Effectiveness of Internet-Based Reading Apprenticeship Improving Science Education (*iRAISE*)

A REPORT OF A RANDOMIZED EXPERIMENT IN MICHIGAN AND PENNSYLVANIA

12/16/2016

Andrew P. Jaciw

Adam M. Schellinger

Li Lin

Jenna Zacamy

Megan Toby

Empirical Education Inc.

ACKNOWLEDGEMENTS

We are grateful to the teachers and staff in participating schools in Michigan and Pennsylvania for their assistance and cooperation in conducting this research. The research was conducted under a subcontract with WestEd as part of their 2012 Investing in Innovation (i3) Development grant (Award Number U411C120094). As the independent evaluator, Empirical Education Inc. was provided with independence in reporting the results.

ABOUT EMPIRICAL EDUCATION INC.

Empirical Education Inc. is a Silicon Valley-based R&D firm that uses its deep knowledge of education and its expertise in analytics, research design, and engineering, to help K-12 schools improve their programs and processes. We work directly with state and local education agencies as well as with the U.S. Department of Education, education technology providers, foundations, and leading research organizations that study and serve schools.

Executive Summary

In 2012, WestEd received a "Development" grant from the U.S. Department of Education's Investing in Innovation (i3) competition to develop and implement Internet-based Reading Apprenticeship Improving Science Education (*iRAISE*). *iRAISE* was implemented in Michigan and Pennsylvania and was provided to over 100 teachers who served approximately 20,000 students during the grant period. This report presents findings from the randomized control trial of *iRAISE*, which took place during the 2014-15 school year and investigated the impact of the program on teacher and student outcomes.

Overview of the Intervention

iRAISE brings Strategic Literacy Institute (SLI)'s 65-hour biology-based, face-to-face literacy professional development (PD) to an online format, with the hope of cutting the cost of previous face-to-face training by half. *iRAISE* is a year-long learning community in which high school science teachers learn about, practice, and refine ways to improve their students' ability to engage in and understand a variety of scientific texts. *iRAISE* builds from the existing materials, protocols, and key design elements of face-to-face Reading Apprenticeship PD and leverages interactive, internet-based technologies to enhance teachers' learning. The course is divided between online synchronous sessions with facilitators and peers and personal, asynchronous work. The PD begins with a 5-day (approximately 20-hour) *iRAISE* Foundations training during the summer prior to classroom implementation. After the start of the school year, teachers participate in monthly follow-up meetings from September through May, allowing them to continuously implement their learning over the year. The follow-up meetings provide three hours of additional support per month in two different formats: whole-group meetings introducing new learning (Ignite sessions) and small-group meetings intended to produce discussion and collaboration (PLC sessions). Through the development grant funding, SLI aimed to create a flexible, accessible, and high-quality online professional learning platform, while preserving the interactive, engaging character of face-to-face Reading Apprenticeship PD.

Research Design

The i3 evaluation of *iRAISE*, conducted by Empirical Education Inc., employed a cluster randomized control trial in which 82 teachers were randomly assigned to receive the *iRAISE* PD (41 teachers) or continue with business-as-usual (41 teachers). This was an intent-to-treat design, with impact estimates generated by comparing student average outcomes for teachers randomly assigned to the *iRAISE* group with student average outcomes for teachers assigned to control group status, regardless of the level of participation in or implementation of *iRAISE* instructional approaches after random assignment.

This report presents key implementation and impact findings from the i3 impact evaluation of the *iRAISE* project. Data sources for this report include teacher surveys; PD observations and attendance records; school district student records; and an assessment of students' literacy skills.

Key Findings About RAISE Implementation

Implementation and contextual factors that may have facilitated or hindered implementation of *iRAISE* were measured through PD observations and attendance records, teacher surveys, and principal

surveys. The following data indicated that *iRAISE* PD and in-school support were delivered as intended.

- Over 90% of the observed PD sessions exhibited the five key design characteristics.
- 32 out of the 33 teachers who agreed to participate in the PD attended at least four days of the Foundations training, with 26 of those attending all five days. However, eight teachers randomized to the *iRAISE* group did not attend any of the PD: one teacher left his teaching position, two teachers declined to participate in the study shortly after randomization, and five teachers agreed to data collection but declined the PD because of other obligations.
- Over 80% of teachers (n = 27) who responded to survey questions about the *iRAISE* PD after attending felt that it "moderately", "more than moderately", or "completely" prepared them to use the set of literacy practices modeled during the training.

iRAISE teachers reported more support for literacy instruction than their control peers and generally held positive views of Reading Apprenticeship and its efficacy. Their survey responses indicated buy-in and commitment to implementing the framework.

- *iRAISE* teachers reported receiving support for literacy instruction at a greater frequency than control teachers, and they rated this support as "very" or "more than moderately" helpful at higher levels than control teachers.
- 43% (n = 13) of teachers reported being fully committed to Reading Apprenticeship at the end of the study.

However, implementation was not without challenges, with most teachers (over 60%) reporting competing priorities that hampered implementation, such as standardized test preparation or addressing content standards.

Key Findings about Impacts on Classroom Practice

Monthly teacher surveys measured the extent to which *iRAISE* had an impact on teacher mediating outcomes, including shifts in instructional practice and confidence in literacy instruction. *iRAISE* had significant impacts on teachers' use of certain core Reading Apprenticeship practices and on their confidence in delivering literacy instruction with effect sizes (ES) ranging from 0.236 to 0.619. The following were areas of impact.

- Teacher confidence in literacy instruction, ES = 0.619, p = .004
- Students practicing comprehension strategies, ES = 0.516, p = .001
- Students practicing metacognitive inquiry, ES = 0.457, p = .003
- Variety of text types, ES = .393, p = .033
- Fostering student independence, ES = 0.382, p = .034
- Traditional instructional strategies, ES = 0.329, p = .066
- Teachers instructing comprehension strategies, ES = 0.316, p = .04
- Student collaboration, ES = 0.285, p = .129

- Teachers modeling metacognitive inquiry, ES = 0.250, p = .086
- Teachers modeling comprehension strategies, ES = 0.243, p = .100
- Teachers instructing metacognitive inquiry, ES = 0.236, p = .095

The analyses of teacher survey data suggest *iRAISE* had an impact on reported attitudes and instructional practices in key areas emphasized by the Reading Apprenticeship framework. *iRAISE* teachers were more likely than control teachers to encourage student-directed learning by using practices that foster student independence, providing opportunities for students to practice various reading strategies, and offering opportunities for peer-to-peer learning and collaboration.

Key Findings about Impacts on Student Literacy Achievement

Student literacy achievement was measured through an online, scenario-based assessment that was developed by Educational Testing Service (ETS) as part of the Reading for Understanding grant funded by the Institute for Education Sciences. The assessment was designed to measure the strategic reading processes that are primary targets of Reading Apprenticeship and closely aligned with the Common Core State Standards. The assessment was designed to be a more rigorous measure of complex reading comprehension than typical state English Language Arts tests. While there was no impact of *iRAISE* on general reading literacy, we did find a differential impact of *iRAISE* based on prior student achievement, favoring students with lower incoming achievement. No differential impact was observed across other student subgroups. Also, based on a correlational analysis, we did not observe a relationship between the posited mediating outcomes and student achievement.

Conclusions

After a one-year implementation with *iRAISE,* we do not find an impact of the program on student achievement. However, we do find that the impact of *iRAISE* on general reading literacy increases with lower incoming achievement. This echoes prior research, as a previous study found effects on student achievement for students reading two to five years below grade level (Kemple et al., 2008). Additionally, we found a positive effect on classroom instructional practices, which replicated the results from the prior RAISE study, with significant impacts on fostering student independence, teachers instructing comprehension strategies, students practicing comprehension strategies, students practicing metacognitive inquiry, use of a variety of text types, and teacher confidence in literacy instruction. These findings are consistent with specific intended goals of *iRAISE*: to provide a high-quality online training that impacts teaching.

