strategy Use & Instruction, Feedback, Classroom Discourse, and Text-Based Instruction—as well as the total score. Shaded rows are sub-elements used to calculate the scores for each element. Each element (and sub-element) could receive a score of 1–4. The total score is a sum of the overall score for each of the four elements. Overall scores represent the highest score achieved over each segment.

Table E3. School 4 (AL) PLATO Scores

	Cohort 3		
Class Type	**ALFA**		
PLATO Element	**Segment 1**	**Segment 2**	**Overall**
Strategy Use & Instruction	1	1	1
Feedback	2	2	2
Classroom Discourse	3	2	3
Uptake of Student Responses	3	2	3
Opportunities for Student Talk	3	2	3
Text-Based Instruction	1	2	2
Use of Authentic Texts	1	2	2
Production of Texts	1	1	1
Total Score			**8 (out of 16)**

Note. ICF researchers virtually observed two to three 15-minute teaching segments for the ALFA Lab class. The white rows represent the four elements that teachers were scored on—Strategy Use & Instruction, Feedback, Classroom Discourse, and Text-Based Instruction—as well as the total score. Shaded rows are sub-elements used to calculate the scores for each element. Each element (and sub-element) could receive a score of 1–4. The total score is a sum of the overall score for each of the four elements. Overall scores represent the highest score achieved over each segment.

Table E4. School 3 (CA) PLATO Scores

	Cohort 1						Cohort 3		
Class Type	**ALFA**			**Comparison**			**ALFA**		
PLATO Element	**Seg. 1**	**Seg. 2**	**Overall**	**Seg. 1**	**Seg. 2**	**Overall**	**Seg. 1**	**Seg. 2**	**Overall**
Strategy Use & Instruction	1	2	2	2	3	3	1	2	2
Feedback	2	2	2	2	2	2	1	2	2
Classroom Discourse	2	3	3	2	2	2	1	3	3
Uptake of Student Responses	2	3	3	2	2	2	1	3	3
Opportunities for Student Talk	2	3	3	2	2	2	1	2	2
Text-Based Instruction	3	3	3	3	3	3	2	3	3
Use of Authentic Texts	3	3	3	3	3	3	1	3	3
Production of Texts	1	1	1	1	3	3	2	2	2
Total Score			**10 (out of 16)**			**10 (out of 16)**			**10 (out of 16)**

Note. In Cohort 1, ICF researchers observed two 20-minute teaching segments for the ALFA Lab class and comparison classes. In Cohort 3, ICF researchers virtually observed two to three 15-minute teaching segments for the ALFA Lab class. The white rows represent the four elements that teachers were scored on—Strategy Use & Instruction, Feedback, Classroom Discourse, and Text-Based Instruction—as well as the total score. Shaded rows are sub-elements used to calculate the scores for each element. Each element (and sub-element) could receive a score of 1–4. The total score is a sum of the overall score for each of the four elements. Overall scores represent the highest score achieved over each segment.

Appendix F: Checklist for ALFA Implementation (CFAI)

We present results from classroom observation using the CFAI for Cohorts 1 and 3 ALFA Lab classes. We did not observe any classes for Cohort 2 due to the COVID-19 pandemic.

Table F1. Checklist for ALFA Implementation Results

Daily Launch						
Item	Cohort 1		Cohort 3			
	School 1 (NY)	School 3 (CA)	School 1 (NY)	School 2 (CT)	School 4 (AL)	School 3 (CA)
Did the teacher introduce the class with a Daily Launch that included either a teaser trailer, a preview of the upcoming reading, an introduction to a new unit, or a lesson on background information needed related to the reading?	Y	Y	Y	N	N	Y
Were the students focused on the teacher during the Daily Launch?	Y	Y	Y	N	Y	Y
Total "Yes"	***2 (of 2)***	***2 (of 2)***	***2 (of 2)***	***0 (of 2)***	***1 (of 2)***	***2 (of 2)***

Overall Room						
Item	Cohort 1		Cohort 3			
	School 1 (NY)	School 3 (CA)	School 1 (NY)	School 2 (CT)	School 4 (AL)	School 3 (CA)
Does the room seem organized (papers off the floor, table materials in neat piles)?	Y	Y	Y	Y	Y	Y
Were there posters on the wall related to ALFA?	Y	Y	Y	Y	Y	Y
How many students are in the room? In each group?	15 total, 2–5 per group	16 total, 3–6 per group	15 total, 3–4 per group	21 total, 5–7 per group	16 total, 5–6 per group	19 total, 3–6 per group

Overall Room						
Item	Cohort 1		Cohort 3			
	School 1 (NY)	School 3 (CA)	School 1 (NY)	School 2 (CT)	School 4 (AL)	School 3 (CA)
Were tables grouped into stations rather than rows or other configurations?	Y	Y	Y	N	Y	Y
Total "Yes"	***3 (of 3)***	***3 (of 3)***	***3 (of 3)***	***2 (of 3)***	***3 (of 3)***	***3 (of 3)***

