2015 Collaborative Regional Education (CORE) i3 Validation Study

Final Report

Implementation and Impact Study Results

December 2020

Submitted to:
Jacksonville State University

Submitted by:
ICF Evaluation Team

Table of Contents

I. Introduction

1. Program Background

Rural educational attainment rates remain below the U.S. average, which has a significant economic impact for future job and wage-earning prospects of those living and working in rural areas, and the abilities of these communities to attract and retain employers. Based on a 2012 study, one in four rural students do not graduate from high school and only 17% of adults in rural areas have a college degree (Byun, Meece, & Irvin, 2012). These challenges underscore the importance of having teachers who are well-prepared for instruction and hold students to high standards (Howley and Hambrick, 2011), to increase their readiness for college and/or the workforce. With funding through the Department of Education's Investing in Innovation (i3) grant, the Collaborative Regional Education (CORE) project addresses obstacles rural schools often face for integrating technology and active learning in classrooms. However, simply making technology and professional development available in rural schools does not provide a solution, as teachers need scaffolded support and peer-to-peer collaboration to effectively use the technology resources made available to them (Blanchard, LePrevost, Tolin, & Gutierrez, 2016).

Under the 2015 i3 validation grant, the goal of CORE is to have a positive impact on rural high school students' college and work readiness outcomes by improving teachers' use of classroom technology and active learning strategies. In the 2015 school-level model, CORE project resources were expanded from an earlier 2013 iteration of one-teacher-per middle and high school model to supporting a multi-disciplinary team of teachers and administrators at rural high schools. Operated by Jacksonville State University (JSU), the project has partnered with five regional universities and 28 schools in a total of four states—Alabama, Louisiana, North Carolina, and Texas—to implement CORE. While JSU provided the professional development courses and support, Regional University Partners (RUPs) were an integral part of the effort to streamline processes for implementing the CORE components, collect documentation of successful administration of the PK-20 partnership, and administer data collection activities.

CORE professional development courses began with the 2gno.me skills assessment during orientation, followed by access to an online learning experience based on the SmarterU learning management system. Professional development resources were available to participating school teams, providing teachers with access to online instructional support to integrate technology and new teaching methodologies in school classrooms that promote individualized student learning, and teacher collaboration through sharing learning objects, lesson plans, and teaching strategies. Through this grant, JSU also leveraged 2013 CORE partners to provide high school teams with diagnostic support tools. Civitas provided access to the Change Diagnostic Index (CDI) tool, report, and debrief to assist administrators in assessing needs related to readiness for school change. CDI identifies stress areas within the school system, and mitigation of organizational instability through leadership, professional development, and planning. EdReady , available through a membership-based group of educators partnering to improve student success known as the NROC project, provides an open-resource preparation tool available for math and English/Language Arts (ELA) teachers to annually assess students' needs related to college readiness in order to provide appropriate supports. Teachers were also provided access to JSU instructional support staff and annual workshop opportunities to share reflections and lessons learned with one another.

2. Description of the Intervention

The CORE model is a comprehensive, systems-based approach that consists of seven components designed to build school capacity to better prepare students for college and career by enhancing their 21st century skills, such as critical thinking, problem-solving, technology skills, collaboration skills, and creativity. The CORE model integrates technology and active learning modules in schools by providing multi-disciplinary teams of teachers and administrators with professional development and support to improve college and career readiness and non-cognitive skills outcomes among grade 10-12 students in primarily high need and rural high schools. Over the course of two school years, all CORE schools will participate in seven Key Components:

1. CORE principals engage in professional learning with school teams.
2. School teams participate in online learning communities.
3. Schools receive CORE resources.
4. School teams participate in CORE instructional professional development services.
5. School teams present during CORE professional development workshops.
6. Schools participate in change-management support through CORE partnership resources.
7. School teams provide students with college readiness advisement and support through use of EdReady™ tool.

CORE RUP liaisons identified and met with chosen school administrators to discuss technology needs and develop plans for CORE implementation. Building relationships with school administrators was vital to the success of this project. Classroom funding for hardware and classroom support and technology were provided to treatment schools to procure items based on their school plan. In conjunction with the school-level resources provided, principals and CORE team teachers participated in an online professional learning experience. SmarterU is the medium through which school teams engage with professional development content, instructional services and support, and collegial networking through a content-focused online community. After completing the 2gno.me skills assessment, teachers were granted access to the CORE Learning Management System course catalog in SmarterU.

Over 86 SmarterU micro-courses were developed by JSU to be aligned with the International Society of Technology in Education (ISTE) Standards for Education Leaders to promote student engagement through active learning-based teaching and differentiation of instruction. Based on the individual 2gno.me (To Know Me) skills assessment results (see https://2gno.me/), learning plans were generated for each team teacher and administrator. Administrators participated in a series of leadership modules and were required to support team teachers in implementation through observations and debriefs upon completion of course modules. Teachers completed a minimum of 11 micro-courses per school year with ongoing support from the JSU CORE instructional staff, school team administrator, and their RUP liaison. Reflections and feedback were shared with the online learning community through the SmarterU system.

Change-management support, via CDI and results reports, was provided to CORE schools to support the shift to new modes of instruction. EdReady™ was used to test students on math and ELA skills, identifying areas for improvement, and bridging the gap for remediation for grade 10-12 students. Providing support for college readiness assessments and other resources is

expected to directly impact students' college and career readiness—leading to positive long-term high school and college outcomes. The CORE program's effect on college and career readiness and non-cognitive skills outcomes is thought to be mediated by schools' use of active learning methods. These relationships are depicted in the study logic model (see Appendix A).

3. Evaluation Overview

JSU has contracted with ICF to conduct a federally mandated third-party implementation and impact evaluation of the 2015 i3 validation grant. A cluster randomized control trial (RCT) impact study was conducted to assess a confirmatory question about CORE program impact on schoolwide college and career readiness outcomes after two years of schoolwide implementation. A second confirmatory question examined the CORE program's effects on schoolwide non-cognitive skills outcomes after two years of schoolwide implementation. Exploratory analyses assessed the impact of CORE on these two outcomes after one year of program implementation and by grade level.

The implementation study of CORE is guided by seven evaluation questions, aligned to each of the Key Components (KCs) specified in the CORE program logic model (see Appendix A).

4. Purpose of this Report

The purpose of this report is to provide an overview of evaluation findings at the culmination of the CORE i3 2015 grant, including findings related to fidelity of implementation and impact.

II. Impact Study

1. Impact Study Introduction

The purpose of this study was to assess the impact of the CORE program on participating schools' mean college/career readiness and non-cognitive skills outcomes for grade 10-12 students after two years of schoolwide implementation. The study team used a cluster-level RCT design, randomly assigning 14 schools to the treatment group and 14 schools to the control group. The two-year study design was longitudinal; participating schools and students were followed for three data points (pretest, mid-test, posttest) over two years, enabling longer-term tracking of students who participated in the entirety of the study. The study focused on program impact by analyzing how student college/career readiness and non-cognitive skills outcomes, between pretest and mid-test and between pretest and posttest, changed for treatment and control groups. The change between pretest and posttest was our confirmatory focus. The change between pretest and mid-test, as well as other additional evaluation questions (discussed below), were considered initial or exploratory findings.

The study considered four student outcomes. The main outcome, operationalized as College and Work Readiness Assessment + (CWRA+) scores, is students' competencies in critical thinking, analytic reasoning, problem-solving, and written communication skills—all 21st century skills that the *Partnership for 21st Century Skills* has deemed critical for college and work environments (2018). The study team also assessed students' noncognitive skills and student engagement and efficacy scores. JSU and ICF collaborated to develop the non-cognitive student scale to measure students' non-cognitive orientation, which may be indicative of

students' academic success. The study continued to use two student orientation measures, student engagement and student efficacy, from the previous i3 CORE study (ICF, 2018). The engagement scale measures the level of student engagement in academic course work and education in general. The efficacy scale measures students' confidence in whether they can excel at school. All the measures were administered to a cohort of sampled students in participating schools at three points in time: pretest in the spring of school year 2017-18, mid-test in the spring of school year 2018-19, and posttest in the spring of school year 2019-20).

The following sections describe evaluation questions, methods, analysis, and results for the impact study.

2. Research Questions

2.1 Main Program Impact Analysis

As previously described, the CORE model provides professional development opportunities and resources for teachers and administrators to enhance engagement with colleagues and to positively impact instructional practices and strategies (e.g., more effective use of technology). By exposing students to enhanced active learning instructional methods, the CORE model sought to improve students' levels of college and work readiness, as measured by CWRA+. CORE also aimed to improve students' non-cognitive skills, engagement, and efficacy.

Exhibit 1 summarizes four of the evaluation questions as originally designed to be conducted as an RCT with the whole analysis sample. The main questions focus on two-year program impact based on students' growth between pretest and posttest. The first confirmatory question is whether the average school-level college and career readiness scores, as measured by the CWRA+, were higher for students from treatment schools compared to those from control schools. The second confirmatory evaluation question asked whether school-level non-cognitive skill outcomes, as measured by a non-cognitive skill measure, were higher for students from treatment schools compared to control schools. The two exploratory questions focused on two other student measures as outcomes: Student Engagement scores and Student Efficacy scores. Like the confirmatory questions, the ICF team assessed the program impact of the two-year intervention on these outcomes.

Exhibit 1. Summary of Main Confirmatory and Exploratory Evaluation Questions: Based on Pretest-Posttest Data (Two-Year Program Impact Analysis)

Confirmatory Evaluation Question 1	What is the impact of two years of schoolwide CORE implementation upon the mean school-level **CWRA+ scores** for Grade 11-12 students compared with the business-as-usual condition?
Confirmatory Evaluation Question 2	What is the impact of two years of schoolwide CORE implementation upon the mean school-level **non-cognitive skill (NCS) scores** for Grade 11-12 students compared with the business-as-usual condition?
Exploratory Evaluation Question 1	What is the impact of two years of schoolwide CORE implementation upon the mean school-level **student engagement scores** for Grade 11-12 students compared with the business-as-usual condition?
Exploratory Evaluation Question 2	What is the impact of two years of schoolwide CORE implementation upon the mean school-level **student efficacy scores** for Grade 11-12 students compared with the business-as-usual condition?

The COVID-19 situation in the spring of 2020 during the last phase of data collection affected the completeness and quality of posttest data collected. Robust findings for the two confirmatory questions could not be obtained due to COVID19-related disruptions (schools were closed, and teachers taught 11th graders primarily online in the spring of 2020 and in many districts, seniors were exempt from online coursework). As explained in further detail later, the confirmatory part of the study became a high-attrition RCT and the design was considered a quasi-experimental design (QED).

The study team addressed several other exploratory questions regarding CORE's one-year impact based on data collected from pretest to mid-test (Spring 2018 and Spring 2019). Findings from the pretest and mid-test data became more important (than previously considered) for providing some indication of program outcomes, as these data were collected earlier and not affected by COVID-19-related disruptions. As shown in Exhibit 2, these questions are almost identical to the confirmatory questions. The subjects were the same students who were 10th and 11th graders at the time of mid-test data collection. The difference, again, is the duration of the intervention, which was one year. These questions were addressed by the data collected between pretest (Spring 2018) and mid-test (Spring 2019).

Exhibit 2. Summary of Exploratory Evaluation Questions: Based on Pretest-Mid-test Data (One-Year Program Impact Analysis)

Exploratory Evaluation Question 3	What is the impact of one year of schoolwide CORE implementation upon the mean school-level **CWRA+ scores** for Grade 10-11 students compared with the business-as-usual condition?
Exploratory Evaluation Question 4	What is the impact of one year of schoolwide CORE implementation upon the mean school-level **non-cognitive skill (NCS) scores** for Grade 10-11 students compared with the business-as-usual condition?
Exploratory Evaluation Question 5	What is the impact of one year of schoolwide CORE implementation upon the mean school-level **student engagement scores** for Grade 10-11 students compared with the business-as-usual condition?
Exploratory Evaluation Question 6	What is the impact of one year of schoolwide CORE implementation upon the mean school-level **student efficacy scores** for Grade 10-11 students compared with the business-as-usual condition?

2.2 Subgroup Impact Analysis

The following set of exploratory questions (Exhibit 3) examined how CORE program impact may be associated with various student characteristics, after one year and two years of program implementation. These questions explored possibilities that the program may have different levels of effectiveness, as measured by student outcomes, for subgroups defined by gender, race (white vs. minority students), parent education level (at least one parent graduated college vs. the rest), pretest CWRA+ scores (low and high based on percentiles), and regions where students attended school (as defined by school affiliation with RUPs). As the RCT study was not designed to confirm these hypotheses and the sample size was too small to sustain sufficient statistical power, these questions are posed as exploratory questions. Initial findings may encourage future confirmatory investigations.

Exhibit 3. Summary of Exploratory Evaluation Questions Related to Subgroup Impact

Exploratory Evaluation Question 7a and b	How does one-year CORE program impact on students' outcomes vary by **gender**? How does two-year CORE program impact on students' outcomes vary by **gender**?
Exploratory Evaluation Question 8a and b	How does one-year CORE program impact on students' outcomes vary by **minority status**? How does two-year CORE program impact on students' outcomes vary by **minority status**?
Exploratory Evaluation Question 9a and b	How does one-year CORE program impact on students' outcomes vary by **parents' education level?** How does two-year CORE program impact on students' outcomes vary by **parents' education level?**
Exploratory Evaluation Question 10a and b	How does one-year CORE program impact on students' outcomes vary by **pretest CWRA+ scores (Low and High)**? How does two-year CORE program impact on students' outcomes vary by **pretest CWRA+ scores (Low and High)**?
Exploratory Evaluation Question 11a and b	How does one-year CORE program impact on students' outcomes vary by **region**? How does two-year CORE program impact on students' outcomes vary by **region**?

2.3 Additional Exploratory Analysis of Program Implementation and Student Outcomes

2.3.1 Data Sources for Additional Exploratory Analyses

ICF examined how CORE program impact varied by other intervention characteristics. Available data for these exploratory analyses included:

- **Exposure data (student and teacher link):** For each treatment student, data were available indicating whether students were taught by CORE program participant teachers (i.e., teachers participating on CORE teams). ICF requested that teachers report whether they taught students participating in the study.
- **Implementation data (Key Components 1 to 7):** As described later in the Implementation Study section, program fidelity of implementation is captured through seven Key Components.
- **Teacher 2gno.me data:** Treatment and control teachers completed the 2gno.me pretest and mid-test, designed to measure their experience in seven areas: learner, leader, citizen, collaborator, designer, facilitator, and analyst. Change over time in 2gno.me scores was compared across the two teacher groups.
- **Change-management data:** Data provided insight on teachers' flexibility and openness to organizational changes. The data were collected at four timepoints from all treatment schools and some comparison schools.

The ICF team explored how three of the four data sources related to program impact. The change-management data had a limitation in that different treatment schools participated in the survey at different timepoints (Fall 2018, Spring 2019, Spring 2020) and the control schools participated at a different timepoint (Fall 2019). The data provided useful diagnostics for participating schools; however, ICF decided not to use this for evaluation analysis. In the

following sections, questions explored through the exposure data analysis, the implementation data analysis, and the teacher 2gno.me analysis are described. Findings from these analyses are reported in the Implementation Study Results section. The resulting patterns are informative and relevant for future program implementation considerations.

2.3.2 Exploratory Analysis of Exposure Data and Student Outcomes

This analysis focuses on students' "exposure to the intervention" data. As mentioned above, students at treatment schools were taught by varied numbers of treatment teachers. Some students in treatment schools were never taught by any of the teachers participating in CORE, while others have been taught by as many as five or more CORE teachers (e.g., when teachers on CORE teams changed at the same school). The outcome of interest was students' change in CWRA+ scores. The level of students' exposure to the intervention was measured by the number of team teachers who reported that the student was in their courses. We expected that if students are exposed to more CORE teachers through their courses, they may be more likely to exhibit gains in CWRA+ scores. Evaluation questions explored are as follows:

- EQ12a: How is students' level of exposure to CORE teachers during the pretest-to-mid-test phase (over one year) related to changes in CWRA+ scores?
- EQ12b: How is students' level of exposure to CORE teachers during the pretest-to-posttest phase (over two years) related to changes in CWRA+ scores?

2.3.3 Exploratory Analysis of Implementation Data (Key Components) and Student Outcomes

ICF also explored whether treatment schools' level of fidelity to program implementation was associated with students' growth in CWRA+ scores. Treatment schools (and the attending treatment students) were classified into different implementation levels based on school-level results regarding fidelity to the seven Key Components (see details in the Implementation Study Analysis section). Evaluation questions explored are as follows:

- EQ13a: How do schools' fidelity to program implementation over one year relate to pretest-to-mid-test changes in students' CWRA+ scores?
- EQ13b: How do schools' fidelity to program implementation over two years relate to changes in students' pretest-to-posttest CWRA+ scores?

The general expectation is that program impact on the outcome is greater when treatment schools had a higher level of fidelity to program implementation. In other words, schools that were implementing the program as intended should be more likely to experience impact. Note that the study was not designed to treat this question as confirmatory. Findings for this analysis should be interpreted as suggestive of future study direction.

Data were collected from teachers and administrators to understand the degree to which treatment schools had implemented the seven CORE program components. As detailed in the Implementation Study section, the seven components are (1) principal engagement, (2) teachers' active participation in online program activities, (3) school resources, (4) professional development activities, (5) school teams' presentations during professional development workshops, (6) change management, and (7) use of EdReady™. Per each KC, the evaluation team classified the 14 treatment schools into three levels of program implementation (low, medium, and high). To understand how implementation and student outcomes are correlated,

the team derived school-average student outcome scores (change in CWRA+ scores between pretest and mid-test and between pretest and posttest) by the three levels of implementation (low, medium, and high). As a reference, the average student outcome scores for the control group were also calculated.

2.3.4 Analysis of Teacher 2gno.me data and Student Outcomes

ICF analyzed the 2gno.me assessment data collected from treatment and control teachers and examined the following exploratory questions.

- EQ14a: How do the measures of the treatment and comparison teachers change by measurement points (pretest, mid-test, posttest)?
- EQ14b: How are the school-average 2gno.me measures correlated with school average changes in CWRA+ scores?

The CORE program is expected to encourage teachers to grow in the teacher traits measured by 2gno.me and thus, the averages of treatment teachers should be higher than those of control teachers at mid-test and posttest data collection points. In terms of how the school-level 2gno.me scales were correlated with student outcomes, the analysis team examined how the school average change in CWRA+ scores (pretest to mid-test; mid-test to posttest) are correlated with the school average 2gno.me scores. The analysis is highly descriptive because the units of analysis are schools and thus the number of cases is limited.

3. Impact Study Methodology

3.1 School Randomization

Randomization of the 28 recruited schools to study conditions (treatment or control) was conducted in July 2018. Because each school was recruited and supported during the study by a RUP, ICF treated RUPs (n=5) as blocks within which the random assignment of schools to conditions occurred. Blocking ensures a reasonable balance of treatment and control schools will be identified within each RUP.

Exhibit 4. School Randomization Status Results by Regional University Partner

School Randomization Results		
RUP	**School Name**	**Assignment Status**
Fayetteville State University	Rocky Mount High School	Control
	E. E. Smith High School	Treatment
	Massey Hill Classical High School	Control
	Westover High School	Treatment
	West Bladen High School	Control
	Cross Creek Early College High School	Treatment
Jacksonville State University	Chilton County High School	Control
	Talladega High School	Treatment
	Moody High School	Control
	Lawrence County High School	Treatment
	Skyline High School	Control
	Haleyville High School	Treatment
	Cherokee High School	Control

School Randomization Results		
RUP	**School Name**	**Assignment Status**
	Colbert County High School	Treatment
	Ragland High School	Control
	Woodville High School	Treatment
	North Jackson High School	Control
	Central High School of Clay County	Treatment
	Springville High School	Control
Tarleton State University	Lipan High School	Treatment
	De Leon High School	Control
	Mart High School	Treatment
	Bosqueville High School	Control
Louisiana Tech University	Tensas High School	Treatment
	Pleasant Hill High School	Control
West Texas A&M University	Booker Junior/High School	Treatment
	Hereford High School	Control
	Brownfield High School	Treatment

3.2 Definition of Team Teachers and Team Composition Changes

Teachers selected for CORE teams in treatment schools participated in program activities. To be part of the team at each treatment school, teachers needed to complete a consent form and to participate in the CORE orientation session. The membership of teachers changed occasionally as some teachers left their CORE teams (without leaving their schools) at some point during the two years of the study, and other teachers left their schools entirely. When teachers joined the team later during the school year, they were given a link to a recorded orientation session about the program. As a school-level randomized study, these team membership changes do not affect study quality, as turnover is expected. ICF tracked team changes as far as leavers and replacements and used CORE team teacher and student interactions (i.e., whether students were taught by team teachers and how much) as a variable that potentially contributes to student outcomes (discussed earlier as an exploratory analysis of students' exposure to the team teachers). Exhibit 5 below describes the change in numbers of treatment and control teachers overall in each year of the study.

Exhibit 5. Changes in Treatment and Control Teacher Groups Year to Year

Condition	School Year	Team Members	Leavers n	Arrivers n	Leavers %	Arrivers %
Control	2018-2019	104	7	0	7%	0%
	2019-2020	104	3	10	3%	10%
Treatment	2018-2019	110	16	2	15%	2%
	2019-2020	107	9	24	8%	22%

3.3 Sample Identification

This section describes how the CWRA+ was administered to students in the 28 participating treatment and control schools as defined in the original sample, as well as how attrition was calculated when the analysis sample was constructed. Essentially, all students who were enrolled at the time of pretest (Spring 2018) became study participants. Any students missing from this initial student roster were not part of the study sample. Schools were considered "attrited" when they dropped out of the CORE program or did not supply data for analysis. The analysis sample did not include any "joiners." The What Works Clearinghouse (WWC) definition states that subjects who join the study after random assignment may affect the integrity of an RCT study to the extent that they self-selected to be at school for receiving the intervention (What Works Clearinghouse, 2013). There were no joiners, as no one was eligible to become part of the analysis sample if they did not provide pretest data. This helped minimize potential bias, as WWC indicates that introducing joiners can be a source of bias.