Reference this report: Jaciw, A. P., Schellinger, A., Lin, L., Zacamy, J., & Toby, M. (2016). *Effectiveness of Internet-Based Reading Apprenticeship Improving Science Education (iRAISE): A Report of a Randomized Experiment in Michigan and Pennsylvania.* (Empirical Education Rep.No. Empirical_ iRAISE-7027-FR1-Y2-O.3). Palo Alto, CA: Empirical Education Inc. Retrievable at http://empiricaleducation.com/pdfs/iRAISEfr.pdf

Table of Contents

Introduction 1

READING APPRENTICESHIP 1

NEED FOR AN ACCESSIBLE & LOW COST SOLUTION: *IRAISE* 2

IMPACTS OF *IRAISE*: WHAT AND FOR WHOM 2

Methods 4

EXPERIMENTAL DESIGN 4

How the Sample was Identified 4

Randomization 5

What Factors May Moderate the Impact of *iRAISE*? 6

What Factors May Mediate Between *iRAISE* and the Outcome? 6

SITE DESCRIPTION 7

IRAISE LOGIC MODEL AND OVERVIEW OF IMPLEMENTATION STUDY 8

SCHEDULE OF MAJOR MILESTONES 11

DATA SOURCES AND COLLECTION 11

Teacher Training Observations 12

Teacher Surveys 12

Principal Survey 15

Teacher Interviews 16

District/School Data Requests 16

ETS Literacy Assessment 16

FORMATION OF THE EXPERIMENTAL GROUPS 17

Baseline Sample 18

Analytical Sample 20

ANALYSIS AND REPORTING ON THE IMPACT OF *IRAISE* 23

Results 25

IMPLEMENTATION OF *IRAISE* 25

Implementation of Core Program Components 25
Contextual Factors of Implementation 33
Commitment to Reading Apprenticeship and Overall Impressions 37
IMPACT RESULTS 38
Overview 38
Impacts on Classroom Instructional Outcomes 38
Impact on Students 42
Moderation of the Impact 43
Teacher Mediating Outcomes 51
Discussion 56
OVERVIEW 56
IMPACTS ON STUDENTS 56
IMPLEMENTATION RESULTS 57
CONCLUSION 57
References 58
Appendix A. Considerations for Statistical Power 60
Appendix B. Details of the Approach to Estimating Impacts 62
Appendix C. Reporting the Results 65
Appendix D. A Post-Experimental Method to Assessing Impact under Strong Implementation 67
Appendix E. Fidelity of Implementation 68
Appendix F. Teacher Survey Constructs 71
Appendix G. Teacher Survey Construct Trends 71

Introduction

Empirical Education Inc. is the independent evaluator of WestEd's 2012 Investing in Innovation (i3) Development grant for Internet-based Reading Apprenticeship Improving Science Education (*iRAISE*). This report presents the results of a randomized control trial (RCT) during the 2014-2015 school year. The RCT measured the impact of *iRAISE* on classroom instructional practices (as measured by teacher surveys) and student reading literacy (as measured by the Educational Testing Service (ETS) literacy assessment) in high school science classes in 27 schools in Michigan and Pennsylvania.

READING APPRENTICESHIP

The Strategic Literacy Institute (SLI) at WestEd began developing the Reading Apprenticeship instructional framework over 20 years ago. Reading Apprenticeship impacts student learning styles, literacy skills, and content knowledge by utilizing four interconnected dimensions of classroom learning culture: social, personal, cognitive, and knowledge-building. The model is built on routines that create metacognitive conversation. These conversations take place both internally, as teachers and students learn and incorporate strategies for understanding complex texts, and externally, as students and teachers construct knowledge together, accompanied by a shift from teacher instruction to teacher modeling and student-to-student learning in pairs and groups. Metacognitive conversation is supported by an increase in both the opportunities to engage with—and the variety of—complex reading materials.

Grounded in 20 years of research and development (Greenleaf et al., 2008; Greenleaf et al., 2011; Somers et al., 2010), Reading Apprenticeship's inquiry-based professional development (PD) is designed to change teachers' understanding of their role in adolescent literacy development and to build capacity for literacy instruction in the academic disciplines. The PD model addresses the complexity of literacy and learning with disciplinary texts through the following.

- Experiential learning that mirrors the instructional environment and practices
- Learning how the framework supports students' literacy and learning
- Applying specific pedagogical practices
- Carrying out formative assessment focused on student reading, thinking and learning

Reading Apprenticeship aims to address several challenges facing 21st century education in the U.S. Across the country, two-thirds of high school students are unable to read and comprehend complex academic materials, think critically about texts, synthesize information from multiple sources, or communicate what they have learned (NAEP, 2013), while the new Common Core Standards call for all students to demonstrate advanced literacy proficiency not only in English classes, but also in academic subjects such as science (NCCSSO & NGA, 2010). Unless targeted at the high school level, students can expect to struggle with complex academic texts in secondary and post-secondary education (ACT, 2012). Currently, teachers report that little time is devoted to supporting reading comprehension beyond basic summarizing (Ness, 2008, 2009; Vaughn et al., 2013), particularly in content areas. In science education, the Reading Apprenticeship approach is premised on the idea that

to support the shift to Next Generation Science Standards, students need to move beyond memorization of facts and towards a deep understanding of science knowledge and practices (NGSS, 2013). Literacy skills are essential for this understanding. Through Reading Apprenticeship's metacognitive conversations, students are expected to gain the skills needed to move beyond rote interaction with scientific literature to actively building their own knowledge and engaging in science.

NEED FOR AN ACCESSIBLE & LOW COST SOLUTION: *iRAISE*

Facing the need for accessible, low-cost, high-quality PD that prepares teachers and students for literacy skills in a STEM-focused world, *iRAISE* brings SLI's 65-hour biology-based face-to-face literacy PD to an online format, with the hope of cutting the cost of previous face-to-face training by half. *iRAISE* is a year-long learning community in which high school science teachers learn about, practice, and refine ways to improve their students' ability to engage in and understand a variety of scientific texts. *iRAISE* builds from the existing materials, protocols, and key design elements of face-to-face Reading Apprenticeship PD and leverages interactive, internet-based technologies to enhance teachers' learning. The *iRAISE* course is divided between online synchronous sessions with facilitators and peers, and asynchronous assignments. *iRAISE* PD begins with a 5-day (approximately 20-hour) *iRAISE* Foundations training during the summer prior to classroom implementation. Each day includes four hours of synchronous work with a large group of teachers (roughly 20 in each group), as well as an hour and a half of personal, asynchronous time for reading, reflection, and posting on the discussion board. After the start of the school year, teachers participate in monthly follow-up meetings from September through May, allowing them to continuously implement their learning over the year. The follow-up meetings provide three hours of additional support per month in two different formats: whole-group meetings introducing new learning (Ignite sessions) and small-group meetings intended to produce discussion and collaboration (PLC sessions). The online content itself is presented across multiple platforms, including BlackBoard Collaborate for synchronous work, Canvas for course management, and YouTube and GoogleDocs for resource storage and sharing. Use of these interactive spaces encourages the collaborative nature of Reading Apprenticeship, wherein teachers become students and learn alongside each other. With their development grant, SLI aimed to create a flexible, high-quality online platform that can cut costs for schools and districts while meeting the needs of modern school systems.

SLI piloted the *iRAISE* program in 2013-14 with a group of 25 teachers in Michigan and Pennsylvania, several of whom had previously attended the face-to-face RAISE training. We conducted a formative evaluation during this pilot to provide feedback on program components, and teachers' impressions of the program were overwhelmingly positive. Additionally, attendance at the PD sessions and implementation of core performance measures exceeded the expectations of the program developers.