Main Station						
Item	Cohort 1		Cohort 3			
	School 1 (NY)	School 3 (CA)	School 1 (NY)	School 2 (CT)	School 4 (AL)	School 3 (CA)
Did the teacher introduce text by activating prior knowledge?	Y	Y	Y	Y	Y	N
Did the teacher introduce text by giving a purpose for reading?	Y	N	N/A	N	N	N
Did the teacher introduce key words and/or help students build background knowledge?	Y	Y	Y	N	Y	Y
Did the teacher model effective use of reading strategies?	Y	Y	N/A	Y	Y	N
Did the teacher give students time to read a short text?	Y	Y	N/A	N	N	Y
Did the teacher reference strategies included on the Menu of Strategies poster?	Y	Y	N/A	N	N	N
Did the teacher model the use of instruction/learning tools such as graphic organizers?	N	N	N/A	N	N	N
Did the teacher pause the reading in order to provide instruction, as opposed to reading completely through the text?	Y	Y	N/A	Y	Y	Y
Were students actively listening to the teacher?	Y	Y	Y	Y	Y	Y

Main Station						
Item	Cohort 1		Cohort 3			
	School 1 (NY)	School 3 (CA)	School 1 (NY)	School 2 (CT)	School 4 (AL)	School 3 (CA)
Are students engaging in the discussion?	Y	Y	Y	Y	Y	Y
Is the teacher asking questions to check for understanding?	Y	Y	Y	Y	Y	Y
Total "Yes"	***10 (of 11)***	***9 (of 11)***	***5 (of 11)***	***6 (of 11)***	***7 (of 11)***	***6 (of 11)***

Satellite Stations						
Item	Cohort 1		Cohort 3			
	School 1 (NY)	School 3 (CA)	School 1 (NY)	School 2 (CT)	School 4 (AL)	School 3 (CA)
Did the teacher or assistant establish, explain, and post procedures for discussion behaviors?	N	Y	N	Y	N/A	N
Did the assistant or teacher use a timer set for 20 minutes to ensure timely rotations?	Y	N	N/A	N	N	N
Did the assistant or teacher give a 1–3-minute warning before end of 20-minute rotation?	N	N	N/A	N	N	Y
Did transitions during rotation happen smoothly?	N	Y	N/A	Y	N/A	Y
Are students staying at their stations during the duration of each station?	Y	Y	Y	Y	Y	Y
Were students staying in the same groups as they rotate among the stations?	Y	Y	Y	Y	Y	Y
Are students demonstrating proper social skills?	Y	Y	Y	Y	Y	Y
Are students engaging with students at their tables around the assignments?	N	Y	Y	N	Y	Y

Satellite Stations						
Item	Cohort 1		Cohort 3			
	School 1 (NY)	School 3 (CA)	School 1 (NY)	School 2 (CT)	School 4 (AL)	School 3 (CA)
Were students having discussions in their groups?	N	Y	Y	Y	Y	Y
Was there a Wordology Station?	Y	Y	Y	Y	Y	Y
Were students engaged in Wordology?	Y	Y	Y	Y	Y	Y
Was there a Comprehension Connection Station?	Y	Y	Y	Y	Y	Y
Were students engaged in Comprehension Connection?	N	Y	Y	Y	Y	Y
Was there a Media Madness Station?	Y	Y	Y	Y	Y	Y
Were students engaged in Media Madness?	Y	Y	Y	Y	Y	Y
Was there a computer station?	Y	Y	Y	Y	Y	Y
Was there a listening station?	Y	Y	N/A	N/A	Y	Y
Is the assistant present?	Y	Y	Y	Y	Y	Y
Did the assistant check the progress of students at each station?	Y	Y	Y	Y	Y	Y
Total "Yes"	***13(of 19)***	***17(of 19)***	***14(of 19)***	***15(of 19)***	***15(of 19)***	***17(of 19)***
Grand Total	***31 (of 35)***	***28 (of 35)***	***24 (of 35)***	***23 (of 35)***	***26 (of 35)***	***28 (of 35)***

Note. In Cohort 3, ICF researchers virtually observed classes; N/A indicates items that could not be ascertained in the recording. No observations were conducted in Cohort 2 due to virtual learning environments as a result of COVID-19.

Appendix G: School Profiles & Contexts

For the high schools that participated in implementing ALFA Lab during the 3 years of this study, the following section provides a snapshot of their contextual factors that may have had an impact on the implementation of ALFA Lab in individual schools as they relate to their years of participation.

School 1 (NY)

School 1 (NY) is a public high school located on New York's Long Island, serving a high percentage of minority students with a large proportion of them coming from Hispanic or Latino/a backgrounds. The school also serves a high percentage of students who are economically disadvantaged and receives Title I funding. A majority of the student body— approximately 64%—is classified as English Language Learner (ELL). School 1 participated in all three cohorts of ALFA Lab intervention during the 2019–2020, 2020–2021, and 2021–2022 school years.