The impact study relied on two analysis samples. Evaluation questions addressing two-year program impact relied on the sample of students who took both the pretest and posttest. This was a two-year timespan, with COVID-19 disruptions affecting the final point of data collection. Thus, this sample suffered a significant amount of attrition. In contrast, evaluation questions addressing one-year program impact focused on the sample of students who took the pretest and mid-test (at the end of year 1). This latter dataset was more complete than the final full sample, due to the shorter amount of time between data points and because COVID-19 disruptions had not yet occurred.

The full sample of students collected for assessing the two-year program impact experienced both high school-level and student-level attrition. As details will follow in a later section, the baseline equivalence of data could not be established based on pretest CWRA+ test scores. To establish baseline equivalence and define and select comparison and treatment students for analysis, ICF used propensity score matching (PSM). Based on the similarity in predictors such as pretest CWRA+ scores and other student characteristics, PSM created the matched comparison sample in which each student in the CORE group were matched to a student in the comparison growth with a similar propensity score that quantified the multiple predicators used in matching.

3.4 CWRA+, the Student Outcome Measure and Test Administration

The CWRA+ is a standardized assessment, developed by the Council for Aid to Education (CAE), designed to measure student mastery of 21st century skills that are necessary for success in postsecondary education and workforce settings (e.g., critical thinking). The assessment includes both performance task (PT) and selected response questions (SRQs). To minimize test administration time, only SRQs were administered. The SRQ score represents students' cumulative performance related to three 21st century skills: (1) scientific and quantitative reasoning, (2) critical reading and evaluation, and (3) critiquing an argument. These are skills hypothesized to be positively influenced by teachers' exposure to and use of instructional strategies learned through participation in CORE. The SRQ has sufficient internal consistency (Cronbach's $\alpha = .73$). For more information about the CWRA+ and SRQ score measures, see the CAE solutions website https://cae.org/wp-content/uploads/2020/07/Client-Case-Studies-Curriculum-Efficacy-Study.pdf.

The CWRA+ SRQ items were administered to two grade levels of students (9th and 10th graders at pretest; 11th and 12th graders at posttest) who were enrolled in the participating teachers' classrooms on or before the beginning of each school year at participating treatment schools. ICF requested that schools test all 9th and 10th graders enrolled. All participating students were asked to complete the study outcome measure at pretest (Spring 2018), mid-test (Spring 2019), and posttest (Spring 2020).

3.5 Non-cognitive Student Scores, Student Engagement and Self-Efficacy Scores for Exploratory Outcome Analysis

When completing the CWRA+ at pretest and posttest, students were also asked to respond to items from the students' non-cognitive scale, student engagement scale, and efficacy scale. The non-cognitive student scale was developed and pilot-tested by JSU and ICF (see Appendix C, Exhibit AC1, for the listing of all survey items). This part of the survey asked students ten questions, such as "I can prioritize my work to ensure I am completing tasks in a timely manner," "I am confident I will complete any task assigned," and "I see more than one correct answer to many questions." The purpose of this instrument is to measure students' academic and schooling orientation. Students who score higher on this scale are considered to likely be better prepared for success in academic performance, high school completion, and readiness for college. The reliability (Cronbach alpha) of the non-cognitive scale is .81.

The study team also used student engagement and efficacy scales (see Appendix C, Exhibit AC1 for the items). The four student engagement questions were adapted from the Consortium on Chicago School Research Academic Engagement Scale (CCSR/AES) (Consortium on Chicago School Research, 2007). Five self-efficacy questions were adapted from the Patterns of Adaptive Learning Scales (PALS, Midgley et al., 2000). Based on pretest data, the two measures had sufficiently high Cronbach's alpha values of .71 and .84. It was hypothesized that students in classrooms with teachers who were participating in CORE would have higher scores on these three student scales than students in control classrooms. Students rated themselves on each item on a Likert-type scale of *strongly agree, agree, neither strongly disagree or agree, agree,* and *strongly agree.* The five response values were coded, respectively, as 1, 2, 3, 4, and 5. The average value of the survey items were derived as a student-level score and used for analysis.

4. Impact Study Analysis Model

ICF used the Hierarchical Linear Modeling (HLM) framework for multivariate regression modeling to estimate program impact (Raudenbush & Bryk, 2002). Program impact estimates were adjusted for pretest CWRA+ scores, grade level, gender, race and ethnicity, and parents' college education. School differences were also adjusted in the model as random effects. A standard assumption of residuals being normally distributed was not attainable as the data were correlated by schools as clusters. To address this clustering issue, the HLM framework estimated the intercepts (i.e., school effects) as random effects. The program effect was estimated as the coefficient of the treatment status (1 if treatment, 0 if control) and the standardized effect size was presented to facilitate interpretation (standard deviation of the analysis sample was used to standardize the program impact coefficient). The following equation summarizes the model described above:

$$Posttest_{ij} = \beta_{00} + \beta_{10} * pretest_{ij} + \beta_{20} * treatment_j + ... + r_{ij} + u_j$$

where:

- Posttest represents posttest outcome scores
- Pretest represents baseline scores (of the posttest outcomes)
- Postscripts *i* and *j*, respectively represent student and teacher
- βs are parameters to be estimated
- The three ellipses (i.e., "…") indicate that the model will include multiple predictors and corresponding parameters; predictors are gender, grade levels (9th and 10th at pretest), race and ethnicity, parents' college education (if at least one parent earned BA degree 1; else 0).
- Treatment represents the treatment status (1 if treatment group; 0 if control group)
- *r* and *u* are independently and identically distributed residuals with a mean of 0.

This model was applied for evaluation questions (CEQ 1 and 2; EEQ1 — 11). The outcome variables were switched per each evaluation question.

5. Impact Study Results

5.1 Attrition Analysis

This section describes findings from the attrition analysis, the baseline equivalence analysis, and the PSM analysis. To state the conclusion first, only one of the contrasts retained its methodological status as a low attrition RCT (randomized controlled trial). This contrast examined evaluation question 3 (the pretest-mid-test analysis of CWRA+ outcome). Due to data attrition and baseline equivalence problem, seven other main contrasts became a QED (quasi-experimental design study). Exhibit AD14 summarizes the results of attrition analyses, baseline equivalence analyses, and PSM (propensity score matching) analyses.

Exhibit 6 summarizes the results of the overall school-level attrition rate and differential attrition rates. For the pretest-to-mid-test CWRA+ sample, school-level attrition was kept within the threshold specified by WWC standards. The study began with 14 schools each in treatment and control groups. One control school did not administer the CWRA+ pretest. Thus, when mid-test data were collected, the number of participating schools were 14 and 13, respectively, for treatment and control groups. Another control school did not submit the student scale part of the CWRA+, which resulted in the loss of one control school. Thus, the student scale sections of the study (non-cognitive scale, engagement scale, and efficacy scale) became a high-attrition study. For the second phase of the study, data collection was not conducted at all participating schools for the pretest-to-posttest sample due to COVID-19-related school closures. Data were obtained from six treatment schools and five control schools. The overall attrition rate and differential attrition rate were larger than what WWC considers acceptable. In conclusion, the pretest-mid-test CWRA+ study is a low-attrition RCT, but all other contrasts (pretest-mid-test student scales and pretest-posttest CWRA+ and student scales) were considered high-attrition RCTs.

Exhibit 6. Summary of Sample Sizes and Cluster-level Attrition Information

Cluster Level	Roster (Pretest Takers)			Analysis Sample			Attrition Analysis (%)					
	Tx	Con-trol	Sub-total	Tx	Con-trol	Sub-total	Over all	Tx	Con-trol	Differential Attrition Rate (DAR)	WWC liberal boun-dary	Attri-tion Level
Pretest+ Mid-test (One-Year Program Impact Analysis)												
CWRA+	14	14	28	14	13	27	3.6	.00	7.1	7.1	10.3	Low
Student Survey Scales	14	14	28	14	12	26	7.1	.00	14.3	14.3	10.8	High
Pretest + Posttest (Two-Year Program Impact Analysis)												
CWRA+	14	14	24	6	5	11	60.7	57.1	64.3	7.1	1.4	High
Student Survey Scales	14	14	24	6	5	11	60.7	57.1	64.3	7.1	1.4	High

Exhibit 7 shows the results of student-level attrition rates. Because CWRA+ scores and the three student sub-scales— (1) students' non-cognitive scale, (2) student engagement scale, and (3) efficacy scale—had slightly different patterns of missing values, attrition rates were calculated separately. Per WWC guidelines, students were included in attrition calculation only when their schools participated in the data collection. This is why only six treatment schools and five control schools were included in the attrition calculation for the pretest-to-posttest analysis. The individual-level attrition rate for the pretest-posttest CWRA+ contrast surpassed the WWC threshold.

Exhibit 7. Student-level Attrition Calculations by School Year (Pretest and Mid-test Data)

Sub-cluster (student) Level	Roster (Pretest Takers)			Analysis Sample			Attrition Analysis (%)					
	Tx	Control	Sub-total	Tx	Control	Sub-total	Over-all	Tx	Control	DAR	WWC Bound-ary	Attrition level
From pretest to mid-test analysis (interim analysis)												
CWRA+	1,624	1,853	3,477	1,105	1,187	2,292	34.1	32.0	35.9	4.0	7.4	Low
Student Survey Scales	1,624	1,853	3,477	946	1,031	1,977	43.1	41.7	44.4	2.6	5.3	Low
From pretest to posttest analysis (final analysis)												
CWRA+	507	705	1,212	273	423	696	42.6	46.2	40.0	6.2	5.6	High
Student Survey Scales	507	705	1,212	268	400	668	44.9	47.1	43.3	3.9	5.1	Low

Note: Per WWC guidelines, students from the school that did not submit pretest data (n=1) were excluded from the sub-cluster attrition analysis. DAR means "differential attrition rate."

5.2 Baseline Equivalence Analysis

The attrition analysis found that the only pretest-to-mid-test CWRA+ analysis sample (Exploratory Evaluation question 3; see Exhibit 2) a low-attrition sample (both at school and student levels), making the analysis a valid RCT (Randomized Controlled Trial) study. Other contrasts had the school-level attrition problem and thus were downgraded to be QED (quasi-experimental design) studies. For the latter contrasts to remain valid per WWC guidelines, it was necessary for baseline equivalence to be demonstrated. Exhibit 8 shows the results of baseline equivalence tests on pretest CWRA+ and pretest variables for other student scales. The pretest-posttest samples encountered the problem. The two groups were not equivalent in terms of the average pretest CWRA+ scores as Hedge's g statistic was greater than 0.25 (which WWC uses as a threshold).[1] Because baseline equivalence was not established for this important variable (used for the confirmatory analysis), ICF decided to conduct the PSM analysis on the pretest-posttest sample.

Exhibit 8. Baseline Equivalence Test Results for the Analysis Samples

		Treatment Group			Control Group			WWC Baseline Test		
	Total N	N	Mean	SD	N	Mean	SD	Mean difference	Hedge g (absolute value)	Result
Pretest and Mid-test Data										
CWRA+	2292	1105	917.90	170.88	1187	932.21	175.40	-14.31	0.08	Low attrition RCT
Non-cognitive Skill	1977	946	3.80	0.55	1031	3.78	0.57	0.02	0.03	BE satisfied
Engagement	1977	946	3.46	0.76	1031	3.44	0.77	0.02	0.03	BE satisfied
Efficacy	1977	946	3.95	0.71	1031	3.94	0.73	0.02	0.03	BE satisfied
Pretest and Posttest Data										
CWRA+	696	273	1,006.32	194.82	423	922.45	163.31	83.87	0.475	BE not satisfied
Non-cognitive Skill	668	256	3.87	0.48	373	3.89	0.57	-0.02	0.04	BE satisfied
Engagement	668	256	3.47	0.77	373	3.57	0.78	-0.10	-0.13	BE satisfied
Efficacy	668	256	4.09	0.57	373	4.01	0.74	0.08	0.12	BE satisfied

5.3 The Propensity Score Matching (PSM) Analysis

The final sample for the pretest-to-posttest analysis needed to be reconstructed as the original sample did not establish the treatment and control group's baseline equivalence without propensity score matching. The analysis team took a quasi-experimental approach of using a PSM sample for all four outcomes because the two groups of students were found to be different at baseline on the outcome of interest for this study. The PSM model included the

[1] Hedge's g is a measure of effect size, which conveys how much one group differs from another.

following predictors: pretest CWRA+ scores, pretest noncognitive student scores, gender, race (minority vs. white students), and parents' education levels. To prioritize the size of the resulting sample, no exact criterion was specified, meaning students in the treatment group can be matched with any students in the control group regardless of schools. As shown in Exhibit 9, the two groups were created to maximize equivalence between the two groups on pretest variables.

Exhibit 9. Baseline Equivalence Test Results for Matched Data (Pretest and Posttest Data)

		Treatment Group			Control Group			WWC Baseline Test		
	Total N	N	Mean	SD	N	Mean	SD	Mean difference	Hedge g (absolute value)	Result
Pretest CWRA+	474	237	968.46	173.74	237	965.70	173.67	2.76	0.02	Satisfied BE
Pretest Non-cognitive Skill	443	225	4.07	0.57	218	4.04	0.69	0.03	0.04	Satisfied BE
Pretest Engagement	443	225	3.51	0.75	218	3.57	0.71	-0.06	0.08	Satisfied BE with statistical adjustment
Pretest Efficacy	443	225	3.87	0.48	218	3.92	0.51	-0.05	0.10	Satisfied BE with statistical adjustment

PSM successfully established baseline equivalence for the pretest-posttest analysis sample; however, the sample size became smaller than originally anticipated. The analysis team compared the sample characteristics between the two study phases. As shown in Appendix Exhibit AD13, the sample size for the CWRA+ analysis from the first study phase (pretest-mid-test) was 2,292. When PSM was applied to the data collected during the COVID-19 pandemic, the sample size for the second study phase (pretest-posttest) became 474. The two samples showed differences in academic, demographic and background characteristics. The second phase sample (after PSM) had a higher pretest CWRA+ average (967) than the first phase sample (925). The second phase sample (compared to the first phase sample) included fewer male students (47% vs. 40%), fewer white students (49% vs. 39%), and more students with college-educated parents (31% vs. 38%).

5.4 Confirmatory Analysis of Program Impact on CWRA+ Outcomes

ICF used the HLM framework for multivariate regression modeling to estimate program impact on student outcome scores (CWRA+, Non-cognitive score, Engagement score, and Efficacy score). The derived program impact estimates were adjusted for pretest CWRA+ scores, grade levels (9^{th} or 10^{th} at pretest), gender, race and ethnicity, and parents' college education. The between-school outcome differences were also included in the model as random effects. The ICF team assessed separately the results from the one-year exposure analysis of the pretest-to-mid-test sample and the two-year exposure analysis of the pretest-to-posttest sample.

Exhibit 10 summarizes program impact analysis findings.

Program impacts based on the two-year analyses were consistently higher than those from the one-year analysis, suggesting that longer-term exposure to the program is important for increased impact. Looking further at one-year program effects, none of the effects were large (standardized effect sizes range from 0.00 to 0.12) and none were statistically significant. The largest standardized effect size of 0.12 was found with analysis of the CWRA+ outcome.

When two-year program impacts were considered, effect sizes were larger than one-year estimates. The program impacts on students' CWRA+ scores and the three student scales ranged from 0.22 to 0.32. The education evaluation literature typically considers an effect size around .25 meaningful.

Three of the four program effects were close to this threshold and the program effect for student efficacy surpassed this threshold. Note that the two-year analysis was based on a smaller dataset created by a PSM technique; however, it is important to consider that effect sizes were reasonably large, suggesting that program effects may have been more apparent if COVID-19 had not affected the data collection effort. In terms of statistical significance, the results for the non-cognitive skill scale and the efficacy scale were found to be statistically significant in the two-year analysis; however, these outcomes lacked between-school variance and HLM did not converge. The statistical tests were derived from the fixed effect model (Ordinary Least Square regression) and thus these results may be overly optimistic.[2]

Exhibit 10. Summary of Program Impact Analysis Results for CWRA+ Scores

	N of Schools	N of Students	Program Impact	Std. Error	P	Sig.	Standardized Effect
Pretest-Mid-test Analysis (One-Year Program Impact)							
CWRA+	27	2,292	20.94	20.11	0.31	ns	0.12
Non-cognitive Skills Scale	27	1,977	0.02	0.03	0.52	ns	0.03
Engagement	27	1,977	0.00	0.05	1.00	ns	0.00
Efficacy	27	1,977	0.05	0.04	0.18	ns	0.07
Pretest-Posttest Analysis (Two-Year Program Impact)							
CWRA+	11	474	38.22	32.01	0.25	ns	0.22
Non-cognitive Skill Scale	11	443	0.13	0.05	0.01	* (OLS)	0.22
Engagement	11	443	0.18	0.11	0.16	ns	0.23
Efficacy	11	443	0.24	0.06	0	*** (OLS)	0.32

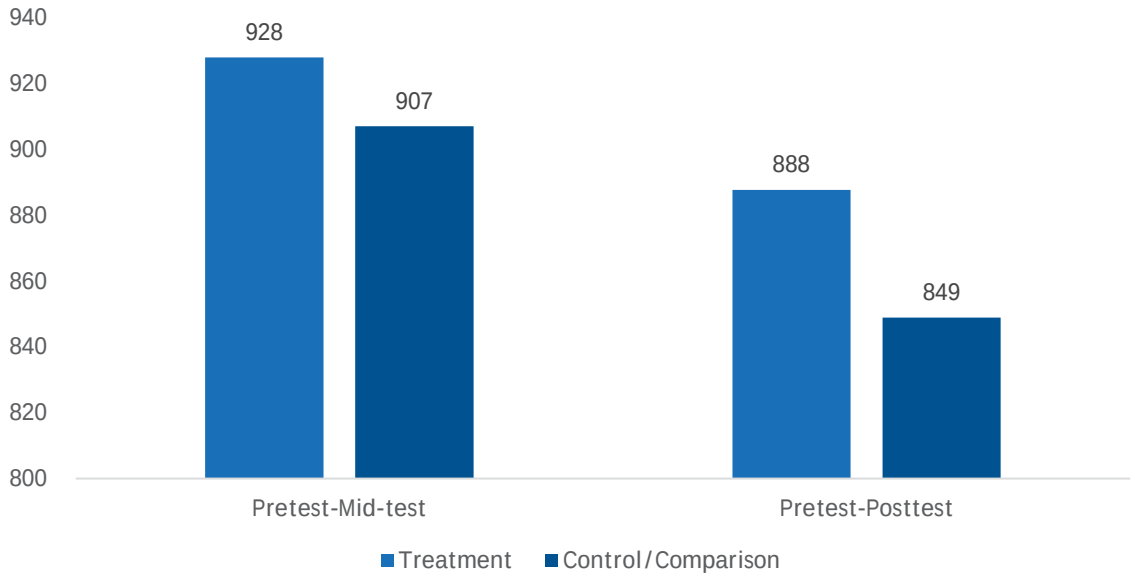
Note: Statistical significance (2-tail test): * = p<.05, ** = p<.01, *** = p<.001. "ns" means the results were not statistically significant. The estimates were adjusted for pretest CWRA+ scores, grade levels, gender, race and ethnicity, and parents' college education. See Appendix D for full results and descriptive statistics.

[2] These two outcomes did not have sufficient between-school variance (i.e., adequate differences between student scores across schools) to sustain the multi-level modeling and thus the estimates were derived from the fixed-effect model (ordinary least square regression model). The results from fixed models tend to be overly optimistic when it comes to finding significance (Raudenbush and Bryk 2002).

The following section discusses program impact findings (already reported in Exhibit 10) by graphically representing the estimated group averages. Per each outcome (CWRA+ and three student survey scales), the adjusted averages of the treatment group and control group were represented, and the results were separately shown for the first phase of the study (pretest-to-mid-test) and for the final phase of the study (pretest-to-posttest). The adjusted averages are the values derived from the HLM models and adjusted for predictors included in the model. For ease of interpretation, the value of the control group was fixed at the control group's unadjusted average scores.

Exhibit 11 graphically summarizes the results of impact analysis for CWRA+ outcomes. The control/comparison group values (907 and 849) were based on the unadjusted averages of the group and the treatment group values (928 and 888) were based on the findings from the multivariate models[3]. As previously discussed, **the treatment group had a higher average score than the control group** for both analysis phases (pretest-mid-test and pretest-posttest). As previously reported in Exhibit 10, standardized effect sizes were .12 and .22, respectively, for the pretest-mid-test analysis and for the pretest-posttest analysis.

Exhibit 11. Program Impact Analysis Results for CWRA+ Adjusted Average Scores per Group (Pretest-Mid-test and Pretest-Posttest)



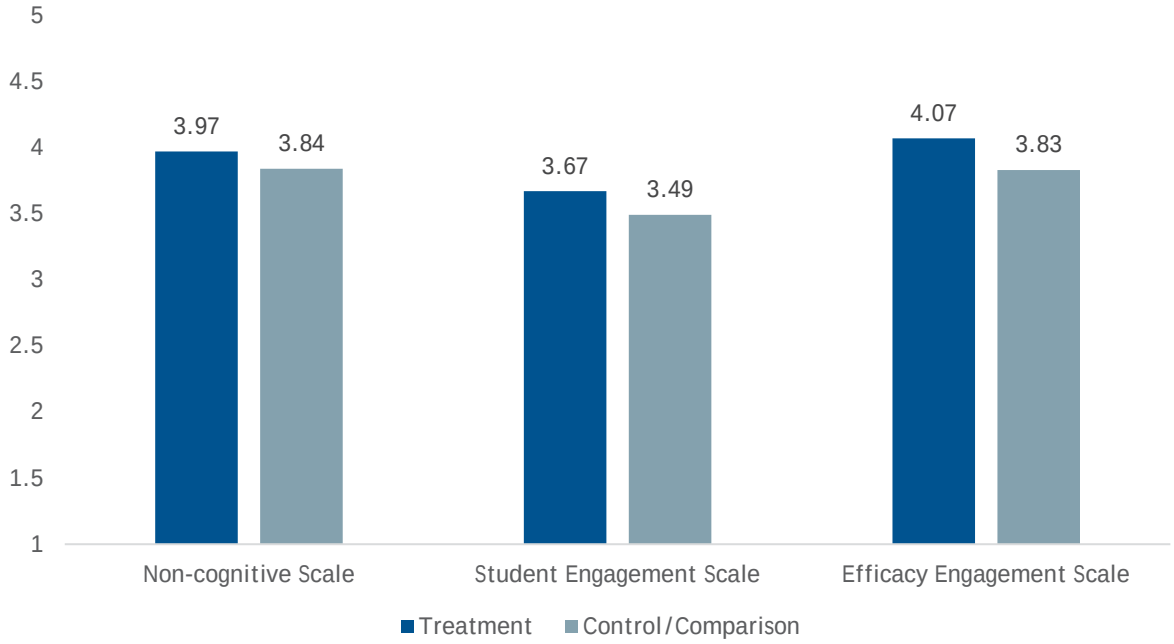
Note: The number of cases per group from left to right were:1,105, 1,187, 237, and 237.