IMPACTS OF *iRAISE*: WHAT AND FOR WHOM

The impact study that followed on the pilot work involves analysis to assess the effect of *iRAISE* on classroom instructional practices and high school students' general reading literacy skills after one year. We were also interested in whether we would find advantages with Reading Apprenticeship for low performing students, a finding that would replicate an earlier study by Kemple et al. (2008), as

well as the related question of whether impact increases for students with lower incoming achievement. We also were interested in whether impacts varied by socioeconomic status and English learner status, although for the latter, the low number of English learners in the final sample did not support firm conclusions.

This study drew on a larger RCT of the Reading Apprenticeship Improving Secondary Education (RAISE) project that began four years earlier, funded through an i3 validation grant (Fancsali et al., 2015). The current study used the same outcome measure that was used and validated through the earlier RCT. General reading literacy was assessed using the biology form of a test developed by ETS. The test was designed to measure general reading literacy, not content knowledge, across different subjects, with forms developed for history, English Language Arts (ELA), and biology. In this report, the assessment will henceforth be referred to as the ETS assessment.

The RAISE study, which ran from 2010-2015, evaluated the impact of a face-to-face version of the Reading Apprenticeship PD in ELA, history, and biology, utilizing similar research questions to this study. Teacher classroom practices were evaluated using constructs from the monthly surveys that are also used in this study. The RAISE study found a positive impact with an effect size of 0.32 on the ETS assessment for students in science classes, and found an impact on several teacher mediating outcomes, including employing practices that foster student independence; providing opportunities for students to practice metacognitive conversations; providing opportunities for students to practice comprehension strategies; providing opportunities for student collaboration; and teacher confidence in literacy instruction. In that study, impacts were assessed after two years.

For *iRAISE*, we were able to take advantage of analyses conducted with the RAISE data in order to refine the questions to be asked in *iRAISE* prior to examining and beginning analysis of those data (Jaciw, Newman, Lazarev, Lin, & Ma, 2016). The questions were finalized at the beginning of the study to ensure that the analyses were not prioritized based on a post-hoc "fishing" of results. The following are the research questions.

1. Is there a positive impact of *iRAISE* on classroom instructional practices, after one year, as measured by teacher surveys?
2. Is there a positive impact of *iRAISE* on general reading literacy outcomes, after one year, as measured through an ETS assessment of the construct?
3. Is there a differential impact of *iRAISE* on general reading literacy, after one year, depending on student English Language Learner (ELL) status, socioeconomic status (SES), or prior achievement?
4. Are impacts of *iRAISE* on student general reading literacy, after one year, mediated through impacts on teacher literacy instructional practices?

In addition to this, we ran several analyses to further clarify certain results. A question of importance not directly addressed in this research is whether the lower-cost, online *iRAISE* training is as effective as the face-to-face training used more commonly for Reading Apprenticeship and analyzed in the RAISE RCT. That is, we did not randomize teachers between online and face-to-face. Additionally, the majority of the results from the RAISE RCT are from the *second* year of implementation: teachers

receive an initial 5 days of training in the summer, 2 more days in the following winter, and the final 3 days of training in the next summer, with teacher and student outcome data taken from the *subsequent* year. In this study, teachers implement as they attend the PD and report on their classroom practice concurrently. Nevertheless, we were interested in whether we would see similar impacts on teacher classroom practices in the two studies. This would provide an indirect indication that the two modalities were fostering similar processes. This report will provide comparisons between the two studies when appropriate, offering a chance to compare the modalities.

For this experimental study, Empirical worked with *iRAISE* program managers and state coordinators towards the goal of initially recruiting 100 teachers to participate. 117 teachers expressed interest in the research study, but after eligibility criteria and teacher consent, the randomized sample included 82 teachers. We divided these 82 teachers into two groups: a group of teachers who would train to and use *iRAISE* (*iRAISE* group) and a group of teachers who would continue with their existing program (control group), that is, "business as usual." First, we paired teachers, primarily within each school, and then, we used a random number generator to determine which teacher in each pair would join the *iRAISE* group and which teacher would be a control.

An RCT eliminates a variety of biases that could otherwise compromise the validity of the research. For example, it ensures that teachers in both groups were not selected on the basis of their interest in trying *iRAISE* or their ability to take advantage of the new program. Random assignment to experimental conditions does not, however, ensure that we can generalize the results beyond the schools where the research was conducted, and the results are not applicable to schools with practices and populations different from those in this experiment. This report includes a rich description of the conditions of the implementation to provide the reader with an understanding of the context for our findings.

Methods

This section outlines the experimental design and explains how we made decisions with regard to how many teachers to recruit and how teacher pairs were formed for the randomization process. Our experiment results in a comparison of outcomes for teachers who were randomly assigned to *iRAISE* and teachers using the schools' current methods, where the outcomes of interest are classroom instructional practices, as measured through teacher surveys, and student test scores on the biology form of the ETS literacy assessment. This section details the methods we used to assess the impact of *iRAISE*. We begin with a description and rationale for the experimental design and go on to describe the program, the research sites, the sources of data, and the composition of the experimental teacher groups.

EXPERIMENTAL DESIGN

How the Sample was Identified

The *iRAISE* sample was one of convenience, chosen from school districts that responded to invitations from the *iRAISE* State Coordinators in two states: Michigan and Pennsylvania. Initial recruitment materials were sent around in early 2014, and interested districts were given until the end of March to

submit an application. Empirical met regularly with the state coordinators to discuss potential districts and obtain progress updates. For this experiment, we recruited schools that had at least two teachers interested in participating in the research study. Interested districts assigned a point of contact responsible for obtaining contact information for interested teachers and consent from district-level personnel. Eligibility criteria were established: eligible teachers would teach at least one of the five major science topics (physics/physical science, chemistry, biology, earth/environmental science, and general/integrated science) in at least one regular (not AP/honors, ELL, or special education) section. While some schools had teachers with limited amounts of exposure to Reading Apprenticeship concepts, any teachers who had previously attended the 10-day RAISE training were ineligible.[1] Additionally, 5 of the 27 school sites may have contained teachers who participated in Reading Apprenticeship training during the prior RAISE grant. We do not know if students of study (either *iRAISE* or control) teachers in these schools may have been exposed to Reading Apprenticeship through classes other than their target class they were enrolled in during the study. However, since participating teachers and their classes were randomly assigned, we can assume that such exposure will be balanced between conditions.

Randomization

Twenty-eight schools in 27 districts submitted applications, for a grand total of 117 teachers. After applying eligibility criteria and obtaining consent from teachers, principals, and district personnel, the sample randomized was 82 teachers. With one exception, all schools were in different districts, for a total of 26 schools in 25 districts. To achieve adequate statistical power, teachers were randomized within schools. (Randomization of schools would have resulted in a low number of units of randomization and could easily have led to an underpowered study and a high probability of not rejecting the null hypothesis, even with an appreciable effect.)[2] Appendix A provides a detailed

[1] There were three control teachers and four treatment teachers who said they had received between 1 – 3 days of "RA training" from their district or IU. Additionally, six of the 27 school sites may have contained teachers who participated in Reading Apprenticeship training during the prior RAISE grant. We do not know if students of study (either *iRAISE* or control) teachers in these schools may have been exposed to Reading Apprenticeship through classes other than their target class they were enrolled in during the study. The effect of this could be to reduce impact in those schools if controls had exposure to previously trained teachers who are continuing to use Reading Apprenticeship (although this boost received by control students could be offset by the students of treatment teachers getting an additional dose of the program through those same previously trained teachers.) Regardless, we addressed the issue by re-running the main impact analysis after removing the 6 schools that may have contained previously exposed teachers. Where randomization is of teachers within schools (which is true in most cases with *iRAISE*), each school may be considered a mini-experiment, and the removal of entire schools does not compromise the equivalence achieved through randomization in the remainder of the sample, so that the remaining sample of schools provides a fully experimental assessment of impact. (Two teachers in the six schools were paired with teachers in schools other than the six. This is unlikely to affect the result.) In the case of *iRAISE*, the analysis without the six schools did not change the result.