Prior to the COVID-19 pandemic in Year 1 of ALFA Lab implementation, School 1 lost its reading specialists to middle and elementary schools in the area owing to a districtwide policy change. This impacted how the school was able to effectively serve their ELL students. To address this, the school attempted to recruit new staff in Year 1 as teaching aides to the ALFA Lab classroom.

Year 2 saw School 1 grapple with the effects of the COVID-19 pandemic. The school implemented ALFA Lab in the fall 2020 semester as a hybrid model with the option of full-time remote learning available. Students were split into two cohorts, with each cohort attending in-person learning 2 days of the week; the remaining 3 days were designated as remote learning. While students were learning remotely, they joined classes synchronously from home. In the spring semester, all students in the school transitioned back to remote-only learning due to an exponential increase in the local COVID-19 positivity rate caused by a new variant.

In Year 3, School 1 returned to complete in-person instruction a couple of months after the school year began. However, the school was still facing challenges associated with the pandemic. Indeed, staff noted that the pandemic exacerbated and compounded the issues that students were already experiencing, particularly with respect to high school graduates who struggled to access support services and resources after graduation. Moreover, students were disengaged and easily distracted as they struggled to adapt to regular in-person instruction given the extended block schedule (70–90 minutes) that School 1 followed.

School 2 (CT)

School 2 (CT) is a private Catholic college preparatory school located in a coastal suburb of Connecticut. The school serves approximately 500 male students who come from diverse racial and ethnic backgrounds along with students from diverse economic backgrounds. Students' literacy skills at the

school were also quite varied, where teachers supported classrooms of students with varied literacy levels. School 2 offered a support class for students with low literacy levels. This class focused on developmental reading and offered reading programs, including ALFA Lab, to students. The school participated in Cohort 3 of the ALFA Lab intervention during the 2021–2022 school year.

At the time of participation in the ALFA Lab intervention, the school was still grappling with the effects of the COVID-19 pandemic specific to student disengagement. Reports of low student engagement and concerns surrounding shorter attention spans as students acclimatize to regular in-person instruction were shared with the ICF evaluation team. School 2 returned to in-person instruction at the start of the 2021–2022 school year.

School 3 (CA)

School 3 (CA) with over 100 years of history is located in the western Mojave Desert in California. This public high school has a high percentage of minority students with a large proportion of them coming from Hispanic or Latino/a backgrounds. Additionally, a majority of the student body—approximately 95%—come from economically disadvantaged backgrounds. Many of the students come from families in which the parents are not college graduates, with some also not having a high school diploma. Staff noted that some students come from neighborhoods and communities surrounded by a strong gang presence, live in foster homes, or are experiencing homelessness. These factors affect student literacy levels since literacy is not a high priority given the lived experiences of these students. Furthermore, the school faces a high student turnover in which student enrollment fluctuates often and can create a challenge with course programming to address students' learning needs. School 3 participated in all 3 years of ALFA Lab intervention during the 2019–2020, 2020–2021, and 2021–2022 school years.

Prior to the onset of the COVID-19 pandemic in Year 1 of ALFA Lab implementation, the biggest contextual factor that contributed to the challenge of implementing ALFA Lab was related to the structure of the semester and their curriculum pacing guidelines. Because the semester was structured in a compact manner, staff had little time during instruction to dedicate to exploring topics in more detail. In the spring semester, the school switched to asynchronous remote learning as a function of the pandemic.

For most of Year 2, School 3 instituted remote synchronous instruction with shortened blocks of instructional time (i.e., 37-minute blocks) following by an "extended hours" block. By March 2021, remote learning was still in place, but students were encouraged to come back to campus for extended hours. This option was available for students with unique needs, including ELL students and those with disabilities; the ALFA Lab teacher opted to offer this virtually.

During Year 3, some COVID-19 protocols were still in place, specifically social distancing, which affected how students experienced station work; that is, students did not move around stations but rather the stations moved. This structure necessitated a significant amount of preparation for teachers. Students in the school continued to struggle with the effects of the pandemic. Staff noted the increase in issues related to students' social-emotional well-being and student disengagement as a result of remote learning.

School 4 (AL)

School 4 (AL) is a public high school in rural Alabama. The student body is mostly White, with a majority from economically disadvantaged backgrounds. The school does not receive Title I funding. The school participated in Cohort 3 of the ALFA Lab intervention during the 2021–2022 school year. Before the onset of the COVID-19 pandemic, the main challenge affecting academic outcomes was reported to be the high poverty rate.

At the time of participation in the ALFA Lab intervention, the school was still struggling with the effects of the COVID-19 pandemic. School 4 returned to in-person instruction at the start of the school year but with COVID-19 protocols in place. During the ICF evaluation team's site visit, staff expressed concern about the low student engagement after the return to regular in-person classes. This, coupled with the high poverty rate, posed a challenge to ALFA Lab implementation.