Exhibit 12 graphically summarizes the results for the three student scales. Findings from the pretest-mid-test analysis were not substantial in terms of effect sizes (range: 0.00 ~ 0.07) and thus only those from the pretest-posttest were presented. The values represented for the control/comparison group (3.84, 3.49, 3.83) were the unadjusted averages of the group and the values used for the treatment group (3.97, 3.67, 4.07) were derived from the multivariate

[3] For example, the pretest-mid-test treatment group value 928 was derived as the sum of the program impact estimate 21 (from Exhibit 10) and the control group average 907.

models[4]. The treatment group students had a higher average score than the comparison group students on all three scales. The standardized differences, as shown earlier in Exhibit 10, were 0.22 (Non-cognitive scale), 0.23 (Engagement scale), and 0.32 (Efficacy scale).

Exhibit 12. Program Impact Analysis Results for Three Student Scales Adjusted Average Posttest Scores per Group (Pretest-Posttest Only)



Note: The number of cases for the treatment and comparison groups were, respectively, 225 and 218

5.5 Subgroup Impact Analysis

To explore the possibility that CORE program impact on students' CWRA+ scores and student scales may be more substantial within certain subgroups, we conducted a series of impact analyses on subgroups. The analytical model was the same as the confirmatory and exploratory analyses with the same set of covariates. There were four outcome variables, six subgroups (male, female, minority, white, students with a college-educated parent or parents, and students without college-educated parents) and regions. Similar to the confirmatory analysis described above, there were two analysis approaches: the one-year program analyses (pretest-to-mid-test) and the two-year program analyses (pretest-to-posttest). Due to numerous contrasts, or multiple statistical tests for within-group differences, the analyses are exploratory and only standardized effect sizes equal to or greater than 0.25 are considered important. The ICF team also considered consistency of findings from the one-year analysis (pretest to mid-test) to the two-year analysis (pretest to posttest) important in deciding which results to highlight, as the consistency over time may indicate a more reliable pattern. Statistical tests were conducted and provided as a reference; however, these subsamples were statistically underpowered.

[4] For example, the pretest-mid-test treatment group value for non-cognitive scale 3.97 was derived as the sum of the program impact estimate 0.13 (from Exhibit 10) and the control group average 3.84 (Exhibit AD5).

5.5.1 Gender

The finding for gender was not conclusive, as the results differed depending on the contrast.

No differences by gender were found for CWRA+ scores, as one form of program impact. As shown in Exhibit 13, the pretest-mid-test finding suggested a potential pattern, with 0.09 and 0.16 as the standardized effects for male students and female students, respectively. These were both small effects, but the impact on female students as shown by the larger effect size at mid-test was greater than the impact on male students. However, this pattern was not replicated at the time of posttest: the program impact for male students was 0.30 and that for female students was 0.26.

Findings on the three student scales were mixed: the sizes of program impact on these variables were all close to zero from the pretest-mid-test analysis, and inconsistent from the pretest-posttest analysis. Program impact on non-cognitive scores seemed greater for males than for females at posttest. The result for student efficacy score was reversed (the program impact appeared greater for females than for males).

Exhibit 13. Gender-Specific Program Impacts on Student Outcomes

	Male Sample		Female Sample			
	Standardized Effect	S. Test	Standardized Effect	S. Test	Average Group Difference	Statistical Model
CWRA+ Score						
Pre-Mid	0.09	ns	0.16	ns	-0.06	HLM
Pre-Post	0.30	ns	0.26	ns	0.04	HLM
Non-cognitive Skills Score						
Pre-Mid	-0.03	ns	0.07	ns	-0.10	OLS
Pre-Post	0.29	ns	0.19	ns	0.10	OLS
Engagement Score						
Pre-Mid	-0.05	ns	0.06	ns	-0.11	HLM
Pre-Post	0.32	*	0.25	ns	0.07	OLS
Efficacy Score						
Pre-Mid	0.00	ns	0.11	ns	-0.10	OLS
Pre-Post	0.27	ns	0.35	**	-0.09	OLS

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. Statistical model column indicates whether the model is HLM (Hierarchical Linear Modeling) or OLS (Ordinary Least Square) model. Number of cases: Male sample (from top to bottom row), CWRA+ pre-mid: 1,077; pre-post 190; Student scales pre-mid: 924; pre-post 174; Female sample, CWRA+ pre-mid: 1,155; pre-post 274; Student scales pre-mid: 1,012; pre-post 259.

5.5.2 Race and Ethnicity

Findings for this subgroup analysis suggest that program impact was more substantial for white students, as demonstrated through CWRA+ scores. For other student scales, the findings are mixed.

As shown in Exhibit 14, program impacts on students' CWRA+ scores appeared to be greater for white students than for minority students, with some caveats. Specifically, none of the program impacts were statistically significant. However, the differences across student groups was replicated in both samples: the pretest-to-mid-test program effects for minority and white

students were, respectively, 0.10 and 0.19. The pretest-to-posttest program effects were 0.12 and 0.25.

The findings for the three other student scales were mixed and inconsistent. All pretest-to-mid-test impact estimates were not large, with 0.11 being the largest effect on the efficacy scale from pretest-to-mid-test. For pretest-to-posttest scores, the program effect on students' non-cognitive scores was about the same for the two groups.

Contrary to the CWRA+ findings, program effects on engagement and efficacy scores were greater for minority students than for white students. The differences between program effects on engagement scores across the two student groups were not statistically significant, while the differences on efficacy scores were significant for the pretest-to-posttest sample. While interesting, this finding is not consistent with the pretest-to-mid-test findings, and thus should be interpreted with caution. When findings are not consistent across the two samples analyzed for the study (pretest to mid-test; pretest to posttest), there is less confidence in the existence of a clear pattern or trend.

Exhibit 14. Race-specific Program Impacts on Student Outcomes

	White Students		Minority Students			
	Standardized Effect	S. Test	Standardized Effect	S. Test	Average Group Difference	Statistical Model
CWRA+ Score						
Pre-Mid	0.19	ns	0.10	ns	0.09	HLM
Pre-Post	0.25	ns	0.12	ns	0.12	HLM
Non-cognitive Skills Score						
Pre-Mid	0.10	ns	-0.02	ns	0.12	HLM
Pre-Post	0.23	ns	0.26	ns	-0.02	OLS
Engagement Score						
Pre-Mid	0.08	ns	-0.07	ns	0.15	HLM
Pre-Post	0.21	ns	0.36	ns	-0.15	OLS
Efficacy Score						
Pre-Mid	0.11	*	-0.01	ns	0.12	OLS
Pre-Post	0.28	*	0.37	**	-0.09	OLS

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. Statistical model column indicates whether the model was HLM (Hierarchical Linear Modeling) or OLS (Ordinary Least Square) model. Number of cases: White sample (from top to bottom row), CWRA+ pre-mid: 1,098; pre-post 188; Student scales pre-mid: 956; pre-post 179; Minority sample, CWRA+ pre-mid: 1,141; pre-post 278, Student scales pre-mid: 983; pre-post 256.

5.5.3 Parents' Education Level

Most contrasts examined showed that there was no program impact difference between students whose parents had different levels of education. However, results suggest that program effects are potentially greater on non-cognitive skills scores for students with at least one parent who graduated college, compared to those who do not have at least one parent who graduated college. This was consistent for the two analysis samples (pretest to mid-test; pretest to posttest). For the pretest-mid-test analysis, program effects were 0.11 for students with a college-educated parent and 0.01 for students without a college-educated parent (see Exhibit 15). For the pretest-posttest analysis, the effects were 0.31 and 0.13 respectively, for students

with at least one college-educated parent and those without. Other contrasts did not show a pattern indicative of differential program impact.

Exhibit 15. Parents' Education Level-Specific Program Impacts on Student Outcomes

	BA Parent		No BA Parent			
	Standardized Effect	S. Test	Standardized Effect	S. Test	Average Group Difference	Statistical Model
CWRA+ Score						
Pre-Mid	0.18	ns	0.11	ns	0.07	HLM
Pre-Post	0.25	ns	0.24	ns	0.01	HLM
Non-cognitive Skills Score						
Pre-Mid	0.11	ns	0.01	ns	0.10	HLM
Pre-Post	0.31	*	0.13	ns	0.18	OLS
Engagement Score						
Pre-Mid	0.11	ns	-0.04	ns	0.16	HLM
Pre-Post	0.21	ns	0.28	ns	-0.07	OLS
Efficacy Score						
Pre-Mid	0.11	ns	0.07	ns	0.05	HLM
Pre-Post	0.29	ns	0.33	**	-0.04	HLM

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. Statistical model column indicates whether the model was HLM (Hierarchical Linear Modeling) or OLS (Ordinary Least Square) model. Number of cases: BA parent sample (from top to bottom row), CWRA+ pre-mid: 714; pre-post 179; Student scales pre-mid: 632; pre-post 170; No BA parent sample, CWRA+ pre-mid: 1,504; pre-post 283; Student scales pre-mid: 1,288; pre-post 261.

5.5.4 Pretest CWRA+ Levels (Low and High)

Most contrasts for this exploratory analysis supported the presence of differential program impact; however, there were no significant findings indicating a relationship between CWRA+ pretest scores and mid-test or posttest scores, or with non-cognitive skills scores or efficacy scores. There was a non-significant difference on students' engagement scores: program effects appeared to be greater for students whose pretest CWRA+ score was lower than the median than for students whose scores were equal to or higher than the median (see Exhibit 16).

Exhibit 16. Pretest CWRA+ Level-specific (Low and High) Program Impacts on Student Outcomes

	Low Pretest CWRA+		High Pretest CWRA+			
	Standardized Effect	S. Test	Standardized Effect	S. Test	Average Group Difference	Statistical Model
CWRA+ Score						
Pre-Mid	0.13	ns	0.18	ns	-0.05	HLM
Pre-Post	0.28	ns	0.23	ns	0.06	HLM
Non-cognitive Skills Score						
Pre-Mid	0.05	ns	0.05	ns	0.01	OLS
Pre-Post	0.17	ns	0.21	t	-0.04	OLS
Engagement Score						
Pre-Mid	0.08	ns	-0.01	ns	0.09	HLM
Pre-Post	0.47	ns	0.08	ns	0.39	HLM

	Low Pretest CWRA+		High Pretest CWRA+			
	Standardized Effect	S. Test	Standardized Effect	S. Test	Average Group Difference	Statistical Model
Efficacy Score						
Pre-Mid	0.04	ns	0.10	ns	-0.06	OLS
Pre-Post	0.29	*	0.32	ns	-0.03	OLS

Notes: Statistical significance (2-tail test): ns = not significant, * = p<.05, ** = p<.01, *** = p<.001. Statistical model column indicates whether the model was HLM (Hierarchical Linear Modeling) or OLS (Ordinary Least Square) model. Number of cases: Low Pretest CWRA+ sample (from top to bottom row), CWRA+ pre-mid: 1,143; pre-post 191; Student scales pre-mid: 971; pre-post: 176; High Pretest CWRA+ sample, CWRA+ pre-mid: 1,149; pre-post 283; Student scales pre-mid: 1,006, pre-post 267.

5.5.5 Gender, Race and Ethnicity

Additional subgroup analyses classified students into four categories based on gender and race: minority female students, white female students, minority male students, and white male students. Based on the same HLM framework, the program effects on CWRA+, the non-cognitive skills scale, engagement scale, and efficacy scale were estimated per each subgroup. Findings for the pretest-mid-test analysis were mixed and inconsistent (see Appendix C, Exhibit AC2). The findings from the first study phase (pretest-mid-test) showed that all effect sizes were small and not significant. The findings also show that the effect size was slightly greater for white female students, but this was not replicated by the pretest-posttest analysis.

The same analysis was repeated for the pretest-posttest sample, with some interesting findings, particularly for female minority students and male white students (see Exhibit 17). Program effects were larger for female minority students than most other subgroups on CWRA+ scores (0.32), the non-cognitive skills scale (0.38), engagement scale (0.45), and efficacy scale (0.45). White male students also had large program effects on two of the four outcomes: CWRA+ (0.46) and the non-cognitive skills scale (0.38). Minority male students experienced a large program effect on engagement scores (0.43). When all subgroups were considered, program impact seemed greater consistently for female minority students than other groups based on scores on the various assessments.

Exhibit 17. Pretest-Posttest Subgroup Program Impact Estimates for Different Student Subgroups

Subgroups	T	C	T	C	Estimate	Sig.	Standardized Effect	Model
CWRA+ Change Score								
Minority female	5	5	80	76	50.87	ns	0.32	HLM
White female	5	5	57	60	21.99	ns	0.13	HLM
Minority male	6	5	61	59	23.13	ns	0.14	HLM
White male	5	5	32	36	95.53	*	0.46	OLS
Non-Cognitive Skills Change Score								
Minority female	5	4	74	71	0.22	ns	0.38	HLM
White female	5	5	57	56	0.05	ns	0.08	OLS
Minority male	6	5	56	53	0.16	ns	0.25	OLS

Subgroups	T	C	T	C	Estimate	Sig.	Standardized Effect	Model
White male	5	5	31	32	0.23	ns	0.38	HLM
Student Engagement Score								
Minority female	5	4	74	71	0.34	ns	0.45	HLM
White female	5	5	57	56	0.10	ns	0.13	HLM
Minority male	6	5	56	53	0.34	ns	0.43	HLM
White male	5	5	31	32	0.19	ns	0.25	OLS
Student Efficacy Score								
Minority female	5	4	74	71	0.31	**	0.45	OLS
White female	5	5	57	56	0.17	ns	0.22	OLS
Minority male	6	5	56	53	0.21	ns	0.26	OLS
White male	5	5	31	32	0.22	ns	0.28	OLS

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. Statistical model column indicates whether the model was an HLM (Hierarchical Linear Modeling) or an OLS (Ordinary Least Square) model.

5.5.6 RUP Affiliation

ICF explored how program impact varied by the RUPs that recruited local schools for the study and supported participating schools throughout the life of the study. The number of schools and students included in the samples are small at this level of analysis and some analysis samples failed the baseline equivalence test (see Appendix C, Exhibit AC3). Thus, these findings are for exploration purposes only. Only CWRA+ findings are presented below because results from the three student scales produced too many contrasts with mixed, non-interpretable findings.

As shown in Appendix C, Exhibit AC3, (RUP-specific Program Impacts on Student Outcomes), Fayetteville State University's results are interesting. In both pretest-mid-test and pretest-posttest analyses, program effects were larger than those estimated for schools affiliated with other RUPs (pretest-mid-test 0.18 and pretest-posttest 0.56). However, the baseline equivalence for these analyses were not met such that students had higher pretest scores than control/comparison students. The result from Louisiana Tech University was interesting in that the pretest-mid-test program impact was large and negative (-0.53); however, this was based on a small sample and baseline equivalence was not established. Furthermore, the large negative effect finding was not corroborated by the pretest-posttest result (-0.06; negative but close to zero). Tarleton State University also had a large program impact (0.54) from pretest-mid-test analysis. The pretest-posttest data were not available to corroborate this finding.

The analysis team concluded that RUP university-level analyses were not sufficiently powered, and findings should be interpreted with caution.

5.6 Additional Exploratory Analysis of Program Implementation and Student Outcomes

ICF examined how program impact may vary by other intervention characteristics. Available data for these exploratory analyses included:

- **Exposure data (student and teacher link):** For each treatment student, data were available indicating whether they were taught by a program participant teacher (i.e., teachers participating on CORE teams). ICF requested that teachers report whether they taught students participating in the study.
- **Implementation data (Key Components 1 to 7):** As described later in the Implementation Study section, program fidelity of implementation is captured through seven Key Components. Detailed findings on activities related to each component and its associated indicators is described with implementation study results.
- **Teacher 2gno.me data:** One program component involves treatment teachers' completion of the 2gno.me assessment at three timepoints: pretest, mid-test, and posttest; designed to measure their proficiency and growth in seven areas: learner, leader, citizen, collaborator, designer, facilitator, and analyst.
- **Change-management survey data:** A JSU partner, Civitas Learning, provided an online survey to measure the level of organizational change and capacity within CORE schools. Treatment schools were provided a report and debrief of this assessment, to support the ongoing implementation activities.

The ICF team explored how three of the four data sources relate to program impact on students' growth/change in CWRA+ scores (pretest-to-mid-test; pretest-to-posttest). The change-management survey scales were used for informing the treatment schools of their organizational orientation and school culture traits; however, the data were not used for the evaluation analysis. The treatment schools participated in the survey separately at three different timepoints, which made it difficult to use along with outcome measures. Preparatory analysis did not find systematic patterns in the data as the sample size was small per each measurement point. In the following sections, findings from the exposure data analysis, the implementation data (Key Components) analysis, and teacher 2gno.me data analysis are reported. The resulting patterns are informative and relevant for future studies of the CORE program effectiveness.

5.6.1 Analysis of Exposure Data and CWRA+ Outcomes

This analysis focuses on students' "exposure to the intervention" data. As mentioned above, students at treatment schools were taught by varied numbers of team teachers. Some students in the treatment schools may never have been taught by any of the teachers participating in CORE, while others have been taught by as many as five or more CORE teachers. Some CORE teams had more than five teachers participating. In other cases, there may have been teacher turnover on the CORE team, making it more likely that treatment students were taught by a larger number of teachers who had been a part of CORE.

Results for the one-year and two-year program analyses are reported separately in Exhibit 18 and Exhibit 20. The number of team teachers linked to students through course enrollment is coded differently for the two phases. For the one-year program analysis, the number of team teachers per student was categorized as no teacher, one teacher, two teachers, three teachers, four teachers, and five or more teachers. This detailed coding was possible because the sample was a low-attrition sample, and the data were close to intact. However, the pretest-to-posttest dataset was smaller. To summarize students' two years of program experience, the number of teachers students were exposed to from the first and second year were summed. To maximize

the number of cases per group, students were categorized into broader categories: students with zero to three teachers, four to seven teachers, and eight to twelve teachers.

The one-year program analysis focused on pretest to mid-test CWRA+ change scores and the three student survey scales. As shown in Exhibit 18, students' average change scores on these instruments were reported per six treatment subgroups based on the number of teachers who taught treatment students. The control group's average score was reported as a reference point.

These analyses are exploratory, and estimates were not adjusted for covariates. To facilitate interpretation, the standardized average column shows the standardized average scores. The control group's average score was fixed as zero and each subgroup's z-score estimate was considered a deviation from the zero. T-test results were provided with an asterisk (alpha, 0.05). A discussion of results follows.