[2] With the number of teachers that were available for this study, we estimated that the smallest effect size we can detect is an absolute difference of seven percentile points for the ETS literacy assessment for a student who performs at the median of the distribution. This effect size is what we would see if we took a student who performs at the 50th percentile of the distribution of posttest performance for the *iRAISE* group and found that student's score to be seven percentile points higher (i.e., at the 57th percentile) or seven percentile points lower (i.e., at the 43rd percentile) than the median score for the control distribution. We can also express this difference as a standardized effect size, that is, in units of the standard deviation of posttest performance. In that metric, with the expected sample size, we would be able to detect an impact of 0.19. (The factors we considered in running the power analysis are described in Appendix A: *How Large a Sample Do We Need*.)

description of the sample size and power analysis. For the randomization process, teachers in schools with an even number of participants were first paired together based primarily on the subjects they taught in the 2013-2014 school year, and secondarily on their years of teaching experience. The process was then extended to individual teachers who were left unpaired because of an odd number of participants in the school, including cases where a teacher was the only eligible participant at the school. For these remaining teachers, pairs were formed across schools with similar district-level demographics.

To meet the resource constraints of the grant, it was necessary to select one section per teacher as the target class. This class would be the section tested on the ETS literacy assessment and the focus of the monthly teacher surveys. The selection was made after fall 2014 rosters were determined. Sections were chosen without knowledge of whether the teacher had been randomly assigned to *iRAISE* or control to prevent potential for biased section selection. As much as possible, sections were selected: to maintain the subject similarity of members within matched pairs, using several criteria (science subject, AP/Honors, ELL, and Special Ed status), and to reflect balance across the included science subjects and ensure a representative sample of each teacher's science course content.

What Factors May Moderate the Impact of *iRAISE*?

We assessed whether the impact of *iRAISE* varies for different kinds of students. In other words, we evaluated whether characteristics of students moderate the impact of the program. In this study, we explored the program's effectiveness based on ELL status, SES, and pretest scores. We chose these particular moderators because of their prior inclusion in the RAISE i3 Validation study and previous findings in the Enhanced Reading Opportunities study (Kemple et al., 2008; Somers et al., 2010). To provide additional context for the results, we also explored whether impacts varied across the science subject areas in which *iRAISE* was implemented: biology, chemistry, physics, and a category consisting primarily of earth science but that included also some environmental and general science classes. Further, we analyzed whether impacts varied depending on levels of program implementation by teachers. We examined the effects of the moderating characteristics one at a time, rather than analyzing them simultaneously.

What Factors May Mediate Between *iRAISE* and the Outcome?

A mediator is an intermediate outcome, such as an instructional practice, which is impacted by the program and that facilitates impact on more distal outcomes such as student achievement. A mediator can either block or intensify the effect of an intervention, either entirely or in part. For purposes of this study, we can think of a mediating process as occurring in two steps, from random assignment to the mediator, such as an instructional practice, and from the mediator to the distal outcome, such as student achievement. Mediation analyses tell us about plausible causal pathways between randomization and impact on student achievement.

There are several formal approaches to mediation analysis. Generally, they are considered to have limited statistical power and require very large samples. We chose to use a descriptive approach, with analysis performed in two steps: first, by examining the impact of *iRAISE* on each of the mediators,

and then, by assessing the relationship between each mediator and the ETS literacy assessment, while controlling for the effects of a series of covariates. We deemed this two-step exploratory approach to be potentially more informative for supporting hypothesis-building. That is, looking at the two steps individually would tell us about each process and allow for more power to analyze each component. In the two-step process, we were interested in identifying the mediators that were both impacted by *iRAISE* and related to student achievement. The approach may be considered a mixture of causal and correlational analyses: while the first step tells us about the causal impact of the program on the mediator, the second step describes the association between the hypothesized mediating variable and achievement (after adjusting for the effects of other variables that may be associated with the mediator and the outcome).

SITE DESCRIPTION

The 27 study schools are spread equally across the two states, with 13 in Michigan and 14 in Pennsylvania, and nearly equally across the four National Center for Education Statistics (NCES) locale designations, with slightly more suburban and fewer urban schools. Table 1 shows the school-level averages for the 27 schools from publically available NCES data and district data provided on the research application.

TABLE 1. AVERAGE DEMOGRAPHICS OF PARTICIPATING SCHOOLS

Demographics	
Locale Designations	
Rural	25%
Town	25%
Suburban	29%
City/urban	21%
Full-time equivalent teachers	58
Student to teacher ratio	16.3
Student Characteristics	
Student population	974
Mobility rate	17%
Dropout rate	4%
Free and reduced price lunch eligible	54%
Graduation rate	87%
Special education	16%
ELL [a]	3%
White	70%
Black	12%
Hispanic	16%
Asian	1%
American Indian / Native Alaskan	0.3%
Multi racial / No response	2%

[a] ELL stands for English Language Learners

Source. NCES 2012-2013 school year and data provided by school districts on study applications

Note. Percentages may not add up to 100% due to rounding of decimals.

iRAISE LOGIC MODEL AND OVERVIEW OF IMPLEMENTATION STUDY

During the 2013-2014 school year, Empirical worked with the developers of *iRAISE* to develop a program logic model and identify the key components of the intervention (Figure 1 below). There are three key components: delivery of the PD by the program developers, attendance of the PD by teachers, and adherence of the PD to Reading Apprenticeship principles, representing the inputs and outputs of the logic model. These are intended to impact teacher classroom use of reading comprehension and metacognitive strategies, thereby increasing student collaboration, engagement, and motivation in literacy practices, which would in turn increase achievement on literacy assessments, especially among low-performing students.

As a requirement of the National Evaluation of i3 (NEi3), we have calculated fidelity of implementation (FOI) scores for each component of the *iRAISE* program. The implementation study applies mixed methods to assess the key components of the logic model, including: presence of inputs, such as the delivery of PD by SLI; the quality of inputs measured through the alignment of the PD to five key characteristics (content focused on science literacy, collective participation, coherence, active learning, and metacognitive inquiry); and recorded levels of activities in terms of outputs, such as teacher attendance at Foundations training, Ignite and PLC sessions, and completion of monthly assigned work. We have assessed implementation fidelity in terms of the following components: (1) SLI delivery of PD, (2) teachers' participation in *iRAISE* professional learning activities, and (3) PD adherence to the principles of Reading Apprenticeship. Program components and fidelity indicators are shown in the fidelity matrix, and a longer description can be found Appendix E.

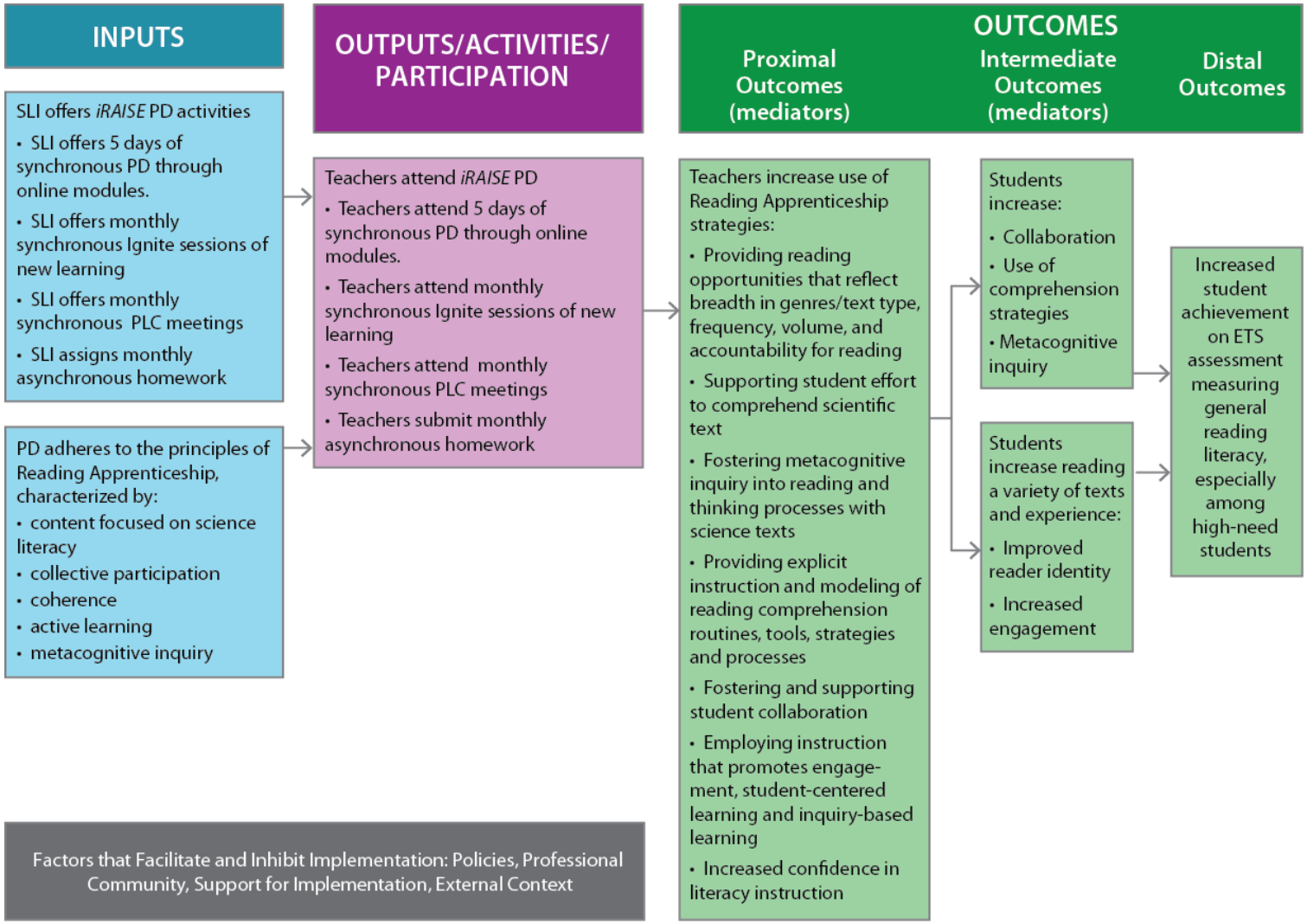

FIGURE 1. LOGIC MODEL

SCHEDULE OF MAJOR MILESTONES

Table 2 lists the major milestones in this study and associated dates.

TABLE 2. MILESTONES

Target date	Data collection event / Planning activity
2013 - 2014 school year	*iRAISE* pilot study
January 2014	State Coordinators send recruitment materials and answer initial questions about *iRAISE* RCT
March 28, 2014	Final deadline for *iRAISE* RCT district applications
April 18, 2014	Final deadline for district agreements
May 16, 2014	Final deadline for teacher consent forms
May 16 - 30, 2014	Empirical forms matched pairs of teachers within schools, solicits principal feedback, and finalizes matched pairs
May 31, 2015	Empirical randomizes teachers to *iRAISE* or control group
June 2, 2014	Empirical provides results of randomization to SLI and participating teachers and schools
June 2014	SLI contacts teachers assigned to the *iRAISE* group about the summer PD and scheduling
August 11 - 15, 2014	Reading Apprenticeship Science Foundations Training
August 18 - 22, 2014	Reading Apprenticeship Science Foundations Training
August 2014 - May 2015	Empirical deploys monthly teacher surveys during the school year
September 2014 - May 2015	Monthly Ignite Sessions and Professional Learning Communities (PLCs)
Spring 2015	Empirical coordinates ETS literacy assessment administration and obtains student posttests
Summer 2015	Empirical collects student demographic and historical achievement data from districts

Source. Empirical Education staff

DATA SOURCES AND COLLECTION

The data for this study were provided by the school districts and collected by Empirical Education. In addition to achievement and demographic data, we collected implementation data over the entire period of the experiment, beginning with the teacher trainings in August 2014 and ending with the schools' academic calendars in June 2015. Data collected through teacher background forms, training observations, multiple teacher surveys, principal surveys, *iRAISE* log data, and teacher interviews were used to provide evidence of the implementation. In addition, we have reviewed various program documents and materials. Table 3 outlines the timeline of the major data collection phases.

TABLE 3. DATA COLLECTION SCHEDULE FOR THE *iRAISE* STUDY

	2014-2015 school year									
Data collection elements	**Aug**	**Sept**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**April**	**May**
Training observations	[X]	[X]	[X]	[X]	[X]	[X]	[X]	[X]	[X]	
Teacher surveys	[X]	[X]	[X]	[X]	[X]	[X]	[X]	[X]	[X]	[X]
Principal survey										[X]
Teacher interviews										[X]
District/school data request (for student demographic and prior achievement data)				[X]			[X]			
ETS assessment									[X]	[X]

Source. Empirical Education staff

Teacher Training Observations

We observed the initial *iRAISE* Foundations training and asked questions about the initial training on teacher online surveys. The *iRAISE* group was divided into two courses, each with 16 or 17 participants, and all five days were observed for both courses. We also observed a sample of the monthly Ignite sessions (randomly choosing one of the courses each month, September through May) and coded the PD content for adherence to the model as part of our implementation study.

Teacher Surveys

Prior to randomization and the initial training for the research study, teachers attended an informational session (through a webinar) outlining the study requirements. Teachers then received a Participant Information Packet as part of an initial online survey. This packet provided general information about the research study, data collection activities, and participant responsibilities, in addition to the study consent form. The survey also included teacher background questions for teachers to answer, providing researchers with information about their teaching history and contact information. We used this information to help describe the context of implementation and to inform our selection of matched pairs at the start of the trial.

Further surveys were deployed on a monthly basis to participating teachers beginning in August 2014 and ending in May 2015. Table 4 outlines the survey schedule and response rates for the control and *iRAISE* teachers participating in the study.

TABLE 4. SURVEY PARTICIPATION RATES

Survey	Date	Response rates		
		Control group	*iRAISE* group	Total
Consent/background	May 20014	100%	100%	100%
Training	August 2014	N/A	100%	100%
Monthly Survey 1	September 2014	100%	92%	96%
Monthly Survey 2	October 2014	95%	87%	91%
Monthly Survey 3	November 2014	97%	89%	93%
Monthly Survey 4	December 2014	97%	82%	90%
Monthly Survey 5	January 2015	97%	82%	90%
Monthly Survey 6	February 2015	97%	82%	90%
Monthly Survey 7	March 2015	97%	82%	90%
Monthly Survey 8	April 2015	92%	79%	86%
Monthly Survey 9	May 2015	92%	79%	86%
Total		97%	87%	92%

n = 76 total, *n* = 38 each for *iRAISE* and control groups

The teacher surveys were extensively piloted, both through the prior RAISE RCT and the prior year's *iRAISE* pilot study, and revisions were made to capture more detail on the variation in time spent in classroom activities and the level of student engagement. Typical questions asked about the number of minutes or class periods spent on literacy strategies during a specific week each month or the students' level of engagement in different types of activities in their target class. Along with the target class questions, teachers reported in each monthly survey on the context for literacy instruction, including support from other teachers and administrators.

The *iRAISE* logic model (see Figure 1) hypothesizes that the intensive 65 hours of PD, including the ongoing nature of support through Ignite sessions and PLCs, will have an impact on teachers' instructional practices and routines. Informed by the RAISE and *iRAISE* models, we used the teacher survey data to create 12 constructs intended to capture the effects of Reading Apprenticeship on the following dimensions of teacher behavior and attitudes (see Table 5). These are the same constructs used in the RAISE RCT to measure impacts on classroom practices.