Exhibit 18. The One-Year Program Analysis: Pretest to Mid-test CWRA+ Change Score by Subgroup Defined by CORE Teacher Exposure

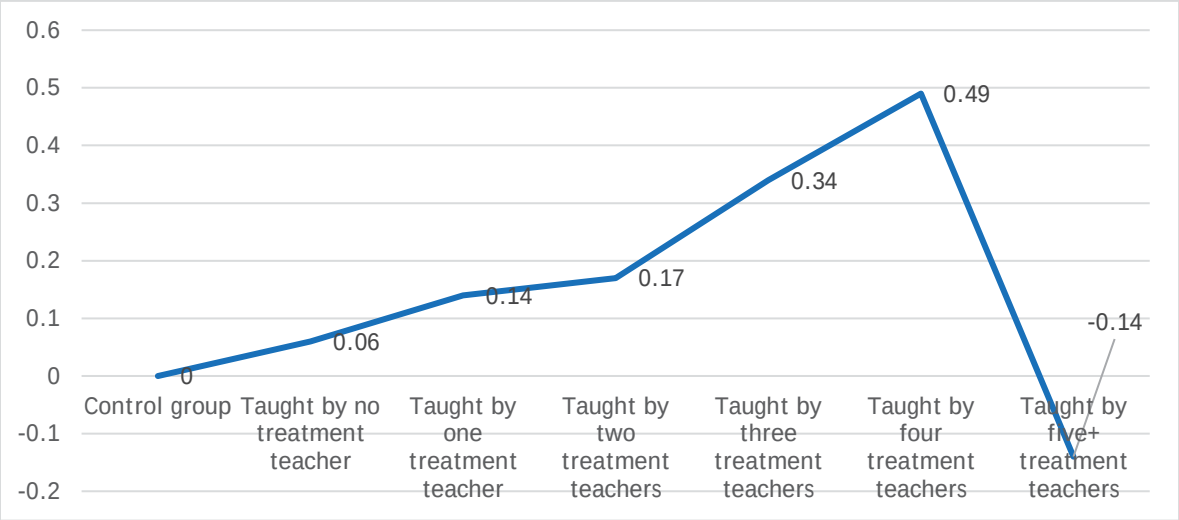
	N	Mean	Std Dev	Standardized Average	Statistical Test
CWRA+ Change Score					
Control group	1,187	-25.22	165.72	0.00	
Treatment -Taught by no treatment teacher	169	-14.63	164.47	0.06	
Treatment -Taught by one treatment teacher	341	-1.35	161.44	0.14	*
Treatment -Taught by two treatment teachers	205	2.77	168.15	0.17	*
Treatment -Taught by three treatment teachers	133	31.35	160.42	0.34	*
Treatment -Taught by four treatment teachers	120	56.78	196.71	0.49	*
Treatment -Taught by five+ treatment teachers	135	-47.65	161.40	-0.14	
Non-Cognitive Skills Change Score					
Control group	1,031	0.00	0.60	0.00	
Treatment -Taught by no treatment teacher	145	-0.07	0.57	-0.12	
Treatment -Taught by one treatment teacher	295	0.06	0.55	0.09	
Treatment -Taught by two treatment teachers	177	0.03	0.57	0.04	
Treatment -Taught by three treatment teachers	116	0.01	0.60	0.01	
Treatment -Taught by four treatment teachers	112	-0.02	0.60	-0.04	
Treatment -Taught by five+ treatment teachers	99	0.02	0.46	0.03	
Student Engagement Score					
Control group	1,031	0.00	0.82	0.00	
Treatment -Taught by no treatment teacher	145	-0.08	0.79	-0.10	

	N	Mean	Std Dev	Standardized Average	Statistical Test
Treatment -Taught by one treatment teacher	295	0.06	0.84	0.07	
Treatment -Taught by two treatment teachers	177	0.01	0.84	0.01	
Treatment -Taught by three treatment teachers	116	0.01	0.78	0.02	
Treatment -Taught by four treatment teachers	112	-0.07	0.79	-0.09	
Treatment -Taught by five+ treatment teachers	99	0.04	0.77	0.04	
Student Efficacy Score					
Control group	1,031	-0.01	0.79	0.00	
Treatment -Taught by no treatment teacher	145	-0.09	0.82	-0.10	
Treatment -Taught by one treatment teacher	295	0.04	0.73	0.07	
Treatment -Taught by two treatment teachers	177	0.09	0.76	0.13	
Treatment -Taught by three treatment teachers	116	0.08	0.79	0.12	
Treatment -Taught by four treatment teachers	112	0.00	0.74	0.02	
Treatment -Taught by five+ treatment teachers	99	-0.01	0.56	0.00	

Note: The values in the z-score column can be interpreted as standardized effect sizes. Before analysis, student outcomes were standardized with a control group mean of zero and control group standard deviation of one. The average z-score values for the control group, therefore, are zeroes and other groups' values are standardized deviation values from the control group value. T-tests were also conducted. An asterisk indicates that the group average was different from the control group estimate with statistical significance (alpha level 0.05).

The results from the pretest-to-mid-test CWRA+ change scores were mostly consistent with the expectation that the more exposure to treatment teachers, the higher the student outcome change scores. For ease of interpretation, Exhibit 19 graphically represents the CWRA+ finding.

Exhibit 19. Graphical Representation: Pretest to Mid-test Student Average CWRA+ Change Outcomes by Subgroup Defined by CORE Teacher Exposure



The standardized z-score averages were larger as the number of teachers increased, except for the last group (five+ treatment teachers)[5], and most of the contrasts were statistically significant (see asterisks). When students were taught by three treatment teachers or by four treatment teachers, the standardized differences from the control group's estimate were, respectively, 0.34 and 0.49. These values are greater than 0.25, which WWC considers "substantively important.[6] This simple analysis supports the idea that program impact may be mediated by the direct link (though course enrollment) between students and the teachers who directly participate in the CORE program. It is interesting that the last group of students with the largest number of teachers linked to them had a lower average-change score. The ICF team hypothesized that students exposed to a larger number of teachers who participated on CORE teams may be in schools that experienced a higher teacher turnover rate from CORE teams. When teachers stopped being part of the team during the school year, new teachers were recruited, which inflated the number of teachers students were exposed to in enrolled courses. The program hours that these teachers experienced were most likely shorter than teachers in other schools, weakening program impact on the nature of their teaching and resulting student outcomes. The data, however, did not affirmatively prove this explanation. Students with high levels of exposure (those linked to a greater number of CORE team teachers) were not always found in the schools with a high teacher-turnover rate on CORE teams. Still the explanation here is theoretically reasonable and a future research study should further investigate how high turnover from the CORE team may affect the intervention's effectiveness.

Results from the three student scales were mixed and none of the between-group average score differences were large in terms of standardized group score differences.

The same analysis was applied to the two-year program sample (pretest-to-posttest data). As previously mentioned, this dataset is substantially smaller than the one-year sample and thus

[5] The evaluation team further analyzed the data to explain the 5+ teacher category being negatively associated with CWRA+ outcomes. We did not find the reason for this in the data pattern.
[6] Interpretation requires caution as the exposure variable was not created prior to randomization. When reevaluated in the more complex multivariate modeling analysis, the results were not statistically significant.

the analysis is statistically underpowered. Because the study period covered two school years, the number of teachers involved was greater. The sum of the number of first year and second year teachers was used as a predictor variable. The number of students per subgroup (defined by the number of team teachers who taught each student) was maximized by using broader categories as previously described: (1) zero to three teachers, (2) four to seven teachers, and (3) eight to twelve teachers. As shown in Exhibit 20, the average pretest-to-posttest change scores on CWRA+ and the three student scales were analyzed by the subgroups defined by the number of team teachers who taught students. Exhibit 21 shows a graphical presentation of the same findings.

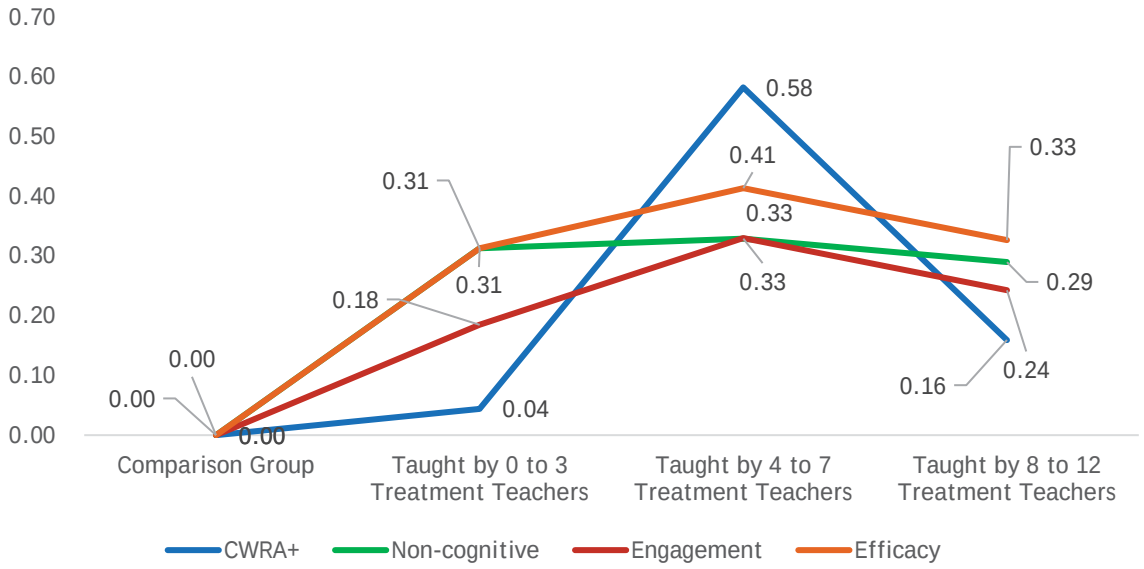
Exhibit 20. The Two-Year Program Analysis: Pretest to Posttest CWRA+ Change Score by Subgroup Defined by CORE Teacher Exposure

	N	Mean	Std Dev	Standardized Deviation	Statistical Test
CWRA+ Change Score					
Comparison group	237	-116.83	163.08	0.00	
Treatment — Taught by **zero to three** teachers	38	-109.66	132.40	0.04	
Treatment — Taught by **four to seven** teachers	91	-21.87	191.21	0.58	*
Treatment — Taught by **eight to twelve** teachers	43	-90.84	182.41	0.16	
Non-cognitive Skill Change Score					
Comparison group	218	-0.08	0.60	0.00	
Treatment — Taught by **zero to three** teachers	37	0.11	0.51	0.31	*
Treatment — Taught by **four to seven** teachers	84	0.12	0.56	0.33	*
Treatment — Taught by **eight to twelve** teachers	40	0.10	0.66	0.29	
Student Engagement Score					
Comparison group	218	-0.08	0.85	0.00	
Treatment — Taught by **zero to three** teachers	37	0.07	0.67	0.18	
Treatment — Taught by **four to seven** teachers	84	0.20	0.88	0.33	*
Treatment — Taught by **eight to twelve** teachers	40	0.13	0.94	0.24	
Student Efficacy Score					
Comparison group	218	-0.21	0.73	0.00	
Treatment — Taught by **zero to three** teachers	37	0.02	0.60	0.31	*
Treatment — Taught by **four to seven** teachers	84	0.09	0.61	0.41	*
Treatment — Taught by **eight to twelve** teachers	40	0.03	0.81	0.33	

Note: T-tests were also conducted. An asterisk indicates that the group average was different from the control group estimate with statistical significance (alpha level 0.05).

A similar trend, as found in the previous section, was found with the pretest-to-posttest analysis. Exhibit 21 graphically summarizes the findings.

Exhibit 21. Graphical Representation: Pretest to Posttest Student Average Change Outcomes by Subgroup Defined by CORE Teacher Exposure



The average pretest-to-posttest CWRA+ change score was the lowest for the comparison students (they are referred to here as comparison students as the study is no longer a RCT but a QED study). The average was slightly higher for the group taught by zero to three treatment teachers (0.04) and it was substantially higher for the group taught by four to seven treatment teachers (0.58). Like the pretest-to-mid-test finding, the group taught by eight to twelve teachers had a lower average score (0.18) than other treatment subgroups.

The three student scales followed the same pattern (see Exhibit 21 above), supporting the idea that the program effect (measured by positive score change on the student scales) seems positively correlated with the number of CORE teachers who teach them. This was not contradictory to the pretest-mid-test result but was not exactly consistent in the magnitude of standardized effect sizes. The previous section showed that the average treatment subgroups' student survey scores did not differ substantially from the comparison group (the highest standardized average change score was only 0.13 for one of the treatment subgroups for the efficacy score).

Consistent with the one-year program analysis, the "highest" group (students taught by eight to twelve treatment teachers) had a lower change score in all four outcomes; CWRA+ scores and all three student scales. Again, it is possible that this group is represented by the schools with a high turnover rate of faculty members, although we cannot verify. We did not find an evidence for this explanation. The team inspected that data and did not find a large correlation between team teacher turnover rate and the number of teachers that students were linked to via course enrollment.

5.6.2 Analysis of Implementation Data (Key Components) and Student Outcomes

The goal of this analysis is to understand how student outcomes and the level of program implementation by treatment schools are correlated. As previously described and covered in more detail in the Implementation Study section (Section III), these data were collected from teachers and administrators to understand the degree to which schools have implemented the seven CORE program components as intended. Per each KC, the evaluation team classified the 14 treatment schools into three levels of program implementation (low, medium, and high). KC3 and KC5 were excluded from this analysis due to the lack of variation in the predictor variable. In other words, all schools attained high fidelity on these two KCs, thus no associations could be determined between different levels of implementation and student outcomes. To understand how the three levels of implementation are correlated with student outcome changes (CWRA+ and the three student scales), we derived average student change scores by the three levels of implementation (low, medium, and high). As a reference, average scores for the comparison group were also calculated. As the analyses were exploratory in nature and the number of students in the pretest-to-posttest sample was small, we used simple descriptive statistics; statistical tests were provided only as a reference.

Like other analyses in this report, we report the pretest-to-mid-test findings and the pretest-to-posttest findings separately. Because there were four student outcomes, seven Key Components, and three levels of implementation, this analysis generated many statistics for inspection and interpretation. We focus only on students' CWRA+ change scores (growth scores from pretest to mid-test) because the results included those of substantive importance (standardized group difference greater than 0.25). The results for the three other student survey scores (non-cognitive skills, student engagement, and student efficacy) were negligible in group differences (results are available upon request).

As shown in Exhibit 22, there did not appear to be a link between KC1 (principal engagement) and changes in student outcome scores. From pretest-to-mid-test, the low implementation school had an even lower average score (-0.25), while the medium-level implementation showed the highest CWRA+ pretest-mid-test change scores (0.24). This is followed by the highest implementation group's average score being not so different from the control group (-0.01). Similarly, the pretest-to-mid-test findings showed that lower implementation levels across the two school years (medium-to-low and medium-to-medium implementation over time) had a higher level of outcome changes than other subcategories, such as the medium to high level and the high to high level. The pretest-to-posttest estimates are based on a small number of students and thus are likely unstable and unreliable.

Exhibit 22. KC1 and CWRA+ Change Score Averages (Pretest to Mid-test)

KC1: *CORE principals engage in professional learning with school teams.*

Pretest-to-Mid-test Analysis					
Implementation Level	**N**	**Mean**	**SD**	**Standardized**	**Sig.**
Control Schools	1,187	-25.22	165.72	0.00	
Low	45	-66.76	193.01	-0.25	
Med	803	15.16	166.63	0.24	**

KC1: *CORE principals engage in professional learning with school teams.*					
High	256	-27.00	166.35	-0.01	
Pretest-to-Posttest Analysis					
Comparison Schools	237	-116.83	163.08	0.00	
Med-Low	27	-27.19	141.92	0.55	**
Med-Med	93	-24.47	190.21	0.57	**
Med-High	63	-90.95	140.77	0.16	
High-High	54	-125.28	171.50	-0.05	

Note: In the pretest-to-posttest analysis column, school categories are represented by their level of implementation on KC1 in year 1 and year 2, as this KC was measured on an annual basis (e.g., "med-low" indicates a medium level of implementation in year 1, and low implementation in year 2). Statistical significance (2-tail test): * = $p<.05$, ** = $p<.01$, *** = $p<.001$.

As shown in Exhibit 23, the pattern for KC2 (teachers' active participation in online program activities) did not follow the expectation that implementation level would be positively correlated with student outcome changes. The pretest-to-mid-test finding showed that the low implementation school had a higher average CWRA+ change score than students in high implementation schools. Likewise, the pretest-to-posttest analysis, based on a smaller number of cases, showed that the lowest implementation schools had the highest change score average.

Exhibit 23. KC2 and CWRA+ Change Score Averages (Pretest to Mid-test)

KC2: *School teams participate in online learning communities.*					
Pretest-to-Mid-test Analysis					
Implementation Level	**N**	**Mean**	**SD**	**Standardized Averages**	**Sig.**
Control Schools	1,187	-25.22	165.72	0.00	
Low	386	36.84	151.21	0.37	**
High	718	-16.67	175.20	-0.05	
Pretest-to-Posttest Analysis					
Comparison Schools	237	-116.83	163.08	0.00	
Low-Low	27	-27.19	141.92	0.55	**
High-Low	33	-100.61	150.53	0.10	
High-High	177	-64.69	180.60	0.32	**

Note: In the pretest-to-posttest analysis column, school categories are represented by their level of implementation on KC2 in year 1 and year 2, as this KC was measured on an annual basis (e.g., "high-low" indicates a high level of implementation in year 1, and low implementation in year 2). Statistical significance (2-tail test): * = $p<.05$, ** = $p<.01$, *** = $p<.001$.

As shown in Exhibit 24, KC4 (professional development activities) followed the expectation that the degree to which this KC is implemented correlates positively with CWRA+ change scores. The pretest-to-mid-test findings show that the low implementation school average (-0.09) was slightly lower than the control school average; however, high implementation schools had a higher level of student change scores (0.26) than other groups. The pretest-posttest results

exhibited a consistent pattern. Students in schools that were lower in implementation during both study years (low-low) had a slightly higher CWRA+ change score (0.11) than control schools. The highest implementation group (high-high) had a change score (0.42) that was substantially higher than the control group.

Exhibit 24. KC4 and CWRA+ Change Score Averages

KC4: *School teams participate in CORE instructional professional development services.*

Pretest-to-Mid-test Analysis					
Implementation Level	**N**	**Mean**	**SD**	**Standardized**	**Sig.**
Control Schools	1,187	-25.22	165.72	0.00	
Low	312	-39.94	159.91	-0.09	
High	792	18.58	169.81	0.26	**
Pretest-to-Posttest Analysis					
Comparison Schools	237	-116.83	163.08	0.00	
Low-Low	79	-98.57	161.72	0.11	
High-High	158	-48.85	176.66	0.42	**

Note: In the pretest-to-posttest analysis column, school categories are represented by their level of implementation on KC4 in year 1 and year 2, as this KC was measured on an annual basis (e.g., "low-low" indicates a low level of implementation in year 1, and low implementation in year 2). Statistical significance (2-tail test): * = $p<.05$, ** = $p<.01$, *** = $p<.001$.

KC6 (change management) followed expectations in that implementation levels were correlated positively with student change scores. The first phase of analysis (pretest-to-mid-test) showed that the low implementation group had an average pretest-mid-test change score (0.03) that was similar to the control group. As shown in Exhibit 25, the high implementation group in year 1 had the highest change score average (0.29). This KC is only measured once during the two years of the study, and no new schools were added to the high-fidelity group in year 2. Thus, the second phase of analysis (pretest-to-posttest) relies on the same school groupings, but fewer students who had available posttest data. However, the pattern found in year 1 was similar in the second year: the lower implementation group (low) had a higher level of student score change (0.28) than the control group, and the higher implementation group (high) had an even higher student change score (0.34). However, the difference between the two implementation groups was rather small (0.06).

Exhibit 25. KC6 and CWRA+ Change Score Averages

KC6: *Schools participate in change-management support through CORE partnership resources.*					
Pretest-to-Mid-test Analysis					
Implementation Level	**N**	**Mean**	**SD**	**Standardized**	**Sig.**
Control Schools	1,187	-25.22	165.72	0.00	
Low	532	-20.33	159.87	0.03	
High	572	22.85	174.79	0.29	**
Pretest-to-Posttest Analysis					
Comparison Schools	237	-116.83	163.08	0.00	
Low	90	-71.82	143.36	0.28	*
High	147	-61.50	189.35	0.34	**

Note: Statistical significance (2-tail test): * = $p<.05$, ** = $p<.01$, *** = $p<.001$.

For both the pretest to mid-test and pretest to posttest analysis phases, KC7 (use of EdReady™) had only a small variation in the predictor, as most schools achieved high fidelity the first year. As shown in Exhibit 26, the low implementation group had a small number of students: 57 for the first year and 60 for the second year). Yet with the little variance available for analysis, the pretest-mid-test findings indicated that KC7 followed expectations: the low implementation group had an average pretest-mid-test change score that was almost the same (0.01) as the control group. The high implementation group had the highest change score average (0.17). The second phase of the study (pretest-to-posttest) did not exactly follow expectations. Both treatment groups (low-low and high-low) had about the same change score averages (0.30 and 0.32). This may be because both of these groups were low-implementing schools in the second study year and thus did not differentiate themselves in terms of program efficacy.

Exhibit 26. KC7 and CWRA+ Change Score Averages

KC7: *School teams provide students with college-readiness advisement and support through use of EdReady™ tool in CORE schools.*					
Pretest-to-Mid-test Analysis					
Implementation Level	**N**	**Mean**	**SD**	**Standardized**	**Sig.**
Control Schools	1,187	-25.22	165.72	0.00	
Low	57	-23.30	162.38	0.01	
High	1,047	3.42	169.39	0.17	**
Pretest-to-Posttest Analysis					
Comparison Schools	237	-116.83	163.08	0.00	
Low-Low	60	-67.57	150.08	0.30	*
High-Low	177	-64.69	180.60	0.32	**

Note: in the pretest-to-posttest analysis column, school categories are represented by their level of implementation on KC7 in year 1 and year 2, as this KC was measured on an annual basis (e.g., “high-low” indicates a high level of implementation in year 1, and low implementation in year 2).

Out of seven KCs, the findings from three KCs were consistent or almost consistent with program expectations, as in greater fidelity of implementation was correlated positively with average CWRA+ change scores. The findings for KC4 (professional development activities) was consistent for both study phases: the low implementation group was similar to the control/comparison group in the change in student outcomes, which suggests that lower levels of fidelity on this KC may have had limited to no effect on students' CWRA+ scores, whereas implementing this KC to a greater degree, with fidelity, may help students experience growth in CWRA+ outcomes. The same implication is true for the first phase of analysis for KC6 and KC7: when fidelity of implementation was higher on these components, students' average CWRA+ scores were higher as well. These findings were not exactly replicated in the second phase of analysis for KC6 and KC7, in part because there was no change in KC6 fidelity groupings in year 2, and fidelity of implementation decreased on KC7 during this time period.

5.6.3 Correlation Analysis of School-Average CWRA+ Change Scores and Teachers' Average 2gno.me Scores.

The 2gno.me system assessed teachers' orientation and skills across seven components: analyst, citizen, collaborator, designer, facilitator, leader, and learner. The expectation was that treatment teachers, through participation in CORE, develop proficiency and achieve higher scores over time in these seven areas as they progress through the program and apply knowledge they have gained to their daily instruction. The ICF team conducted two exploratory analyses. First, ICF described how treatment status relates to teacher growth and proficiency as captured by the 2gno.me assessment. Second, the team examined how the seven 2gno.me criteria are correlated with changes in students' average CWRA+ scores.

To capture changes in the school average 2gno.me scores, the ICF team decided to use all data available at each of three timepoints and assumed that the average of teachers whose data were available at either of three timepoints helped approximate the school-level 2gno.me orientation levels. The ICF team first derived the school average 2gno.me scores and calculated the group average scores (treatment group and control group) for comparison. The other important data for this analysis was student CWRA+ scores. Change scores were first calculated at the individual level for change from pretest to mid-test, as well as for pretest to posttest. The school averages of the individual changes were then calculated and used to understand how school scores correlated with treatment status (control vs. treatment).