- Providing extensive reading opportunities that reflect a variety of genres and text types (measured by construct 1)
- Supporting student effort to comprehend disciplinary text (measured by construct 2)
- Fostering metacognitive inquiry into reading and thinking processes (measured by constructs 4-6)
- Providing explicit instruction and modeling of reading comprehension routines, tools, strategies, and processes (measured by constructs 7-9)

- Fostering and supporting student collaboration (measured by construct 10)
- Employing instruction that promotes engagement, student-centered learning, and inquiry-based learning (measured by construct 11)
- Confidence in delivering literacy instruction (measured by construct 12)

Construct 3—measuring the extent to which teachers employed traditional instructional strategies such as lecture and using quizzes to assess comprehension—represents a contrast to the Reading Apprenticeship approach. Therefore, we did not expect *iRAISE* to have an impact on these strategies. (See Appendix F for further description of the 12 constructs).

TABLE 5. TEACHER SURVEY CONSTRUCTS

Construct number	Construct description	Possible construct range		Reliability
		Min	Max	Alpha
1	**Variety of Text Types** Total number of text types that a teacher asked students to work with over a week, in or outside of class (e.g., newspapers, textbooks, historical documents)	0	7	0.19
2	**Fostering Student Independence** Total number of minutes over a week that a teacher uses practices to foster independence, such as providing guided practice of reading comprehension strategies and having students teach other students	0	12	0.37
3	**Traditional Instructional Strategies** Total number of minutes over a week that a teacher employs traditional strategies, such as direct instruction and giving quizzes to assess comprehension.	0	12	0.76
4	**Teachers Instructing Metacognitive Inquiry** Total number of metacognitive inquiry strategies in which teachers provided instruction over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	0	4	*
5	**Teachers Modeling Metacognitive Inquiry** Total number of metacognitive inquiry strategies that teachers modeled during their class over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	0	4	*
6	**Students Practicing Metacognitive Inquiry** Total number of metacognitive inquiry strategies that students practiced during class over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	0	4	*

TABLE 5. TEACHER SURVEY CONSTRUCTS

Construct number	Construct description	Possible construct range		Reliability
		Min	Max	Alpha
7	**Teachers Instructing Comprehension Strategies** Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) in which teachers provided instruction over a week	0	8	0.37
8	**Teachers Modeling Comprehension Strategies** Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) that teachers modeled during the class over a week	0	8	0.39
9	**Students Practicing Comprehension Strategies** Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) that students practiced during class over a week	0	8	0.42
10	**Student Collaboration** Total number of minutes over a week that teachers had students work on reading and writing activities in pairs, in small groups, and as a whole class	0	15	0.63
11	**Student Engagement** Total of teachers' ratings on the proportion of students in their class completing homework, paying attention in class, and participating in class activities	0	15	0.68
12	**Teacher Self-Confidence in Literacy Instruction** Total of teachers' ratings on their confidence in their ability to provide literacy instruction, such as providing opportunities for reading a variety of texts of different genres and teaching students to analyze their own thinking about texts	0	55	0.86

Source. Empirical Education staff calculations

*A Cronbach Alpha coefficient could not be estimated for these scales.

Principal Survey

All principals who agreed to participate were sent a one-time survey in May 2015, with 58% (n = 15) of principals responding. This survey gathered school-level information on the context for implementation of the *iRAISE* program, including types of support for literacy instruction and factors that may inhibit implementation.

Teacher Interviews

A sample of teachers participated in brief, semi-structured interviews in May 2015. A sample of 25% of the *iRAISE* teachers was identified to be representative of the *iRAISE* group, based on subjects taught, school characteristics, and levels of implementation. Eight interviews were conducted; these interviews gathered valuable information on the context for implementation—including challenges and supports—and provided an opportunity for teachers to reflect on the benefits and drawbacks of the *iRAISE* program.

District/School Data Requests

We requested and collected class rosters from each school in fall 2014 to familiarize teachers with the data collection process and to allow us to track attrition at the student level. We then requested updated spring semester target class rosters in February 2015. For the purposes of the ETS assessment, we requested target class rosters and student IDs from each district. We provided temporary IDs to ETS and then matched the ETS scores with demographic data from the districts, including the standardized assessment pretest scores. These data were required to conduct equivalence tests and moderator analyses. Specifically, we asked the districts to provide the following student data.

- Name
- Unique identifier
- Gender
- Ethnicity
- English proficiency status
- Disability status (whether or not student has a disability or is in special education, but not the specific condition)
- Date of birth
- Grade
- Classroom teacher name and unique identifier
- Course name and section
- School name
- Pretest scores (Science and ELA Michigan Educational Assessment Program, ACT, and WorkKeys in MI; Science and ELA Pennsylvania State Standardized Assessment and Biology and Literature Keystone Exams in PA)

ETS Literacy Assessment

The scenario-based literacy assessment was developed by ETS as part of the Reading for Understanding grant funded by the Institute for Education Sciences (IES, 2010), and used in the RAISE study, as well as multiple other data collection efforts at the secondary level. Based on the Global Integrated Scenario-Based Assessment, ETS designed the assessment to measure how well students read and reason about text sources in a discipline where they have been exposed to content and

strategies for understanding text (O'Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014). There were three forms of the assessment developed: biology, history, and English language arts (ELA). This study used the biology form, which included texts on ecosystems and invasive species. It assessed a variety of purposeful literacy activities in which students are expected to read multiple texts for understanding. The scenarios organized the assessment around a theme and goal for reading; for example, students were asked to imagine they were studying for an exam or preparing for a presentation. They were then asked to participate in a sequence of tasks that would lead to a final goal, such as identifying important ideas and meaning, evaluating sources, or integrating information across multiple sources. The assessments were not designed to assess or be dependent on specific science content knowledge, but rather to assess student literacy skills in the context of science.

The scenario-based assessment was pilot tested in fall 2011 and spring 2012 to collect evidence of its psychometric properties. The assessment displayed adequate reliability for each of the subject-area forms ($r = 0.84$ for biology), and modestly high correlations with available state standardized tests in ELA suggesting that the literacy assessment captured some of the same underlying constructs related to reading comprehension as those state tests. Finally, psychometric testing also showed sufficient range and variability in scores, with no evidence of ceiling or floor effects (O'Reilly et al., 2014). For further information on the ETS literacy assessment, see the final RAISE report (Fancsali et al., 2015).

For the *iRAISE* evaluation, testing using the ETS assessment took place in a window from the beginning of April through the end of May 2015. Empirical Education coordinated with ETS and designated point-of-contacts in each school to ensure the appropriate technology was available for the computer-based test and sufficient training and documentation was provided to all teachers. During the testing window, Empirical monitored the assessment response rate and remained in frequent communication with the point of contacts to prevent attrition.

FORMATION OF THE EXPERIMENTAL GROUPS

This section describes the study sample that we used to assess the impact of *iRAISE*. We started with the baseline sample which consisted of the participating teachers who were randomly assigned to the *iRAISE* or control group and for whom we had information. The sample for which outcomes were analyzed may have been modified somewhat from baseline through attrition or for other reasons that data become unavailable.

Baseline Sample

The *baseline sample* consists of the teachers randomized to *iRAISE* or control, with their students.

Ideally, by randomizing assignment into the two conditions, we create groups that look the same in terms of important characteristics, including demographics and prior achievement. However, by chance, the groups are never exactly balanced and may differ on important characteristics that may affect the outcome. Therefore, in this section we compared the distributions of background characteristics for teachers and students and assessed whether they were balanced between the *iRAISE* and control groups.

In Table 6 and Table 7, we compared the composition of the control and *iRAISE* teachers and students, respectively, at the point we received the rosters (baseline sample). For each of the characteristics of this sample, we conducted a statistical test[3] to determine the probability of observing a difference as large as or larger than the one measured when in fact there is no difference. While the randomization assures us that any imbalance was a result of chance, and is not an indication of selection bias, it is useful to examine the actual groups as formed at baseline to see whether the amount of imbalance is something we would expect to see less than 5% of the time (the standard conventionally used to assess if an effect is statistically significant). We see that balance is achieved on the observed characteristics.