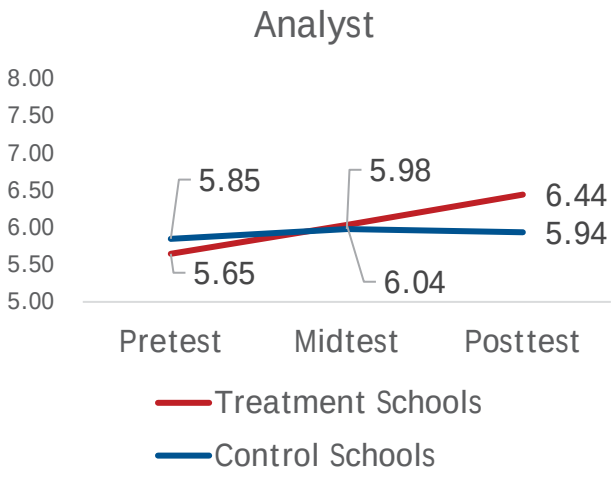
Exhibit 27 describes the whole sample and compares treatment schools and comparison schools on both CWRA+ and 2gno.me variables. Graphical representation of findings and discussion will follow.

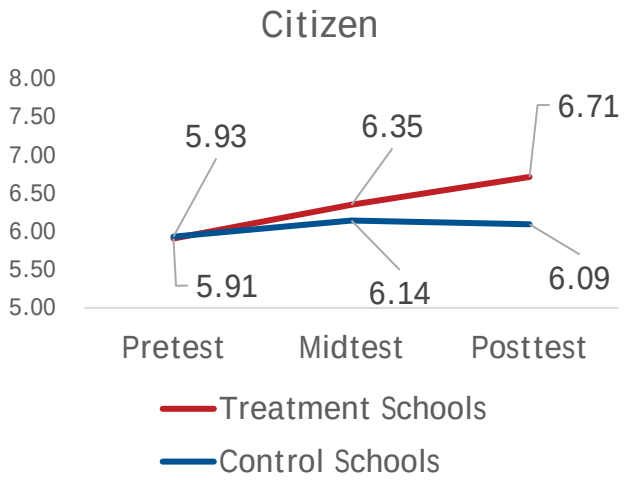
Exhibit 27. Descriptive Statistics of CWRA+ School-Average Change Scores and 2gno.me Scores by Treatment Sample and Control Sample

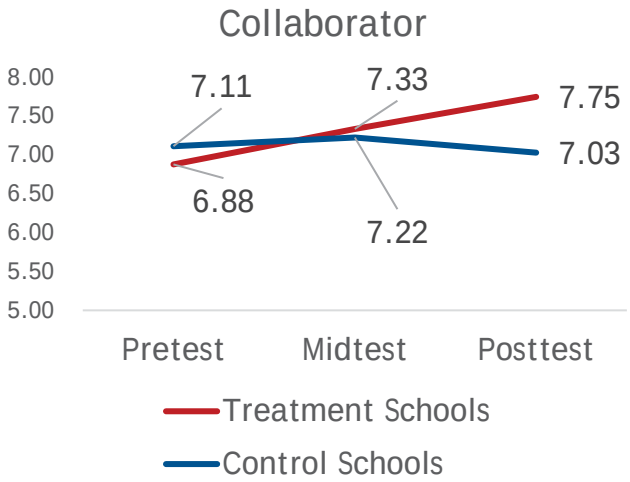
		Treatment Schools			Control Schools		
		N	Mean	SD	N	Mean	SD
CWRA+ Change Scores (pretest-mid-test)		14	-18.08	50.92	13	-36.53	44.11
CWRA+ Change Scores (pretest-posttest)		6	-83.17	54.43	5	-117.36	43.09
Analyst Score	**-Pretest**	**14**	**5.65**	**0.52**	**14**	**5.85**	**0.43**
	-Mid-test	14	6.04	0.52	13	5.98	0.66
	-Posttest	14	6.44	0.73	11	5.94	0.54
Citizen Score	**-Pretest**	**14**	**5.91**	**0.72**	**14**	**5.93**	**0.44**
	-Mid-test	14	6.35	0.77	13	6.14	1.03
	-Posttest	14	6.71	1.03	11	6.09	1.02
Collaborator	**-Pretest**	**14**	**6.88**	**0.55**	**14**	**7.11**	**0.55**
	-Mid-test	14	7.33	0.60	13	7.22	1.05
	-Posttest	14	7.75	0.71	11	7.03	0.70
Designer	**-Pretest**	**14**	**5.81**	**0.56**	**14**	**5.77**	**0.66**
	-Mid-test	14	6.48	0.75	13	5.89	1.07
	-Posttest	14	6.61	0.90	11	5.94	0.59
Facilitator	**-Pretest**	**14**	**5.22**	**0.61**	**14**	**5.45**	**0.82**
	-Mid-test	14	5.79	0.61	13	5.70	0.95
	-Posttest	14	6.10	0.55	11	5.93	0.59
Leader	**-Pretest**	**14**	**7.08**	**0.75**	**14**	**7.41**	**0.54**
	-Mid-test	14	7.60	0.66	13	7.61	1.02
	-Posttest	14	7.78	0.80	11	7.41	0.82
Learner	**-Pretest**	**14**	**6.44**	**1.10**	**14**	**6.28**	**0.62**
	-Mid-test	14	7.06	0.95	13	6.52	0.78
	-Posttest	14	7.33	1.11	11	6.43	1.07

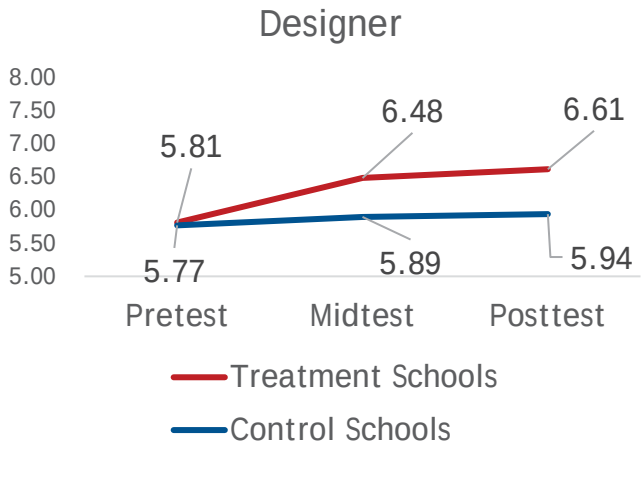
Graphics shown in Exhibit 28 were based on pretest, mid-test, and posttest 2gno.me average scores for teachers in treatment and control schools. All seven indicators exhibited a consistent trend of treatment schools having a higher average around mid-test and posttest points than at pretest. The control schools' trend was rather flat.

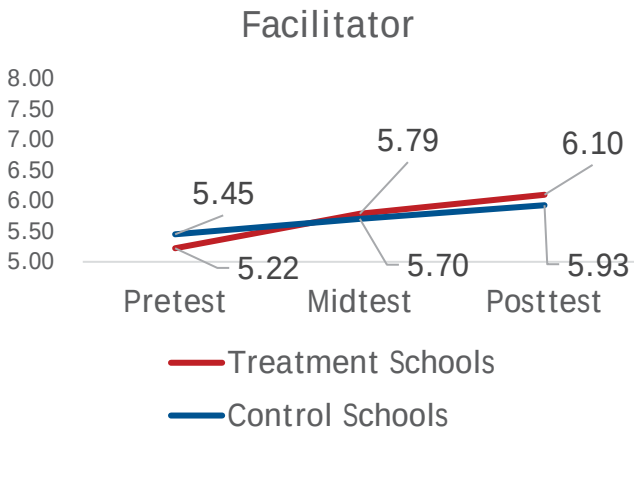
Exhibit 28. Graphical Representation of Findings from the 2gno.me Analysis

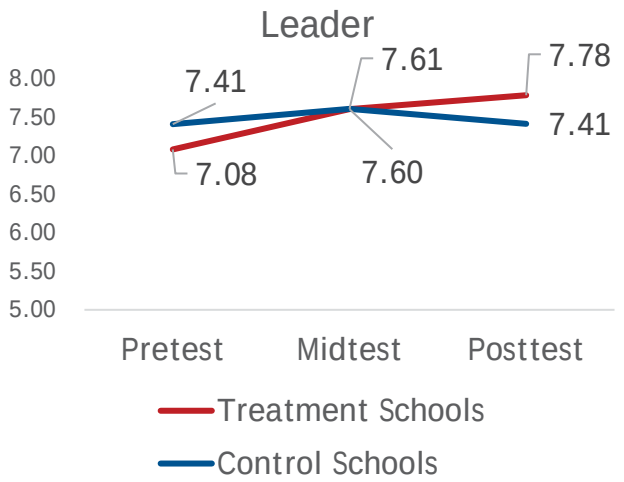


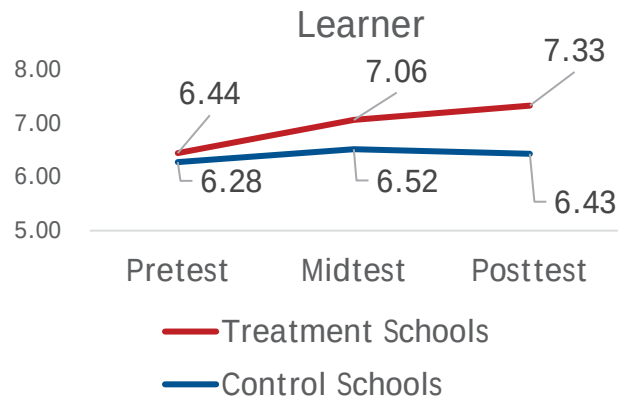












The next exploratory analysis addressed the question of whether schools that had a larger change in average 2gno.me measures also had a greater degree of improvement in student CWRA+ scores. The analysis team derived the correlation statistics between school-average 2gno.me scores and school-average CWRA+ change scores. As shown in Exhibits 29 and 30, both study phases (pretest to mid-test and mid-test to posttest) were considered. Analysis was conducted using the whole sample (reported on the left panel) and the treatment school only sample (reported on the right panel). A limitation was that the data were not sufficiently powered, due to a relatively small number of cases. There was one large correlation value (0.60) for the pretest-to-posttest designer change score and CWRA+ (See Exhibit 30). This, however, was based on six cases and was not replicated in the pretest-to-mid-test finding. Thus, it seems that exploratory analysis findings did not support the idea that teachers' 2gno.me scores are related to students' growth in CWRA+; however; more exploration may be warranted given the small number of cases available for this analysis.

Exhibit 29. Relationship Between Teacher 2gno.me Scores and CWRA+ Average Change Scores: Pretest to Mid-test

Correlation Statistics with Pretest to Mid-test CWRA+ Change Score						
	Whole Data			Treatment Schools		
	N	Pearson Correlation	P-value		Pearson Correlation	P-value
Analyst Change Score	26	0.23	0.26	14	0.13	0.65
Citizen Change Score	26	0.01	0.98	14	-0.06	0.83
Collaborator Change Score	26	0.20	0.33	14	-0.01	0.97
Designer Change Score	26	0.13	0.52	14	-0.07	0.81
Facilitator Change Score	26	0.23	0.26	14	-0.08	0.78
Leader Change Score	26	-0.02	0.90	14	-0.27	0.35
Learner Change Score	26	0.22	0.29	14	0.05	0.86

Exhibit 30. Relationship Between Teacher 2gno.me Scores and CWRA+ Average Change Scores: Pretest to Posttest

Correlation Statistics with Pretest to Posttest_CWRA+ Change Score						
	Whole Data			Treatment Schools		
	N	Pearson Correlation	P-value		Pearson Correlation	P-value
Analyst Change Score	10	-0.26	0.48	6	-0.06	0.91
Citizen Change Score	10	-0.06	0.87	6	-0.33	0.52
Collaborator Change Score	10	-0.26	0.46	6	0.15	0.78
Designer Change Score	10	-0.10	0.77	6	0.60	0.20
Facilitator Change Score	10	0.03	0.94	6	0.38	0.46
Leader Change Score	10	0.11	0.77	6	0.07	0.89
Learner Change Score	10	0.22	0.55	6	0.36	0.48

5.7 Summary of Program Impact Analysis

To summarize, program impact analysis was significantly disrupted by COVID-19 due to the effect of school closures on program implementation and data collection. To compensate for these issues, it was necessary to conduct analysis in two phases, pretest to mid-test and pretest to posttest. The pretest-to mid-test phase did not experience data collection disruptions and considered a low-attrition RCT. These data were collected before COVID-19 and school and student attrition were kept to a minimum. The exploratory hypotheses regarding one-year program impact (EEQ3, 4, 5, 6; see Exhibit 2) returned results that were positive, but not strong, suggesting that potentially greater program impacts could be realized by the end of the two-year intervention, with strong implementation. The standardized program impacts for CWRA+ scores, the non-cognitive skills scale, engagement scale, and efficacy scale, were respectively 0.12, 0.03, 0.00, and 0.07 (see Exhibit 10). These effect sizes were positive, small, and statistically not significant. Program effect on CWRA+ scores (0.12) was interesting in that although small, it was not close to zero.

The study's pretest-posttest phase examined two-year program impact, addressing the main confirmatory hypotheses (CEQ1&2, EEQ1&2; see Exhibit 1). Due to data attrition caused by COVID-19 and the need to use data-matching techniques (PSM), the sample was reduced in size and the study design shifted to a QED. The two-year program standardized effects on CWRA+ scores, the non-cognitive skills scale, engagement scale, and efficacy scale were respectively, 0.22, 0.22, 0.23, and 0.32. Non-cognitive and efficacy scales were statistically significant, but there are caveats to these results—the between-school variance for these two outcomes was too small for running the HLM model, necessitating a regular OLS model instead). Interestingly, the first three outcomes' effect size was close to 0.25, which is considered small but important by WWC Standards. The effect size for the efficacy scale (0.32) was large enough to be considered important.

Subgroup analyses examined how program impact was affected by student characteristics, including gender, race and ethnicity, parents' education level, pretest CWRA+ scores, and RUP affiliation. The only subgroup finding with consistent results in both analysis phases of the study (pretest-mid-test and pretest-posttest) was **the relationship between students' race or ethnicity and program outcomes.** Program impact, as measured by students' CWRA+ scores, was greater for white students than for minority students. There was a consistent pattern year to year: one-year program impacts for white students and minority students were, respectively, 0.19 and 0.10 (the difference 0.09). The two-year program impacts for white students and minority students were, respectively, 0.25 and 0.12 (the difference 0.13).

Three other exploratory analyses were conducted to assess program impact based on other aspects of program implementation: the relationship between student exposure to CORE teachers through instruction and change in student CWRA+ scores, fidelity of program implementation at the school and its effect on CWRA+ scores, and the relationship between teachers' 2gno.me scores and students' CWRA+ scores. The analysis team found that **students who were taught by more CORE team teachers (higher exposure) experienced better results on CWRA+ scores,** up to a point. Specifically, the positive correlation between student exposure and CWRA change scores dropped off in the highest exposure category. This may indicate that in general, students' exposure to CORE teachers is beneficial for student outcomes. However, exposure to the highest numbers of CORE teachers may be indicative of other contextual factors (e.g., a high turnover rate on the CORE team at that school). If a teacher's tenure on a CORE team is short-lived, this may mean less time in general for the teacher to be fully engaged in the CORE model.

When assessing school fidelity of implementation and any correlations to students' CWRA+ change scores, a noteworthy pattern was detected with KC4 (teachers' participation in professional development activities). Control/comparison group students and students from schools with low fidelity of implementation had similar CWRA+ change scores for both analysis phases (pretest-mid-test and pretest-posttest). In contrast, **students from schools in the high implementation group had higher average change scores than those in the comparison group**. This exploratory analysis cannot claim causality; however, this result is noteworthy, since it provides potential emerging evidence of the significance of attaining fidelity to KC4 and the direct link that aspects of this KC (teacher growth and proficiency in 2gno.me) have with student outcomes. These initial findings imply that program effects may be specifically linked to higher levels of KC4 implementation.

The 2gno.me correlational analysis was conducted to assess how teachers' 2gno.me scores and students' CWRA+ change scores were correlated. Findings showed that teachers at treatment schools had higher 2gno.me score averages at posttest compared to teachers at control schools. This is consistent with program expectations that the intervention fosters the growth of teacher skills and capabilities assessed by 2gno.me. However, the analysis did not find a consistent positive correlation between school average 2gno.me scores and school average CWRA+ score changes, thus a strong connection between teacher and student outcomes could not be established at this time.

III. Implementation Study

1. Implementation Study Introduction

A fidelity of implementation study was conducted to measure the extent to which the CORE model was implemented as intended in participating schools. The study was guided by seven evaluation questions aligned to the key intervention components (KCs) specified in the CORE program logic model: (1) meaningful collaboration among administrators and CORE team members, (2) active participation in professional learning communities for teachers, (3) provision of funding resources and support, (4) participation in CORE professional development, (5) active participation in follow-up professional development, (6) support in navigating the change-management process in participating schools, and (7) college-readiness advisement and support using the EdReady™ tool. The sections below summarize the final status of the implementation, based on program activities taking place during SY 2018-19 and SY 2019-20. Findings presented are based on results across both implementation year of the program, which aligns with the time period during which the impact study was conducted. However, note that only a portion of treatment schools have both final impact and implementation data available, due to COVID-19-related school closures in spring 2020. In other words, implementation data are available from all participating schools; however, not all schools provided both impact and implementation data, limiting the pool within which connections between data can be drawn.

2. Implementation Study Methodology

2.1 Implementation Fidelity Measurement System

Fidelity of implementation was measured using a collaboratively developed implementation fidelity measurement system that includes 11 indicators aligned to the seven KCs of the CORE program logic model. In 2017, ICF and JSU identified each initial indicator and set implementation thresholds for the 2015 CORE evaluation. KCs, associated fidelity indicators, and data sources appear in Exhibit 31.

Exhibit 31. Key Component, Indicator, and Data Sources for Fidelity of Implementation Study

Measuring Implementation Fidelity		
Key Component	**Indicator**	**Data Source**
KC1. CORE Principals engage in professional learning with school teams.	1.1 School-level collaboration with principals and the school team	Administrator course evaluation survey
	1.2 Principals participate in and complete at least one online professional learning course	Learning Management System (LMS)
KC2. School teams participate in online learning communities.	2.1 Active participation in online professional learning modules	Learning Management System (LMS)
KC3. Schools receive CORE resources and support.	3.1 Provision of school-level funds for CORE schools	Financial disbursement log
	3.2 Provision of school technology assessments	Technology assessment log

Measuring Implementation Fidelity		
Key Component	**Indicator**	**Data Source**
	3.3 School funds use	School-level summary of school funds use
KC4. School teams participate in CORE instructional professional development services.	4.1. CORE team attendance at CORE Professional Learning services orientation	Orientation attendance roster
	4.2. Use of CORE instructional model	2gno.me assessments
KC5. School teams present during professional development workshops.	5.1. Sharing of learning experiences	Video logs of presentations
KC6. Schools participate in change-management support through CORE partnership resources.	6.1. Participation in change management	Debriefing reports
KC7. School teams provide students with college-readiness advisement and support through use of EdReady™ tool in CORE schools.	7.1 School teams' utilization of EdReady™ college-readiness assessment tool	EdReady™ utilization records

2.2 Data Sources

The implementation study drew data about the 11 indicators in Exhibit 1 from the following sources: (1) attendance records, (2) course evaluation surveys, (3) LMS participation data, (4) fund disbursement and utilization records, (5) technology assessment logs, (6) video logs, (7) teacher 2gno.me pretest/posttest data, (8) change-management debriefing logs, and (9) EdReady™ utilization records.

No changes were made to definitions of other indicators or Key Components during the fidelity of implementation study.

2.2.1 Site Visits

Virtual and on-site focus groups were conducted during the 2019-20 school year with fourteen schools (five control and nine treatment). Site visits included:

- a classroom observation of a teacher from the CORE school team,
- an interview with the teacher participating in the observation,
- an interview with the administrator from the school team, and
- a focus group with the remaining teachers from the school team.

The site visits focused on gathering the information necessary to examine instructional experiences and practices among CORE team members, CORE team perspectives on learning and college and career readiness, and their suggestions for improvement as program participants. Site visits helped the evaluation team gain first-hand perspectives on the intervention and how it is being implemented across treatment schools and regions. Visits with control schools, that had comparable demographics, within each region allowed the evaluation

team to have more context about the likely capacity of schools in the study to implement the intervention.

2.3 Data Collection

All data sources were developed and maintained by JSU, with consultation from ICF. JSU oversaw data collection that began in the summer of 2018 with 2gno.me pretest data from teachers, relevant for Indicator 4.2. Initial data on disbursement of school-level funds (Indicator 3.1) was available in December 2018; additional data for this indicator became available later in the implementation cycle. Data collection for the remaining indicators progressed over SY 2018-19 and SY 2019-20 (e.g., 2gno.me mid-test data were collected at the end of SY 2018-19). Fidelity of implementation was tracked annually, but final status is based on overall implementation progress over the two program years. Final implementation data for SY 2018-19 was transmitted to ICF in the summer of 2020.

2.4 Implementation Study Analysis

Implementation study analysis was conducted after data submission completion. This section describes the analysis that took place. Individual indicator implementation scores were calculated for each of the 14 treatment schools in the study at the conclusion of SY 2019-20.[7] All 11 fidelity indicators were scored for each school. The resulting scores were then coded to represent the extent to which each school met the associated indicator's implementation threshold (typically measured as low, medium, or high). See Exhibit 32 below for descriptions of how high fidelity was operationalized for each KC. Once indicator implementation scores were derived, they were summed within each KC to arrive at a single KC implementation score for each treatment school (typically measured as low, medium, or high).

The percentage of treatment schools meeting the criteria for "high" implementation for each KC was calculated and compared to an established threshold for "high" fidelity at the program level (75% or greater). If the percentage of schools in the entire sample who meet the criteria for "high" implementation meets or exceeds this threshold, fidelity of implementation is considered "met" for the KC at the sample level.