TABLE 6. TEACHER BACKGROUND CHARACTERISTICS BY CONDITION

	Control	*iRAISE*	Less than 5% chance of seeing this imbalance
Male	14 (37%)	12 (32%)	No
Mean years teaching experience	13.2	15.5	No
Mean years science teaching experience	12.6	14.9	No
Bachelor's degree	12 (32%)	8 (21%)	No
Master's degree	19 (50%)	26 (68%)	No
Advanced degree	6 (16%)	3 (8%)	No
Degree in science	36 (95%)	36 (95%)	No
Regular certification	35 (92%)	36 (95%)	No
Prior Reading Apprenticeship exposure	5 (13%)	3 (8%)	No

Source. Empirical Education staff calculations

[3] To assess the baseline equivalence of student characteristics we used a t-test that adjusted for clustering of students in sections. The criterion for significance was set at < .05.

TABLE 7. CHARACTERISTICS OF STUDENT SAMPLE ON ROSTERS RECEIVED (BASELINE SAMPLE)

	Control	*iRAISE*	Less than 5% chance of seeing this much imbalance
Asian[a]	3 (0%)	16 (2%)	No
Hispanic	60 (6%)	88 (9%)	
Black	178 (18%)	160 (17%)	
White	729 (74%)	647 (70%)	
Unspecified[a]	11 (1%)	18 (2%)	
Male	505 (51%)	466 (50%)	No
Receiving free or reduced-price lunch	520 (53%)	505 (54%)	No
Disabled students	150 (15%)	147 (16%)	No
English speaker	978 (< 100%)	921 (99%)	No
Grade 8	1 (0%)	0 (0%)	No
Grade 9	343 (35%)	266 (29%)	
Grade 10	349 (36%)	243 (26%)	
Grade 11	218 (22%)	283 (31%)	
Grade 12	70 (7%)	134 (14%)	
Recalibrated science pretest[b]	0.00	- 0.01	No
Recalibrated reading pretest[b]	- 0.00	- 0.01	No

[a] Given the low counts, the results in this row may be inaccurate and should be interpreted with caution.

[b] Pretests were z-transformed within grade using the mean and standard deviation for controls.

Note. The effect size for the pretest is the mean difference between the *iRAISE* and control group in the pretest scores for students, expressed in units of the pooled within-group standard deviation of the pretest.

Analytical Sample

The analytical sample consists of participants actually used in the impact analysis, since some teachers and students were lost during the course of the experiment. The loss of units randomized—in this case teachers—during the experiment may cause the difference between conditions on the outcome to reflect imbalance on background characteristics, instead of differences caused by being exposed to *iRAISE.*

If the rate of overall attrition is large, even if there is no difference between conditions in the rate of attrition, then a loss of cases may induce bias in the result if those who leave the program group are different from those who leave the control group. If the rate of differential attrition is substantial, even if those who leave the two conditions are not fundamentally different, then the difference in the rate of attrition can induce bias in the result. We adjusted for effects of characteristics of individuals that could potentially produce bias in these ways.

Table 8 shows changes in the samples from the point at which the teachers were randomized to the time when ETS posttests were received.

Immediately after randomization, three control teachers and one *iRAISE* teacher declined to participate in the research for reasons exogenous to the study (all four left their schools). Before the start of the study, two additional *iRAISE* teachers also declined to participate in the study, leaving the sample with 76 teachers, evenly balanced across *iRAISE* and control. Five of the 38 remaining *iRAISE* teachers declined to participate in the PD but still participated in the data collection activities, leaving 33 teachers receiving the full program. In addition to the loss of teachers described above, we considered those teachers, for whom we did not receive student posttests, to be lost to attrition. There were seven such teachers, three in *iRAISE* and four in control. The final analytical sample for assessing impacts on students included 35 *iRAISE* and 34 control teachers. Because teachers were represented in the analysis based on the results of random assignment status, and we include outcomes data from five teachers who declined to participate (i.e., teacher "no-shows"), the impact analysis reflects the effect of being randomly assigned to *iRAISE* compared to business-as-usual, rather than the impact of the program for just those who complied with random assignment to treatment. In other words, we report the results of the analysis of "intent to treat."

TABLE 8. NUMBERS OF UNITS IN THE EXPERIMENTAL GROUPS AND ATTRITION OVER TIME

	Control			*iRAISE*		
Event	**No. of schools**	**No. of teachers**	**No. of students**	**No. of schools**	**No. of teachers**	**No. of students**
Randomization	23	41	n/a	24	41	n/a
(Immediate loss exogenous to assignment status)	0	3	n/a	0	1	n/a
(Additional early attrition)	0	0	n/a	0	2	n/a
(Loss due to lack of posttest)	-	4	235	1	3	223
Final count of units with ETS test	23	34	751	23	35	717

At the student level, while 1,926 students were registered in the ETS system reflecting the number of students on participating teachers' rosters, we received posttests for 1,468 students, which constitutes the analytic sample of students used to estimate the impacts of *iRAISE*. (We lack posttests for 223 out of 940 [23.72%] of *iRAISE* students and 235 out of 986 [23.83%] of control students. This represents a 0.11% rate of differential attrition at the student level.)

As with the baseline sample, we conducted a series of statistical tests to assess baseline equivalence for the analysis sample. Table 9 shows that for each of the characteristics examined, the level of imbalance between conditions was low: none of the differences in average outcomes exceeded levels that we would expect to see by chance 5% of the time. Table 9 shows the equivalence of the analytic sample.

TABLE 9. CHARACTERISTICS OF STUDY SAMPLE (ANALYTICAL SAMPLE FOR ETS LITERACY ASSESSMENT)

Background characteristics	Control	*iRAISE*	Less than 5% chance of seeing this much imbalance	Effect size
Asian[a]	3 (<1%)	15 (2%)	No	n/a
Hispanic	48 (6.41%)	64 (9.04%)		n/a
Black	122 (16%)	113 (16%)		n/a
White	567 (76%)	499 (70%)	No	n/a
Unspecified[a]	9 (1%)	16 (2%)		n/a
Male	384 (51%)	348 (49%)	No	n/a
Receiving free or reduced-price lunch	385 (51%)	374 (53%)	No	n/a
Disabled students	116 (15%)	114 (16%)	No	n/a
English speaker	746 (99%)	700 (99%)	No	n/a
Grade 9	245 (33%)	224 (32%)	No	n/a
Grade 10	296 (40%)	178 (25%)		n/a
Grade 11	168 (22%)	202 (29%)		
Grade 12	40 (5%)	101 (14%)		n/a
Recalibrated science pretest[b]	0.03	0.02	No	- 0.02
Recalibrated reading pretest[b]	0.02	0.00	No	- 0.02

[a] Given the low counts, the results in this row may be inaccurate and should be interpreted with caution.

[b] Pretests were z-transformed within grade using the mean and standard deviation for controls.

Note. The effect size for the pretest is the mean difference between the *iRAISE* and control group in the pretest scores for students, expressed in units of the pooled within-group standard deviation of the pretest.

ANALYSIS AND REPORTING ON THE IMPACT OF *iRAISE*

Approach to Analysis

Before presenting the results, we discuss briefly the approach to analysis.

Impacts on teachers and students were analyzed using standard approaches for cluster randomized trials. The hierarchical structure of the cluster randomized design was reflected in the statistical equations used to estimate the impacts of *iRAISE*: individual observations were nested within teachers and schools. We used SAS PROC MIXED and PROC GLIMMIX (SAS Institute Inc., 2006) as the primary software tools.

The form of the statistical equation used to obtain the impact estimate included the outcome (e.g., ETS assessment) on the left hand side of the equation, and three kinds of variables on the right hand side: (1) a variable indicating whether the outcome was obtained from a teacher randomly assigned to *iRAISE* or control, (2) a set of covariates to increase the precisions of the program impact estimates, and (3) a series of terms representing deviations in performance of individual students, teachers, and matched pairs (i.e., the random fluctuations in the outcomes).[4]

Moderator analyses were used to assess if impacts varied across subgroups of students. For example, we consider whether the program is more effective for higher-performing students than lower-performing students. We estimate this *difference* between subgroups *in the difference* (between the program and control groups) in posttest performance, by including an interaction term in the statistical equation. This term multiplies together the variable that indicates whether the student is in the program group and the variable that indicates the subgroup. The coefficient for this term measures the difference between the subgroups in the impact of the program.