Exhibit 32. Key Components and Definitions of High Fidelity of Implementation

Key Components	Definition of High-Fidelity Implementation
KC1. CORE Principals engage in professional learning with school teams.	1.1: **Principal** agrees or strongly agrees that engaging in professional learning services led to meaningful collaboration 1.2: **Principal** participates in at least one professional learning module during the two years of the study
KC2. School teams participate in online learning communities.	2.1: **Teacher** posts at least 11 reflections on CORE professional learning services. **School-level** fidelity is attained when 66% of

[7] The denominator for fidelity of implementation calculations includes only those teachers/schools that remain in the treatment group at the end of the study.

Key Components	Definition of High-Fidelity Implementation
	teachers have reached high fidelity.
KC3. Schools receive CORE resources and support.	3.1: **Schools** receive $25,000 in program funds in year 1. 3.2: **Schools** complete technology assessments in year 1. 3.3: **Schools** provide summaries of how funds were used.
KC4. School teams participate in CORE instructional professional development services.	4.1: **Teachers** participate in CORE orientation. 4.2: **Teachers** demonstrate proficiency and growth on at least one of the seven 2gno.me components from pretest to posttest.
KC5. School teams present during professional development workshops.	5.1: **CORE teams** present during at least one professional development workshop during the two years of the study.
KC6. Schools participate in change-management support through CORE partnership resources.	6.1: **Schools** participate in a change-management process and survey; each school completes a debrief during the two years of the study.
KC7. School teams provide students with college-readiness advisement and support through use of EdReady™ tool in CORE schools.	7.1: CORE Math and ELA **teachers** conduct assessments with students using the EdReady tool during both years of the study.

3. Implementation Study Results

Exhibit 33 below provides an overview of final implementation status. Details on the status of each component are covered in the following section. Overall, three of the seven KCs achieved high fidelity, and four were unmet at the conclusion of the study.

Exhibit 33. Overall Implementation at a Glance

Key Component	Fidelity Status
KC1. CORE Principals engage in professional learning with school teams.	Did not meet
KC2. School teams participate in online learning communities.	Met
KC3. Schools receive CORE resources and support.	Met
KC4. School teams participate in CORE instructional professional development services.	Did not meet
KC5. School teams present during professional development workshops	Met
KC6. Schools participate in change-management support through CORE partnership resources.	Did not meet
KC7. School teams provide students with college-readiness advisement and support through use of EdReady™ tool.	Did not meet

3.1 Implementation Fidelity by Key Component

Exhibit 34 presents fidelity performance data across both years of the study. The table is organized by KC and lists the corresponding percentage by year and threshold (e.g., met, not met) for the whole treatment school sample (n=14). As previously mentioned, a subset of schools provided both implementation and impact data for the last year of the study. The percentages in parentheses below reflect standing on each KC for just these schools (n=6).

Exhibit 34. Implementation Status by Year: Program Level (Subset of schools providing impact data)

Key Component	Year 1	Year 2
KC1. CORE Principals engage in professional learning with school teams.	29% (33%) Not met	50% (67%) Not met
KC2. School teams participate in online learning communities.	71% (83%) Not met	86% (67%) Met
KC3. Schools receive CORE resources and support.	100% Met	
KC4. School teams participate in CORE instructional professional development services.	71% (50%) Not met	
KC5. School teams present during professional development workshops.	100% Met	
KC6. Schools participate in change-management support through CORE partnership resources.	50% (50%) Not met	
KC7. School teams provide students with college-readiness advisement and support through use of EdReady™ tool in CORE schools.	86% (67%) Met	21% (0%) Not met

Note: Results in parentheses in this table are based on implementation data only for schools providing final-phase impact data: Schools 5, 27, 28, 41, 48, and 66.

3.2 Implementation Fidelity by Indicator

Fidelity to each indicator was assessed using the same scoring criteria established for each indicator's respective KC. For example, the threshold for high fidelity at the program level to KC3 is that 75% of the sample will achieve high implementation fidelity when data are aggregated across indicators 3.1–3.3. To make a fidelity determination separately for each individual indicator (i.e., 3.1, 3.2, and 3.3), we first assessed what percentage of the sample met the criteria for "high" fidelity on each indicator. If at least 75% of the sample met the criteria for "high" fidelity at the indicator level, we determined fidelity was "met" for the indicator.

KC1: High implementation fidelity to KC1, *CORE principals engage in professional learning with school teams*, was <u>not met</u> with high fidelity in year 1 or year 2 of implementation.

KC1 was measured annually. To achieve high fidelity each year, at least 75% of participating principals had to (1) agree or strongly agree with two survey items on the administrator course evaluation survey that focused on principal perspectives on team collaboration and (2) complete at least one professional learning module.[8] High fidelity was considered met for each school on indicator 1.1 when the principal agreed or strongly agreed with both survey items.

In year 1, four principals (about 29%) met both requirements. The percentage meeting both requirements increased to 50% in year 2; however, this is still below the threshold for meeting overall fidelity. In year 1, four principals completed the survey (representing three RUPs), however 13 principals (92%) completed at least one professional learning module. In year 2, eight principals completed the survey, and eight principals completed at least one module (seven principals did both). Nearly all principals responding to the survey in year 2 agreed or strongly agreed that they were provided opportunities to collaborate through CORE, and that they engaged in meaningful collaboration as a result.

Exhibit 35 below illustrates results by school over both years on KC1. Just over a third of schools decreased their fidelity scores over time (from either high to low or medium to low). About another third maintained their fidelity scores from year to year (at medium or at high) and about one-quarter increased their fidelity scores over time (from low to high, or medium to high).

Principal turnover may have played a role in attaining fidelity on this KC—of the five schools that decreased their fidelity score from year 1 to year 2, three had new principals in year 2. These administrators may be occupied with other responsibilities as a part of their new role, which may have contributed to less engagement in the areas captured in KC1.

Exhibit 35. Key Component 1: Fidelity by School

				Fidelity Status by School	
Key Component	Indicators	RUP	School	Year 1	Year 2
KC1. CORE Principals engage in professional learning with school teams.	1.1 School-level collaboration with principals and the school team 1.2 Principals participate in and complete at least one online professional	FSU	School #28	Medium	Medium
			School #32	Medium	Low
			School #40	Low	High
		JSU	School #1	Medium	High
			School #4	Medium	Medium
			School #5	Medium	High
			School #8	High	High

[8] Indicator 1.1 relied on two survey items, which participants indicated their level of agreement with on a five-point scale: "I was provided opportunities to collaborate with my colleagues through the professional learning services offered through the CORE project," and "I engaged in meaningful collaboration with my CORE school team members as a result of the professional learning services."

Key Component	Indicators	RUP	School	Fidelity Status by School: Year 1	Fidelity Status by School: Year 2
	learning course		School #11	High	Low
			School #13	Medium	Low
		LTU	School #27	Medium	Low
		TSU	School #41	High	High
			School #48	Medium	High
		WTAMU	School #60	Medium	Low
			School #66	High	High

KC2: Fidelity to *school teams participating in online learning communities* was <u>not met</u> in year 1 and was <u>met</u> in year 2 of implementation.

For fidelity on this KC, CORE team teachers must demonstrate active participation in online learning communities by posting at least 11 reflections each in the LMS on an annual basis. As described in Exhibit 32, 11 posts is the minimum threshold for high fidelity at the individual level. Two-thirds of the CORE team must meet this requirement to achieve high fidelity at the school level.

In year 1, about three-quarters of the total pool of participants had high implementation scores at the teacher level. Five of these schools achieved 100% high fidelity at the teacher level (i.e., all CORE team teachers had posted 11 or more reflections). Although fidelity fell short of being met in year 1, there was a demonstrated level of engagement among teachers, which was realized in year 2. In the second year, 88% of teachers achieved individual-level fidelity, and 12 of 14 schools attained school-level fidelity. A potential factor that may have influenced completion of this KC is school closures related to COVID-19. The virtual learning environment may have resulted in teachers having more time spent at a computer, as well as created an immediate need for teachers to discuss issues and seek input from their colleagues. The school that met fidelity in year 1 and did not meet fidelity in year 2 may have been influenced by teacher turnover on the team to some extent, as a few members on the team were new for the second year.

Exhibit 36. Key Component 2: Fidelity by School

Key Component	Indicators	RUP	School	Fidelity Status by School: Year 1	Fidelity Status by School: Year 2
KC2. School teams participate in	2.1 Active participation in online professional learning modules	FSU	School #28	High	High
			School #32	High	High
			School #40	High	High
		JSU	School #1	Low	High

				Fidelity Status by School	
Key Component	Indicators	RUP	School	Year 1	Year 2
online learning communities.			School #4	High	High
			School #5	High	High
			School #8	High	High
			School #11	High	High
			School #13	Low	High
		LTU	School #27	Low	Low
		TSU	School #41	High	Low
			School #48	High	High
		WTAMU	School #60	Low	High
			School #66	High	High

KC3: *Schools receive CORE resources and support* was met at the conclusion of the study.

KC3 is comprised of three indicators, all of which must be completed by each school to receive credit for fidelity by the end of the two-year study. As a part of participation in the CORE program, treatment schools each receive $25,000 in school funds at the outset of the first program year (Indicator 3.1). Each school must also participate in a survey to assess the current technology climate at the school (Indicator 3.2). Finally, schools provide reports on how the funds they received are being expended (Indicator 3.3). All schools received funding, participated in the technology assessment, and provided summaries on funding use (see Exhibit 37).

Exhibit 37. Key Component 3: Fidelity by School

				Fidelity Status by School
Key Component	Indicators	RUP	School	Overall
KC3. Schools receive CORE resources.	3.1 Provision of school-level funds for CORE schools 3.2 Completion of school technology assessments 3.3 School funds use	FSU	School #28	High
			School #32	High
			School #40	High
		JSU	School #1	High
			School #4	High
			School #5	High
			School #8	High
			School #11	High

				Fidelity Status by School
Key Component	Indicators	RUP	School	Overall
			School #13	High
		LTU	School #27	High
		TSU	School #41	High
			School #48	High
		WTAMU	School #60	High
			School #66	High

KC4: *Schools teams participate in CORE instructional professional development services* was not met at the end of the two-year study.

KC4 assesses CORE teachers' participation in instructional professional development and assesses their knowledge gains during the program. For high fidelity of implementation on Indicator 4.1, teachers must attend an orientation on professional learning services. Indicator 4.2 provides evidence for CORE teachers' **proficiency** and **growth** in CORE instructional model components (as measured by 2gno.me). This online assessment has seven components, and teachers take part in a pretest (prior to the start of year 1), a mid-test (end of year 1), and a posttest (end of year 2). Teachers receive a proficiency rating on each component on a five-point scale: *no experience, emerging, partially proficient, proficient,* or *advanced.* For teacher-level fidelity, a teacher must, at minimum, demonstrate proficiency and growth from pretest to at least one of the later timepoints (mid-test and/or posttest). In other words, over time, there must be evidence of movement from one proficiency level to the next on at least one component, and a rating of proficient or higher on at least one component. The various scenarios listed below would all result in meeting high-fidelity requirements at the teacher level.

- **Growth and proficiency met pretest to mid-test.** Note that if posttest data are available and fidelity requirements were not met at that time, this teacher would still get a score of high fidelity based on their previous scores (once fidelity criteria are met, the result is considered final).
- **Growth and proficiency met pretest to posttest.** This scenario covers the two-year timeframe of the study and may provide the best representation of change in teacher scores during the study, for those who have these data points available.
- **Growth and proficiency met mid-test to posttest.** Some teachers who may have joined their CORE teams later during the study may not have pretest data available from prior to year 1. For these teachers, the mid-test timepoint effectively serves as their "pretest" score.

Then, for school-level fidelity, two-thirds of teachers on the CORE team need to have attained high fidelity at the individual level as described above.

The seven components included in 2gno.me analysis are learner, leader, designer, collaborator, citizen, facilitator, and analyst. In the first year of the program, CORE teachers were assigned

modules within these components to complete. In the second year, teachers were able to choose which modules they complete, based on professional interest or preference. The differences in how this was implemented may have impacted results, but it is difficult to determine a clear way in which results would have been impacted. For example, some teachers may choose to complete modules that they feel comfortable with and enjoy in the second year. If they are already proficient in these areas, evidence of growth may not be captured. However, others may choose to focus on areas they find challenging— in these situations perhaps growth is more likely to be captured than proficiency. All teachers who were members of CORE teams completed the requirement for fall 2018 orientation, attaining high fidelity on Indicator 4.1.[9] Complete data for Indicator 4.2 (i.e., at least two timepoints) were available from 85 treatment teachers across both years of the study, 61 of whom met growth and proficiency requirements for high implementation on this indicator (72%). At the school level, ten schools achieved high implementation, meaning two-thirds of teachers at these schools met the growth and proficiency requirements on 2gno.me as defined above, falling short of the 75% required for program-level fidelity (see Exhibit 38).

It is important to note that final fidelity scores on Indicator 4.2 are based on available results from up to three timepoints, and findings above are based on all available data taken together as a whole. Refer to Section 3.3 below for a closer look at 2gno.me results at the individual level at each timepoint.

Exhibit 38. Key Component 4: Fidelity by School

				Fidelity Status by School
Key Component	**Indicators**	**RUP**	**School**	**Overall**
KC4. School teams participate in CORE instructional professional development services.	4.1 CORE team attendance at orientation on professional learning 4.2 Use of CORE instructional model as evidenced by scoring proficient or advanced on a minimum of one instructional component per year	FSU	School #28	High
			School #32	High
			School #40	High
		JSU	School #1	Low
			School #4	High
			School #5	Low
			School #8	High
			School #11	High
			School #13	High
		LTU	School #27	Low
		TSU	School #41	High
			School #48	High
		WTAMU	School #60	High
			School #66	Low

[9] A few teachers who took part in orientation and left their schools partway through the year are not included in findings due to missing data.

KC5: *School teams present during professional development workshops* was met at the end of the two-year study.

To achieve high fidelity of implementation on this KC, each CORE team must present on their learning during follow-up workshop sessions with JSU staff and other school teams. This component is measured once during the two years of the study. Schools presented on a staggered schedule as the study progressed; all schools had completed their presentations by the program's end (see Exhibit 39).

Exhibit 39. Key Component 5: Fidelity by School

				Fidelity Status by School
Key Component	Indicators	RUP	School	Overall
KC5. School teams present during professional development workshops.	5.1 Sharing of learning experiences as evidenced by CORE team presentations on learning outcomes during follow-up workshops	FSU	School #28	High
			School #32	High
			School #40	High
		JSU	School #1	High
			School #4	High
			School #5	High
			School #8	High
			School #11	High
			School #13	High
		LTU	School #27	High
		TSU	School #41	High
			School #48	High
		WTAMU	School #60	High
			School #66	High

KC6: *Schools participate in change-management support through CORE partnership resources* was not met at the end of the study.

KC6 measures participating schools' access of change-management support provided by the CORE program. Although this KC does not have separate indicators, each school must meet three requirements to achieve high fidelity of implementation at the school level. Principals must (1) administer a school level change-management survey, (2) receive a report of the results, and (3) participate in a debriefing session about survey results once during the two years of the study.

By the end of year 1, all principals had administered the change-management survey, and half had participated in the follow-up debrief. Debriefs were not completed with the other seven schools in year 2, which resulted in this KC falling short of meeting fidelity (Exhibit 40).

Initial plans to resend results reports and complete remaining debriefs in fall 2019 were delayed at the start of the academic year. In spring 2020, school closures related to the pandemic affected the completion of this task.

Exhibit 40. Key Component 6: Fidelity by School

				Fidelity Status by School
Key Component	Indicators	RUP	School	Overall
KC6. Schools participate in change-management support through CORE partnership resources.	6.1 Participation in change management	FSU	School #28	High
			School #32	High
			School #40	Low
		JSU	School #1	Low
			School #4	Low
			School #5	Low
			School #8	High
			School #11	High
			School #13	High
		LTU	School #27	Low
		TSU	School #41	High
			School #48	Low
		WTAMU	School #60	Low
			School #66	High

KC7: *School teams provide students with college-readiness advisement and support through use of EdReady™ tool in CORE schools* was met in year 1 and was not met in year 2 of the study.

KC7 pertains specifically to math and ELA teachers on CORE teams. All teachers in these subject areas who participate on CORE teams are expected to provide their students with college-readiness support by conducting math and English assessments using the EdReady™ tool on an annual basis.

In year 1, all math and English teachers on CORE teams at 12 of the 14 participating schools had completed these requirements. In year 2, teachers at 3 of 14 schools used EdReady™ with their students (see Exhibit 41).

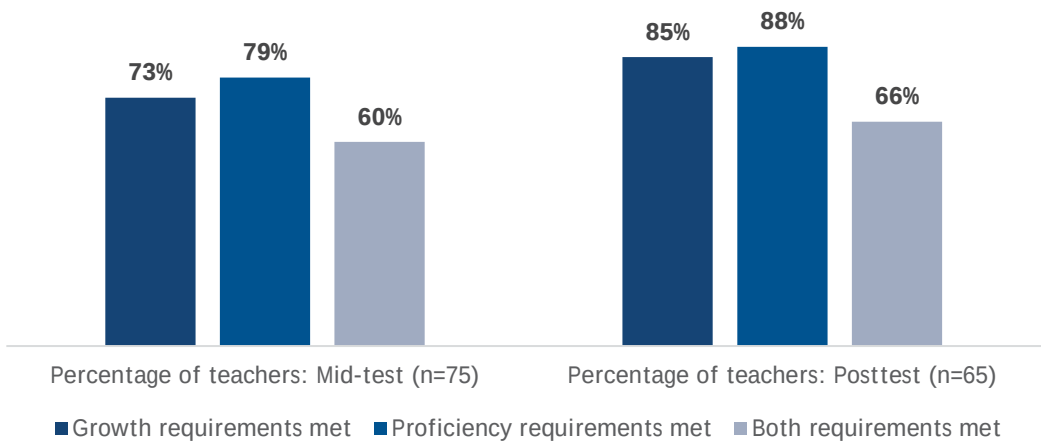
Exhibit 41. Key Component 7: Fidelity by School

				Fidelity Status by School	
Key Component	Indicators	RUP	School	Year 1	Year 2
KC7. School teams provide students with college readiness advisement and support through use of EdReady™ tool in CORE schools.	7.1 School teams' utilization of EdReady™ college readiness assessment tool	FSU	School #28	High	Low
			School #32	High	Low
			School #40	High	High
		JSU	School #1	High	Low
			School #4	High	High
			School #5	High	Low
			School #8	High	Low
			School #11	High	Low
			School #13	High	High
		LTU	School #27	Low	Low
		TSU	School #41	Low	Low
			School #48	High	Low
		WTAMU	School #60	High	Low
			School #66	High	Low

3.3 2gno.me

In this section, results for Indicator 4.2, use of CORE instructional model as evidenced by growth and proficiency on 2gno.me, is explored more in-depth. The figure below illustrates the percentage of teachers meeting growth requirements, proficiency requirements, and both requirements at mid-test (end of year 1) and posttest (end of year 2), based on the number of teachers who had available data at each timepoint.

Exhibit 42. 2gno.me Requirements Met at Mid-test and Posttest (teacher-level)

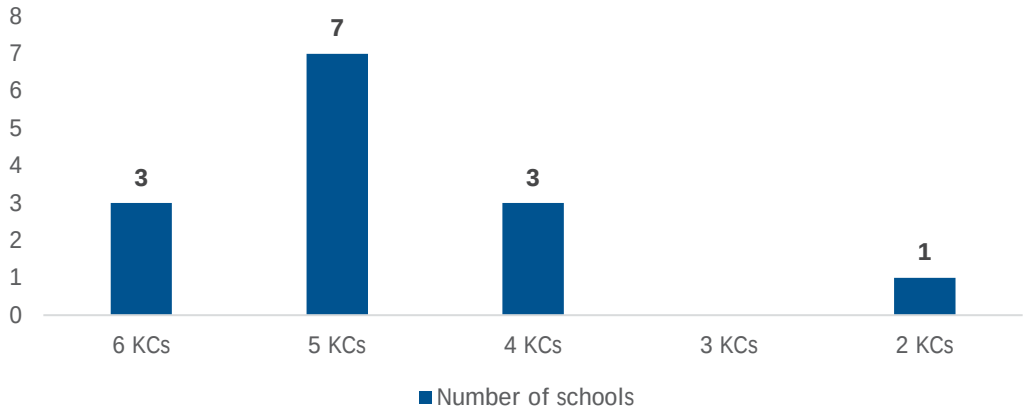


Teacher-level results show that a greater proportion of CORE teachers met requirements by the time of posttest. At both timepoints (mid-test and posttest), a higher percentage of teachers met proficiency requirements for at least one of the 2gno.me components compared to growth requirements. When rolling up individual-level results to the school level, we find that six schools met fidelity requirements by mid-test (i.e., two-thirds of teachers on CORE teams at these schools met individual-level fidelity), and ten schools met fidelity requirements by posttest.

3.4 Implementation Study Summary

Some schools performed well across the board, as far as fidelity of implementation, which may have implications for impact study results where available. It is important to recognize that some KCs may be more relevant for demonstrating the level at which a particular school was engaged with the CORE i3 program. Specifically, KC3 and KC5 were met with high fidelity by all schools. These KCs involved one-time completion of specific activities within the two-year program cycle (in other words, once completed, fidelity was considered met for that school for the entire span of the project). Other KCs (e.g., KC2, KC4) require sustained engagement and involvement throughout the project. KC1 is somewhat unique in that fidelity is based on administrator activities, which can sometimes be linked with teacher perceptions and engagement. Exhibit 43 below shows how many KCs were met by how many schools. No schools met all KCs, or no KCs. Half of schools met five KCs.

Exhibit 43: Number of KCs Met by Schools



Linkages between different KCs are important to acknowledge as well, as they might provide insight on how a particular school implemented CORE i3, and where their implementation strengths are. Findings of interest are included below.

- **Fidelity on KC1, which serves as an indicator of administrator engagement was correlated with teacher engagement (as defined in KC2).** Specifically, all but one of the seven schools that had high fidelity on KC1 also had high fidelity on KC2.
- **Teacher engagement (KC2) was closely linked to teacher growth and proficiency as measured by professional development offerings (KC4).** All schools that attained high fidelity on KC4 also had high fidelity on KC2, except for one school. Conversely, only three of 13 schools that had high fidelity on KC2 did not meet KC4 fidelity requirements.
- **All but one of the schools that met high fidelity for KC6 (participation in change management) also had high fidelity on KC4.** KC6 is more relevant as a school-level

or administrator indicator, whereas KC4 is teacher-based. The connection between these two KCs, and by extension, KC2 to some degree, may reflect a general level of high fidelity across KCs in a subset of schools.

Schools that had high site visit observation scores also had high fidelity on KC4. While the observation score may be linked in some cases to the teacher's level of engagement in the CORE program, interviews with the observed teachers revealed that they were often self-motivated to implement innovative teaching practices and continuously improve. In many cases, these teachers also described additional resources they had access to, provided by their school and/or district. These external factors may have contributed more to high growth and proficiency than the CORE model.

3.4.1 Overview of Fidelity by School

Fidelity of implementation results are described above by KC and indicator. Below in Exhibit 44 is an overview of fidelity for each treatment school. Ultimately, most schools met at least 5 KCs. Fidelity of implementation was weakest on KC1, KC6, and KC7. There were no clear patterns of higher fidelity for schools affiliated with a particular RUP by the end of the two-year study.

Exhibit 44. Fidelity on Key Components by School

RUP	School	KC1	KC2	KC3	KC4	KC5	KC6	KC7
JSU	School #1	Met	Met	Met	Not met	Met	Not met	Not met
	School #4	Not met	Met	Met	Met	Met	Not met	Met
	School #5	Met	Met	Met	Not met	Met	Not met	Not met
	School #8	Met	Met	Met	Met	Met	Met	Not met
	School #11	Not met	Met	Met	Met	Met	Met	Not met
	School #13	Not met	Met	Met	Met	Met	Met	Met
LTU	School #27	Not met	Not met	Met	Not met	Met	Not met	Not met
FSU	School #28	Not met	Met	Met	Met	Met	Met	Not met
	School #32	Not met	Met	Met	Met	Met	Met	Not met
	School #40	Met	Met	Met	Met	Met	Not met	Met
TSU	School #41	Met	Not met	Met	Met	Met	Met	Not met
	School #48	Met	Met	Met	Met	Met	Not met	Not met
WTAMU	School #60	Not met	Met	Met	Met	Met	Not met	Not met
	School #66	Met	Met	Met	Not met	Met	Met	Not met

Note: For KCs that are measured annually, year 2 fidelity status is reflected here.

IV. Discussion

The culmination of the impact and implementation studies of the CORE i3 program leads to some key takeaways, which are described below. Unfortunately, circumstances related to COVID-19 prevented complete data collection for the impact study, necessitating a shift in study design and analysis plans. Arguably, COVID-19 likely affected the strength of program implementation as well, however the repercussions for fidelity are less certain. For example, teachers remained strongly engaged in posting online reflections (an indicator associated with KC2), which may have been easier to do in a virtual learning environment. This KC was not met in year 1, although it was approaching fidelity, and was successfully met in year 2. It is not clear if the aspects of implementation that fell short in year 2 would have been met if COVID-19 was not an issue. Specifically, KC6 (change management) and KC7 (use of EdReady™) did not achieve fidelity, but other contributing factors besides COVID-19 may have influenced these results (e.g., administrator turnover, teacher perceptions about the utility of EdReady™ if they had already implemented it in year 1).

Thus, the primary limitation associated with COVID-19 is the substantial effect on the impact study and the difficulties of drawing strong conclusions about program impact and potential linkages with strength of implementation and program outcomes. Some findings were interesting and might warrant continued exploration in any future iteration of the program. Despite the data challenges caused by COVID-19, the two-year program effect sizes on student outcomes were close to the magnitude considered substantially important in the educational evaluation field. ICF believes a replication of this study is justified to further assess the CORE program's efficacy with particular attention to differential program effects based on students' race and gender. The exploratory findings also pointed to the importance of program exposure for both teachers and students. One important implementation component related to program impact on students was teachers' participation in professional development activities. Furthermore, program impact on students seemed related to students' exposure to CORE teachers. Effective implementation of the CORE program may rest on the program's ability to meaningfully expose a larger number of teachers and their students to program activities and resources.

Another important consideration for CORE program effectiveness is the necessity of buy-in from program participants at the RUP, leadership, and teacher levels. Strong implementation of CORE requires investment and engagement across multiple stakeholders for consistent understanding of program goals and active participation to bring these goals to fruition. Several aspects of CORE program implementation involve sustained participation, engagement, and growth over time. Circumstances common in schools, and particularly in schools served by this project—such as principal and teacher turnover—can have a negative impact on these types of longer-term study activities.[10] For example, active participation in change management (as measured by KC6) can be challenging when there is leadership turnover and competing priorities.

[10] Teacher turnover in this case does not necessarily mean teachers leaving their schools, but rather leaving the CORE team while remaining as a teacher at the school. On average, nearly a quarter of teachers left and were replaced on CORE teams during the two years of the study (see Appendix B).

Although the official study period is concluding, it is important to reflect on lessons learned to sustain progress made through the program and maximize the impact of resources after conclusion of the grant. Unlike the previous CORE model, this modified version allowed for whole school utilization and access to an online platform with useful tools and best practices to customize CORE teams' professional development experiences. However, this approach limited access to in-person support and training, a recognized benefit of the previous CORE model. In the initial proposal, JSU and RUPs were responsible for facilitating an in-person or virtual CORE Academy. But, due to capacity constraints, RUPs did not implement the CORE Academy in their respective regions. Overall, this integral component would have been useful in grounding teachers and administrators upfront in the CORE model and increasing overall understanding and buy-in of the model within schools. ICF recommends debriefing with RUP leaders, school administrators, and CORE teachers to gain a deeper understanding of their experiences in the program from their unique vantage points. Consider if a concluding discussion panel or other opportunity for sharing might benefit program participants as far as sharing successes, challenges, and lessons learned. CORE teams may have ideas for extending content from their professional development workshop presentations to ensure the knowledge they gained lives on in some way through CORE programming. If group reflection is not feasible, the JSU team may consider individual debriefs with specific RUPs or school leaders, to get their perspectives on what worked well as far as program implementation and communication, and where there may be areas for improvement for any future similar initiatives.

Ultimately, the CORE program seeks to facilitate significant and sustained change in teacher instructional strategies at participating schools to improve student outcomes. However, it is important to understand that the process of change is complex and may be difficult to pin down into something measurable. It may be helpful to consider more concrete questions related to implementation. What components need to be in place for a school to take ownership of the change process? Who is responsible for ensuring change takes place, and what is their role at the school? What kinds of supports should be in place to minimize negative impacts of turnover or other extenuating circumstances that may arise?

These are all questions that warrant further exploration through stakeholder feedback as the CORE program evolves and is implemented in other contexts.

V. References

Blanchard, M. R., LePrevost, C. E., Tolin, A. D., & Gutierrez, K. S. (2016). Investigating technology-enhanced teacher professional development in rural, high-poverty middle schools. *Educational Researcher, 45*(3), 2016.

Byun, S., Meece, J. L., & Irvin, M. J. (2012). Rural–nonrural disparities in postsecondary educational attainment revisited. *American Educational Research Journal, 49*(3), 412–437. http://eric.ed.gov/?id=EJ968047

Consortium on Chicago School Research at the University of Chicago. (2007). *Sample high school detailed report.* Retrieved September 19, 2015 from http://ccsr.uchicago.edu

Howley, C. & Hambrick, K. (2011). *Attachment and aspiration: What influences rural youths' educational and residential plans?* White Paper. Retrieved October 10, 2018 from http://eric.ed.gov/ ERIC Number: ED532584

ICF Study Team (2018). 2013 JSU CORE Impact and Implementation Study Final Report submitted to Department of Education under contract number ED-IES-10-C-0064. Retrieved October 10, 2018 from http://www.icf.com/-/media/files/icf/reports/2020/CORE-i3-final-report

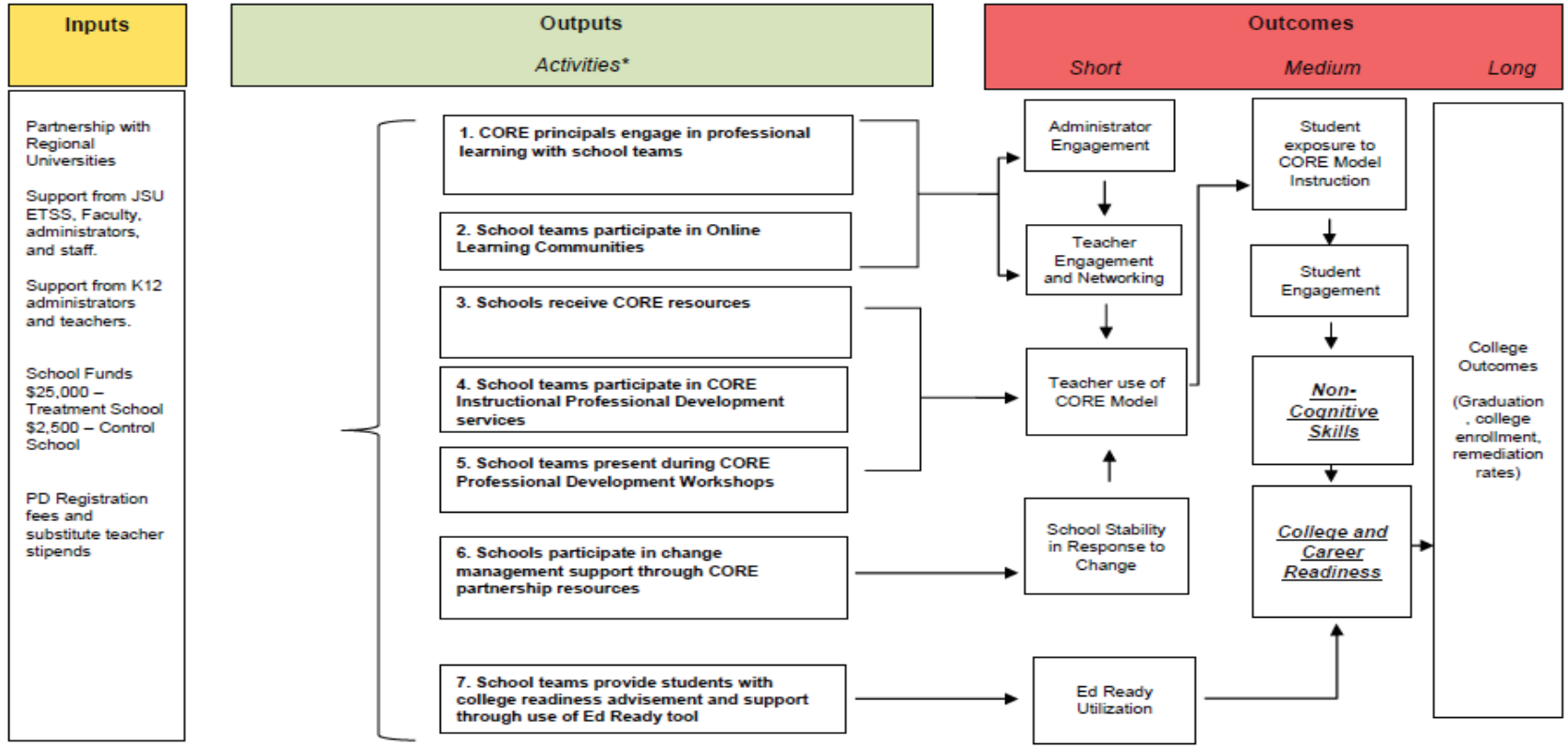
Midgley, C., Maehr, M. L., Hruda, L., Anderman, E. M., Freeman, K. E., Gheen, M., & Urdan, T. (2000). *Manual for the patterns of adaptive learning scales (PALS).* Ann Arbor, MI: University of Michigan.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models* (Second Edition). Thousand Oaks: Sage Publications.

What Works Clearinghouse. (2013). *What Works Clearinghouse: Procedures and standards handbook* (Version 3.0). Retrieved September 1, 2018 from http://whatworks.ed.gov

P21.org Partnership for 21st Century Skills. Web. 16 Sept 2018. http://www.p21.org/about-us/our-mission also retrieved 10 Oct 2020 from https://files.eric.ed.gov/fulltext/ED519462.pdf

Appendix A: Logic Model



Appendix B: Data Management

1. Impact Study

In alignment with the overall research design employed in this study, data management was organized into three distinct levels of a naturally occurring hierarchy: schools, teachers and teams within those schools, and the students attending those schools who are instructed by participating teachers. The participation status of schools randomly assigned to treatment and control conditions was tracked over the course of the project. The status of participating teachers and team members, as well as teacher-level outcomes, was also tracked over time, ensuring that data from students instructed by participating teachers were analyzed according to students' level of exposure to the intervention. Finally, student enrollment in participating schools was tracked annually to assess levels of attrition and adequate representation of school-level impact estimates. All data elements were organized and stored inside a relational database to facilitate access by evaluation team members and to automate the calculation of fidelity of implementation metrics.

As outlined earlier in this report, 14 schools were randomly assigned to the intervention group and 14 schools were randomly assigned to participate as comparison schools. These schools all remained in the study for the life of the project. Despite the continued participation of schools, the project did experience "attrition" in the form of missing data, the result of COVID-19 closures and access issues; some schools were unable to collect outcome data at various timepoints.

Teacher mobility and desire to participate in CORE i3 required longitudinal tracking of participating school team members (which often included Math, ELA, Science, Social Studies, Fine Arts/Career Tech/Foreign Language teachers, an administrator and a school point-of-contact). This required a two-step process: (1) obtain updated team member participation and orientation status from RUPs and (2) generate student-teacher rosters for review by schools, allowing schools to correct student-teacher assignment allocations and participating teachers. Because the research design is longitudinal, teachers that left a school, chose to no longer participate, or were assigned to a new role within a school were tagged as "leavers," along with a leave date. New team members were tagged as "arrivers" along with an arrival date, so that participation records could be tracked backward if necessary.[11]

Exhibit AB1 provides a summary of the team member churn that occurred over the life of the project. Among control schools, the total number of team members (including teachers, administrators, and school points-of-contact) was 104 in both school years. Seven team members left during the 2018-19 school year, yielding about a 7% leaver rate, and three more left in 2019-20 (3%). That same year (2019-20), 10 new team members arrived (10%). Effectively, across both years, 10 team member positions experienced churn within control schools. For treatment schools, 16 of the 110 team members left and two arrived during 2018-

[11] We chose to use "arrivers" and "leavers" language to avoid "attriters" and "joiners" nomenclature typically used in the WWC review process. Complete accurate tracking of dates was not always possible, as there was often a lag between when a team member may have actually arrived and when the evaluation team was informed of the new team member.

19 (15% and 2%, respectively). In 2019-20, nine (8%) and 24 (22%) of the 107 2019-20 team members left and arrived, respectively. Effectively, 25 team positions (approximately 23% of the average number of total positions) experienced churn.

Exhibit AB1. Summary of Team Member Churn: By Condition and School Year

Condition	School Year	Team Members	Leavers n	Arrivers n	Leavers %	Arrivers %
Control	2018-2019	104	7	0	7%	0%
	2019-2020	104	3	10	3%	10%
Treatment	2018-2019	110	16	2	15%	2%
	2019-2020	107	9	24	8%	22%

Student mobility was also an issue in this study. As alluded to in the teacher section above, student-teacher association rosters were generated on an annual basis for review by school-based staff. Staff were asked to confirm students were enrolled in the school, along with their current grade level, and if they were being instructed by school team members. If a student was no longer at the school, school staff were asked to supply a reason (e.g., transferred to another school, dropped out, moved away). Student "joiners" were not tracked for this project; only students attending the school at time of random assignment who were still enrolled at the school each year were included in the analyses. There was a total of 4,305 student records available for the 2018-19 school year and 270 students were tagged for removal from analyses (6.2%), leaving 4,035 students to track forward.[12] None of the 4,035 students were tagged for removal from analysis based on rosters collected during the 2019-20 school year.[13] Within control schools, 518 students did not have corresponding records returned during the 2019-20 roster process (23%). That same year, 213 students in treatment schools did not have corresponding records returned (12%).

Exhibit AB2. Summary of Student Records Lost from 2018-19 to 2019-20

Condition	Leaver n	Student n	Leaver %
Control	518	2,270	23%
Treatment	213	1,765	12%

2. Implementation Study

The components and methods of implementation fidelity are described in more detail in the Implementation Study section. The data-management process is described here. Fidelity of

[12] 265 of the 270 either withdrew from the school or transferred, 2 dropped out, and 3 enrolled in virtual/home school options.
[13] Only 133 supplied any reason at all, such as EC (exceptional child) status/IEP possession or failing a grade, which were not valid reasons for removal from analysis.

implementation data- element tracking was managed through Google Sheets tracking templates populated by JSU for the 2018-19 and 2019-20 school years. Some elements were representations at the school level (i.e., Administrator Survey results for KC1A or administrator participation in learning modules for KC1B) and others at the teacher level (i.e., teacher reflections for online learning modules for KC2). Elements of fidelity were tracked and aggregated to calculate school-level FOI scores. All elements were entered into the relational database where syntactical programming was used to summarize across teachers and schools to provide an overall program-level FOI score for each year, as well as individual FOI scores for each school and school year.

School-level FOI Elements:

- KC1A: School Collaboration Survey
- KC1B: Administrator Learning Module Participation
- KC3A: Monetary Disbursement
- KC3B: School Technology Assessment Completion
- KC3C: School Funds Use Survey Completion
- KC4A: Professional Orientation Completion[14]
- KC5: Workshop Presentation
- KC6: Change-management Survey Completion

Teacher-level FOI Elements:

- KC2: Teacher reflections for online learning modules
- KC4B: Teacher participation/results in 2gno.me assessment
- KC7: EdReady Diagnostic Completion (ELA and/or Math teachers)

[14] KC4A was considered school-level because each school received credit for FOI only if every team member participated in orientation.

Appendix C: Impact Study Section Tables

Exhibit AC1. Student Scales: Non-cognitive Scale, Engagement & Self-Efficacy Items in CWRA+

Question ID	Survey Items
Non-cognitive Skills Student Score	
Q1	I can prioritize my work to ensure I am completing tasks in a timely manner.
Q2	I am confident I will complete any task assigned.
Q3	I see more than one correct answer to many questions.
Q4	I search for solutions instead of adding to the problem.
Q5	I am confident during social interactions in the classroom.
Q6	I am proud to be a part of my school and community.
Q7	I strive to complete each assignment in a timely manner.
Q8	I enjoy working with others in the classroom.
Q9	I process information I receive before thoughtfully responding.
Q10	I set small goals to ensure I meet the overall objective.
Student Engagement Scale	
Q1	I usually look forward to this class.
Q2	I work hard to do my best in this class.
Q3	Sometimes I get so interested in my work in this class that I do not want to stop.
Q4	The topics we are studying in this class are interesting and challenging.
Student Efficacy Scale	
Q1	I'm certain I can master the skills taught in this class this year.
Q2	I'm certain I can figure out how to do the most difficult work in this class this year.
Q3	I can do almost all the work in this class if I don't give up.
Q4	Even if the work is hard in this class, I can learn it.
Q5	I can do even the hardest work in this class if I try.

Exhibit AC2. Pretest-Mid-test Subgroup Program Impact Estimates for Different Student Subgroups

	N. of Schools		N. of Students					
Subgroups	**T**	**C**	**T**	**C**	**Estimate**	**sig**	**Standardized**	**model**
CWRA+								
Minority female	14	13	260	305	18.92	ns	0.11	HLM
White female	12	13	278	303	35.26	ns	0.21	HLM
Minority male	14	13	266	299	18.21	ns	0.11	HLM
White male	11	13	255	243	25.41	ns	0.14	HLM
Non-cognitive Scale								
Minority female	14	12	224	268	0.00	ns	0.00	HLM
White female	12	12	244	268	0.07	ns	0.12	HLM
Minority male	14	12	221	260	-0.04	ns	-0.07	OLS
White male	11	12	223	207	0.02	ns	0.04	HLM
Engagement Scale								
Minority female	14	12	224	268	0.02	ns	0.03	HLM
White female	12	12	244	268	0.08	ns	0.10	HLM
Minority male	14	12	221	260	-0.10	ns	-0.13	HLM
White male	11	12	223	207	0.04	ns	0.05	HLM
Efficacy Scale								
Minority female	14	12	224	268	-0.02	ns	-0.03	HLM
White female	12	12	244	268	0.13	*	0.18	OLS
Minority male	14	12	221	260	0.01	ns	0.01	HLM
White male	11	12	223	207	0.01	ns	0.01	HLM

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. Statistical model column indicates whether the model was an HLM (Hierarchical Linear Modeling) or an OLS (Ordinary Least Square) model.

Exhibit AC3. RUP-Specific Program Impacts on Student Outcomes

	N of Students			N of Schools							
Study phase	**Whole**	**Treatment**	**Control**	**Whole**	**Treatment**	**Control**	**Program impact**	**sig**	**Standardized**	**model**	**Baseline Test**
(14) Fayetteville State University											
Pretest-Mid-test	621	186	435	6	3	3	35.43	ns	0.18	HLM	C
Pretest-Posttest	235	93	142	2	1	1	108.17	***	0.56	OLS	C
(15) Jacksonville State University											
Pretest-Mid-test	1070	667	403	12	6	6	11.32	ns	0.07	HLM	C
Pretest-Posttest	115	31	84	4	1	3	9.61	ns	0.07	HLM	B
(16) Louisiana Tech University											
Pretest-Mid-test	58	26	32	2	1	1	-82.43	ns	-0.53	OLS	C
Pretest-Posttest	38	27	11	2	1	1	-8.72	ns	-0.06	OLS	C
(17) Tarleton State University											
Pretest-Mid-test	172	69	103	4	2	2	95.40	ns	0.54	HLM	B
Pretest-Posttest	65	65	n/a	2	2	n/a	n/a		n/a	HLM	n/a
(18) West Texas A&M University											
Pretest-Mid-test	371	157	214	3	2	1	8.94	ns	0.06	HLM	C
Pretest-Posttest	21	21	n/a	1	1	n/a	n/a		n/a	OLS	n/a

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. Statistical model column indicates whether the model was an HLM (Hierarchical Linear Modeling) or an OLS (Ordinary Least Square) model. The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Appendix D: Impact Study Main Findings Tables

Exhibit AD1. HLM Results for the Assessment of CORE Program Impact on CWRA+ Scores

CWRA+	Pretest-Mid-test Model (n. of students=2,292; n of schools=27)				Pretest-Posttest Model (n. of students=474; n of schools=11)			
	Coeff.	SE	p	Sig.	Coeff.	SE	P	Sig.
Intercept	478.29	23.79	<.0001	***	392.81	49.55	<.0001	***
Treatment (vs. C)	20.94	20.11	0.31	ns	38.82	32.01	0.25	ns
Pretest CWRA+	0.46	0.02	<.0001	***	0.48	0.04	<.0001	***
Grade 10 (vs. 9)	3.06	6.03	0.61	ns	-32.67	13.25	0.01	*
Male	-6.35	6.04	0.29	ns	13.34	13.79	0.33	ns
Gender info missing	-17.13	20.73	0.41	ns	101.25	54.65	0.06	ns
Other race group	-23.80	10.57	0.02	*	-3.13	24.64	0.90	ns
Black	-49.60	9.68	<.0001	***	-29.67	19.87	0.14	ns
Hispanic	-24.56	10.04	0.01	*	-29.54	27.88	0.29	ns
Race info missing	-4.44	23.70	0.85	ns	-50.12	62.05	0.42	ns
Parent w/ BA	15.39	6.73	0.02	*	13.53	14.05	0.34	ns
Parent info missing	-15.01	17.15	0.38	ns	-15.41	41.97	0.71	ns

Notes: Statistical significance (2-tail test): ns = not significant, * = p<.05, ** = p<.01, *** = p<.001. The pretest-mid-test model variance information: Anova model level 1 variance 25,733, level 2 variance 5,256, ICC 0.17; Conditional model level 1 variance 19,667, level 2 variance 2,351, ICC 0.11; Level 1 variance explained 0.24, Level 2 variance explained 0.55. The pretest-posttest model variance information: Anova model level 1 variance 26,565, level 2 variance 3,393, ICC 0.11; Conditional model level 1 variance 19,525, level 2 variance 2,030, ICC 0.11; Level 1 variance explained 0.27, Level 2 variance explained 0.40.

Exhibit AD2. Descriptive Statistics Table: Pretest-Mid-test HLM Results for the Assessment of CORE Program Impact on CWRA+ Scores

	Whole Sample (n=2,292)		Treatment (n=1,105)		Comparison (n=1,187)				
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Mid-test CWRA+ score	913.28	176.42	920.04	179.64	906.99	173.22	13.05	0.07	B
Pretest CWRA+ score	925.31	173.34	917.90	170.88	932.21	175.40	-14.31	0.08	B
Treatment (vs. Control)	0.48	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.48	0.50	0.44	0.50	0.52	0.50	-0.08	0.16	B
Male	0.47	0.50	0.48	0.50	0.46	0.50	0.02	0.04	A
Gender info missing	0.03	0.16	0.03	0.16	0.03	0.16	0.00	0.01	A
Other race group	0.10	0.30	0.10	0.30	0.10	0.30	0.00	0.00	A
Black	0.20	0.40	0.21	0.41	0.19	0.39	0.03	0.06	B
Hispanic	0.19	0.40	0.16	0.37	0.22	0.42	-0.06	0.15	B
Race info missing	0.02	0.14	0.02	0.14	0.02	0.13	0.00	0.02	A
Parent w/ BA	0.31	0.46	0.29	0.45	0.33	0.47	-0.04	0.09	B
Parent info missing	0.03	0.18	0.04	0.19	0.03	0.17	0.01	0.04	A

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD3. Descriptive Statistics Table: Pretest-Posttest HLM Results for the Assessment of CORE Program Impact on CWRA+ Scores

	Whole Sample (n=474)		Treatment (n=237)		Comparison (n=237)				
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Posttest CWRA+ score	875.95	174.07	903.04	180.89	848.86	162.90	54.17	0.31	C
Pretest CWRA+ score	967.08	173.53	968.46	173.74	965.70	173.67	2.76	0.02	A
Treatment (vs. Control)	0.50	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.46	0.50	0.50	0.50	0.42	0.50	0.08	0.16	B
Male	0.40	0.49	0.40	0.49	0.40	0.49	0.00	0.00	A
Gender info missing	0.02	0.14	0.02	0.13	0.03	0.16	-0.01	0.06	B
Other race group	0.11	0.31	0.11	0.31	0.11	0.31	0.00	0.01	A
Black	0.39	0.49	0.34	0.48	0.44	0.50	-0.10	0.20	B
Hispanic	0.09	0.28	0.15	0.36	0.03	0.16	0.13	0.46	C
Race info missing	0.02	0.13	0.02	0.14	0.01	0.11	0.01	0.07	B
Parent w/BA	0.38	0.49	0.37	0.48	0.38	0.49	-0.01	0.03	A
Parent info missing	0.03	0.16	0.03	0.18	0.02	0.13	0.02	0.11	B

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD4. HLM and OLS Results for the Assessment of CORE Program Impact on Non-cognitive Scores

Non-cognitive Scores	Pretest-Mid-test Model (n. of students=1,977; n of schools=26)				Pretest-Posttest Model (OLS) (n. of students=443; n of schools=11)			
	Coeff.	SE	p	Sig	Coeff.	SE	p	Sig.
Intercept	2.00	0.08	<.0001	***	1.79	0.22	<.0001	***
Treatment (vs. Control)	0.02	0.03	0.52	ns	0.13	0.05	0.01	*
Pretest Non-cognitive	0.47	0.02	<.0001	***	0.51	0.05	<.0001	***
Grade 10 (vs. 9)	-0.02	0.02	0.44	ns	0.01	0.05	0.90	ns
Male	-0.01	0.02	0.52	ns	-0.04	0.05	0.49	ns
Gender info missing	-0.14	0.09	0.11	ns	-0.25	0.21	0.24	ns
Other race group	-0.01	0.04	0.81	ns	0.04	0.09	0.70	ns
Black	0.08	0.03	0.02	*	0.04	0.06	0.51	ns
Hispanic	0.04	0.03	0.24	ns	0.02	0.10	0.85	ns
Race info missing	-0.02	0.09	0.86	ns	-0.02	0.24	0.92	ns
Parent w/BA	0.05	0.03	0.06	t	0.13	0.06	0.02	*
Parent info missing	-0.11	0.07	0.10	ns	0.05	0.16	0.77	ns

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. The pretest-mid-test model variance information: Anova model level 1 variance 0.32, level 2 variance 0.00, ICC 0.01; Conditional model level 1 variance 0.25, level 2 variance 0.00, ICC 0.01; Level 1 variance explained 0.22, Level 2 variance explained 0.44. The pretest-posttest model variance information: Anova model level 1 variance 0.37, level 2 variance 0.00, ICC 0.00; Conditional model level 1 variance 0.30, level 2 variance 0.00, ICC 0.00; Level 1 variance explained 0.19, Level 2 variance explained N/A. Due to the lack of between-cluster outcome variance, OLS regression model was used.

Exhibit AD5. Descriptive Statistics Table: Pretest-Mid-test HLM Results for the Assessment of CORE Program Impact on Non-cognitive Scores

	Whole Sample (n=1,977)		Treatment (n=946)		Comparison (n=1,031)				
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Mid-test non-cognitive scale	3.79	0.57	3.81	0.55	3.78	0.58	0.03	0.05	A
Pretest non-cognitive score	3.79	0.56	3.80	0.55	3.78	0.57	0.02	0.03	A
Treatment (vs. Control)	0.48	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.47	0.50	0.44	0.50	0.51	0.50	-0.07	0.13	B
Male	0.47	0.50	0.48	0.50	0.46	0.50	0.02	0.04	A
Gender info missing	0.02	0.14	0.02	0.14	0.02	0.14	0.00	0.01	A
Other race group	0.10	0.30	0.10	0.30	0.10	0.30	0.00	0.01	A
Black	0.20	0.40	0.20	0.40	0.20	0.40	0.00	0.01	A
Hispanic	0.20	0.40	0.17	0.38	0.22	0.42	-0.05	0.12	B
Race info missing	0.02	0.14	0.02	0.14	0.02	0.14	0.00	0.01	A
Parent w/BA	0.32	0.47	0.30	0.46	0.34	0.47	-0.04	0.09	B
Parent info missing	0.03	0.17	0.03	0.18	0.03	0.16	0.01	0.03	A

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD6. Descriptive Statistics Table: Pretest-Posttest OLS Results for the Assessment of CORE Program Impact on Non-cognitive Scores

	Whole Sample (n=443)		Treatment (n=225)		Comparison (n=218)				
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Posttest non-cognitive scale	3.90	0.61	3.95	0.59	3.84	0.63	0.11	0.18	B
Pretest non-cognitive score	3.90	0.49	3.87	0.48	3.92	0.51	-0.05	0.10	B
Treatment (vs. Control)	0.51	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.47	0.50	0.50	0.50	0.43	0.50	0.08	0.15	B
Male	0.39	0.49	0.40	0.49	0.39	0.49	0.01	0.01	A
Gender info missing	0.02	0.15	0.02	0.13	0.03	0.16	-0.01	0.07	B
Other race group	0.10	0.30	0.09	0.29	0.10	0.30	-0.01	0.03	A
Black	0.39	0.49	0.33	0.47	0.44	0.50	-0.12	0.24	B
Hispanic	0.09	0.29	0.16	0.37	0.03	0.16	0.13	0.46	C
Race info missing	0.02	0.13	0.02	0.15	0.01	0.12	0.01	0.06	B
Parent w/BA	0.38	0.49	0.38	0.49	0.39	0.49	0.00	0.01	A
Parent info missing	0.03	0.16	0.04	0.19	0.02	0.14	0.02	0.11	B

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD7. HLM Results for the Assessment of CORE Program Impact on Engagement Scores

	Pretest-Mid-test Model (n. of students=1977; n of schools=26)				Pretest-Posttest Model (n. of students=443; n of schools=11)			
	Coeff.	SE	p	Sig.	Coeff.	SE	p	Sig.
Intercept	1.93	0.09	<.0001	***	2.13	0.20	<.0001	***
Treatment (vs. Control)	-0.00	0.05	0.98	ns	0.18	0.11	0.16	ns
Pretest engagement	0.44	0.02	<.0001	***	0.39	0.05	<.0001	***
Grade 10 (vs. 9)	-0.02	0.03	0.49	ns	-0.01	0.07	0.91	ns
Male	-0.06	0.03	0.05	t	-0.11	0.07	0.14	ns
Gender info missing	-0.18	0.12	0.15	ns	-0.50	0.28	0.08	ns
Other race group	0.09	0.06	0.12	ns	0.11	0.13	0.41	ns
Black	0.17	0.05	0.00	***	0.02	0.10	0.83	ns
Hispanic	0.15	0.05	0.00	**	-0.08	0.14	0.57	ns
Race info missing	0.14	0.13	0.29	ns	0.06	0.32	0.86	ns
Parent w/BA	0.01	0.04	0.74	ns	0.15	0.07	0.05	*
Parent info missing	-0.07	0.09	0.47	ns	-0.09	0.21	0.68	ns

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. The pretest-mid-test model variance information: Anova model level 1 variance 0.59, level 2 variance 0.01, ICC 0.01; Conditional model level 1 variance 0.48, level 2 variance 0.01, ICC 0.02; Level 1 variance explained 0.19, Level 2 variance explained -0.20. The pretest-posttest model variance information: Anova model level 1 variance 0.59, level 2 variance 0.12, ICC 0.03; Conditional model level 1 variance 0.51, level 2 variance 0.01, ICC 0.03; Level 1 variance explained 0.14, Level 2 variance explained 0.33.

Exhibit AD8. Descriptive Statistics Table: Pretest-Mid-test HLM Results for the Assessment of CORE Program Impact on Engagement Scores

	Whole Sample (n=1,977)		Treatment (n=946)		Comparison (n=1,031)				
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Mid-test engagement scale	3.45	0.78	3.46	0.77	3.44	0.78	0.03	0.03	A
Pretest engagement score	3.45	0.76	3.46	0.76	3.44	0.77	0.02	0.03	A
Treatment (vs. Control)	0.48	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.47	0.50	0.44	0.50	0.51	0.50	-0.07	0.13	B
Male	0.47	0.50	0.48	0.50	0.46	0.50	0.02	0.04	A
Gender info missing	0.02	0.14	0.02	0.14	0.02	0.14	0.00	0.01	A
Other race group	0.10	0.30	0.10	0.30	0.10	0.30	0.00	0.01	A
Black	0.20	0.40	0.20	0.40	0.20	0.40	0.00	0.01	A
Hispanic	0.20	0.40	0.17	0.38	0.22	0.42	-0.05	0.12	B
Race info missing	0.02	0.14	0.02	0.14	0.02	0.14	0.00	0.01	A
Parent w/BA	0.32	0.47	0.30	0.46	0.34	0.47	-0.04	0.09	B
Parent info missing	0.03	0.17	0.03	0.18	0.03	0.16	0.01	0.03	A

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD9. Descriptive Statistics Table: Pretest-Posttest HLM Results for the Assessment of CORE Program Impact on Engagement Scores

	Whole Sample (n=443)		Treatment (n=225)		Comparison (n=218)				
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Posttest engagement scale	3.56	0.78	3.64	0.75	3.49	0.80	0.15	0.20	B
Pretest engagement score	3.54	0.73	3.51	0.75	3.57	0.71	-0.06	0.08	B
Treatment (vs. Control)	0.51	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.47	0.50	0.50	0.50	0.43	0.50	0.08	0.15	B
Male	0.39	0.49	0.40	0.49	0.39	0.49	0.01	0.01	A
Gender info missing	0.02	0.15	0.02	0.13	0.03	0.16	-0.01	0.07	B
Other race group	0.10	0.30	0.09	0.29	0.10	0.30	-0.01	0.03	A
Black	0.39	0.49	0.33	0.47	0.44	0.50	-0.12	0.24	B
Hispanic	0.09	0.29	0.16	0.37	0.03	0.16	0.13	0.46	C
Race info missing	0.02	0.13	0.02	0.15	0.01	0.12	0.01	0.06	B
Parent w/BA	0.38	0.49	0.38	0.49	0.39	0.49	0.00	0.01	A
Parent info missing	0.03	0.16	0.04	0.19	0.02	0.14	0.02	0.11	B

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD10. HLM and OLS Results for the Assessment of CORE Program Impact on Efficacy Scores

	Pretest-Mid-test Model (n. of students=1977; n of schools=26)				Pretest-Posttest Model (OLS) (n. of students=443; n of schools=11)			
	Coeff.	SE	p	Sig.	Coeff.	SE	p	Sig.
Intercept	2.30	0.09	<.0001	***	1.52	0.20	<.0001	***
Treatment (vs. C)	0.05	0.04	0.19	ns	0.24	0.06	0.00	***
Pretest efficacy	0.40	0.02	<.0001	***	0.55	0.05	<.0001	***
Grade 10 (vs. 9)	-0.02	0.03	0.54	ns	0.01	0.06	0.82	ns
Male	-0.01	0.03	0.69	ns	-0.08	0.06	0.24	ns
Gender info missing	-0.15	0.11	0.19	ns	-0.49	0.25	0.05	ns
Other race group	0.02	0.05	0.73	ns	-0.01	0.11	0.90	ns
Black	0.12	0.04	0.00	**	0.09	0.07	0.19	ns
Hispanic	0.05	0.04	0.27	ns	0.01	0.12	0.94	ns
Race info missing	-0.06	0.12	0.59	ns	0.46	0.28	0.10	ns
Parent w/BA	0.04	0.03	0.19	ns	0.23	0.07	0.00	***
Parent info missing	-0.11	0.09	0.21	ns	0.20	0.19	0.30	ns

Notes: Statistical significance (2-tail test): ns = not significant, * = $p<.05$, ** = $p<.01$, *** = $p<.001$. The pretest-mid-test model variance information: Anova model level 1 variance 0.49, level 2 variance 0.01, ICC 0.02; Conditional model level 1 variance 0.41, level 2 variance 0.00, ICC 0.00; Level 1 variance explained 0.17, Level 2 variance explained 0.79. The pretest-posttest model variance information: Anova model level 1 variance 0.55, level 2 variance 0.02, ICC 0.03; Conditional model level 1 variance 0.41, level 2 variance 0.00, ICC 0.00; Level 1 variance explained 0.26, Level 2 variance explained N/A. Due to the lack of between-cluster outcome variance, OLS regression model was used.

Exhibit AD11. Descriptive Statistics Table: Pretest-Mid-test HLM Results for the Assessment of CORE Program Impact on Efficacy Scores

	Whole Sample (n=1,977)		Treatment (n=946)		Comparison (n=1,031)		Summary		
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Mid-test efficacy scale	3.95	0.71	3.98	0.68	3.92	0.73	0.05	0.08	B
Pretest efficacy score	3.94	0.72	3.95	0.71	3.94	0.73	0.02	0.03	A
Treatment (vs. Control)	0.48	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.47	0.50	0.44	0.50	0.51	0.50	-0.07	0.13	B
Male	0.47	0.50	0.48	0.50	0.46	0.50	0.02	0.04	A
Gender info missing	0.02	0.14	0.02	0.14	0.02	0.14	0.00	0.01	A
Other race group	0.10	0.30	0.10	0.30	0.10	0.30	0.00	0.01	A
Black	0.20	0.40	0.20	0.40	0.20	0.40	0.00	0.01	A
Hispanic	0.20	0.40	0.17	0.38	0.22	0.42	-0.05	0.12	B
Race info missing	0.02	0.14	0.02	0.14	0.02	0.14	0.00	0.01	A
Parent w/BA	0.32	0.47	0.30	0.46	0.34	0.47	-0.04	0.09	B
Parent info missing	0.03	0.17	0.03	0.18	0.03	0.16	0.01	0.03	A

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD12. Descriptive Statistics Table: Pretest-Posttest OLS Results for the Assessment of CORE Program Impact on Efficacy Scores

	Whole Sample (n=443)		Treatment (n=225)		Comparison (n=218)				
	Mean	SD	Mean	SD	Mean	SD	Diff.	Hedge's g	Baseline test
Posttest efficacy scale	3.96	0.75	4.09	0.70	3.83	0.78	0.26	0.35	C
Pretest efficacy Score	4.06	0.64	4.07	0.57	4.04	0.69	0.03	0.04	A
Treatment (vs. Control)	0.51	0.50	1.00	0.00	0.00	0.00	1.00		N/A
10th grader	0.47	0.50	0.50	0.50	0.43	0.50	0.08	0.15	B
Male	0.39	0.49	0.40	0.49	0.39	0.49	0.01	0.01	A
Gender info missing	0.02	0.15	0.02	0.13	0.03	0.16	-0.01	0.07	B
Other race group	0.10	0.30	0.09	0.29	0.10	0.30	-0.01	0.03	A
Black	0.39	0.49	0.33	0.47	0.44	0.50	-0.12	0.24	B
Hispanic	0.09	0.29	0.16	0.37	0.03	0.16	0.13	0.46	C
Race info missing	0.02	0.13	0.02	0.15	0.01	0.12	0.01	0.06	B
Parent w/BA	0.38	0.49	0.38	0.49	0.39	0.49	0.00	0.01	A
Parent info missing	0.03	0.16	0.04	0.19	0.02	0.14	0.02	0.11	B

Notes: The Baseline test column indicates whether the sample satisfied the What Works Clearinghouse baseline equivalence (BA) requirement: A: Satisfied the requirement (standardized difference <= 0.05), B: Requires statistical adjustment to satisfy the BE requirement (standardized difference <= 0.25), C: Does not satisfy the BE requirement (standardized difference > 0.25).

Exhibit AD13. Comparison of the Two Analysis Samples: Pretest-Mid-test CWRA+ Sample vs. Pretest-Posttest CWRA+ Sample

	Pretest-Mid-test Sample (n=2,292)		Pretest-Posttest Sample (after PSM) (n=474)		
	Mean	SD	Mean	SD	Difference
Outcome CWRA+	913.28	176.42	875.95	174.07	37.33
Pretest CWRA+	925.31	173.34	967.08	173.74	-41.77
	Percentage		Percentage		
Male	47%		40%		7%
Gender info missing	3%		2%		1%
White students	49%		39%		10%
Black students	20%		39%		-19%
Hispanic students	19%		9%		10%
Other race groups	10%		11%		-1%
Race info missing	2%		2%		0%
Parent w/ BA degree	31%		38%		-7%
Parent info missing	3%		3%		0%

Exhibit AD14. Findings from the Attrition Analyses and Baseline Equivalence Analyses

Outcome	School-level Attrition	Student-level Attrition	Baseline equivalence	Propensity Score Matching	Methodological Status
Pretest-Mid-test Analyses					
CWRA+	Low	Low	N/A	No	RCT
Non-cognitive scale	High	Low	BE satisfied	No	QED
Engagement scale	High	Low	BE satisfied	No	QED
Efficacy scale	High	Low	BE satisfied	No	QED
Pretest-Posttest Analyses*					
CWRA+	High	High	BE not satisfied	Yes	QED
Non-cognitive scale	High	Low	BE satisfied	Yes	QED
Engagement scale	High	Low	BE satisfied	Yes	QED
Efficacy scale	High	Low	BE satisfied	Yes	QED

Notes: * The pretest-posttest data collection was affected by the COVID-19 pandemic.