Mediation analyses were used to examine whether an impact of *iRAISE* on student achievement may have been facilitated through prior impact on the intermediate instructional practice outcomes. If an impact is demonstrated on the intermediate variable, and we can also establish an association between the intermediate variable and student achievement—independently of the effect of the program and other covariates—then the intermediate variable may be a mediator of the impact on achievement. The mediation analysis consisted of first estimating the impact of *iRAISE* on a teacher practice variable (the mediator), and then assessing the relationship between student performance on the ETS assessment and the mediator, while controlling for the effects of a series of covariates. Statistical equations like the ones described above for assessing impacts were used at each stage.

Student baseline achievement is an important covariate because it is central to increasing the precision of the estimates, and because it is potentially an informative moderating variable. We obtained pretests from the Michigan Educational Assessment Program Science and Reading tests in Michigan, and the Pennsylvania System of School Assessment Science and ELA scores in Pennsylvania. All pretest scores

[4] With teacher outcomes, random deviations were modeled only at the levels of the teacher and the matched pair.

were assessed at the student level, when students were in 8th grade. Because the pretests were obtained from two different states, we transformed them in order to put them on a common scale.[5]

A further description of the impact model and the approach to handling missing data is described in Appendix B.

[5] A z-transformation was applied by subtracting from each score the mean of performance for the control group and dividing this difference by the standard deviation of the control distribution. This was done separately by state. Each z-transformed score represents how far an individual score lies from mean control group performance in standard deviation units of the control distribution.

Results

IMPLEMENTATION OF *iRAISE*

This section addresses the following research questions.

1. To what extent is *iRAISE* implemented in a way that is consistent with the program model and underlying theory of action?
2. What are the contextual factors that support or hinder *iRAISE* implementation?

The findings related to these research questions provide context for assessing and understanding the measured impacts of *iRAISE* on student and teacher outcomes. This section uses descriptive statistics from *iRAISE* PD attendance records, log data, observations, and teacher survey data to provide context for assessing and understanding the measured impacts of *iRAISE* on student and teacher outcomes. Results from teacher survey questions are reported as percentages of survey respondents, while results from FOI analyses are given two ways: as percentages of the group of teachers randomized to *iRAISE*, and as percentages of the teachers who actively participated in the *iRAISE* program. We have also included comparisons between the *iRAISE* and control groups to give further context to *iRAISE* implementation.

Implementation of Core Program Components

As described in the methods section, FOI was measured for each of the core program components against teacher- and/or program-level thresholds. The core *iRAISE* components include: delivery of the *iRAISE* PD, teacher participation in professional learning activities, and the alignment of the PD with key characteristics. Table 10 provides the teacher and/or program-level thresholds for each indicator of the three components; these thresholds were set in advance by SLI to specify the amounts for delivery of and attendance at PD sessions that were needed to ensure adequate implementation. Component 1 measures whether SLI offered the PD sessions they intended, with separate indicators for the initial five-day Foundations training, the monthly Ignite sessions, the monthly PLC sessions, and the assignment of monthly asynchronous work, which includes notetaker assignments where teachers reflect and build on their recent learning, as well as message board posts, where they engage each other through online asynchronous discussion. The indicators here are measured through observation of the PD sessions and program log data and are applied at the program level. Component 2 measures whether teachers completed the activities related to these same four indicators and uses program log data and course gradebooks. Here, the indicators are measured at the teacher level and then aggregated to the program level. Component 3 measures the alignment of the PD to the five defining characteristics intended by SLI, one indicator for each. This component is measured at the program level using observations of the PD sessions.

Fidelity was met for components 1 and 3—the delivery and alignment of the PD by SLI—but not for component 2: teacher attendance at the PD. The following sections report separately on each indicator.

TABLE 10. FIDELITY MATRIX FOR THE *IRAISE* PROJECT

Key component	Operational definition	Source of information/ Schedule of data collection	Individual-level threshold	Sample-level threshold	Met fidelity?
Component 1: SLI Delivers Professional Development	Indicator 1: 5 days of PD are offered to teachers through online modules	Observations, program log data, and teacher surveys	Not applicable	0: < 5 days of PD offered to teachers 1 = 5 days offered to teachers	1
	Indicator 2: Delivery of monthly whole group synchronous Ignite meetings (2 hours each)	Observations, program log data, and teacher surveys	Not applicable	0 : < 95% of monthly meetings 1 = 95% or more of monthly meetings occur	1
	Indicator 3: Delivery of monthly small-group synchronous PLC meetings (1 hour each)	Observations, program log data, and teacher surveys	Not applicable	0 : < 95% of monthly meetings 1 : 95% or more of monthly meetings occur	1
	Indicator 4: SLI assigns monthly asynchronous activities	Observations, program log data, and teacher surveys	Not applicable	0: SLI assigns at least one asynchronous activity per month 1: SLI assigns one or more asynchronous activities per month	1
	Criteria for implementing Component 1 with fidelity			**Component score ranges from 0-4.** **Score of 0-3 = not with fidelity** **Score of 4 = with fidelity**	**4 = Met fidelity**

TABLE 10. FIDELITY MATRIX FOR THE *iRAISE* PROJECT

Key component	Operational definition	Source of information/ Schedule of data collection	Individual-level threshold	Sample-level threshold	Met fidelity?
Component 2: Teachers Attend Professional Development	Indicator 1: Participation in 5-day *iRAISE* synchronous Foundational training	Observations, program log data, and teacher surveys	Individual score ranges from 0-5, based on number of days teachers attended at least 80% of the session. (Example: 2 = Teacher participated in ≥ 80% of 2 sessions)	Sample-level score ranges from 0-5. (Examples: 2 = 80% or more teachers attend at least two days, 5 = 80% or more teachers attend all five days)	4 (80% of teachers attended at least 4 days)
	Indicator 2: Teachers participation in monthly whole group synchronous Ignite meetings	Observations, program log data, and teacher surveys	0: Teacher participated in < 5 monthly meetings 1: Teacher participated in **≥ 5 monthly meetings**	0: (0%**≤** teachers with a score of 1 <33%) 1: (33%**≤** teachers with a score of 1 < 67%) 2: (67% **≤** teachers with a score of 1<=100%)	1 (63% of teachers attended 5 or more meetings)
	Indicator 3: Teachers participation in once-monthly small-group synchronous PLC meetings	Observations, program log data, and teacher surveys	0: Teacher participated in < 75% of PLC meetings 1: Teacher participated in **≥75% of PLC meetings**	0: (0%**≤** teachers with a score of 1 <33%) 1: (33%**≤** teachers with a score of 1 < 67%) 2: (67% **≤** teachers with a score of 1<=100%)	1 (50% of teachers attended at least 7 PLC meetings)
	Indicator 4: Teachers complete asynchronous assignments	Program log data, access to 'Canvas' platform of work submitted	0: Teacher posted work for 0 – 4 meetings 1: Teacher posted work for 5-9 meetings	0: (0%**≤** teachers with a score of 1 <33%) 1: (33%**≤** teachers with a score of 1 < 67%) 2: (67% **≤** teachers with a score of 1<=100%)	1 (45% of teachers handed in at least 5 assignments)
	Criteria for implementing Component 2 with fidelity			**Component score ranges from 0-11. Score of < 9 = not with fidelity** **Score of ≥ 9 = with fidelity**	**7 = Does not meet fidelity**

TABLE 10. FIDELITY MATRIX FOR THE *IRAISE* PROJECT

Key component	Operational definition	Source of information/ Schedule of data collection	Individual-level threshold	Sample-level threshold	Met fidelity?
Component 3: Adherence of PD to the principles of Reading Apprenticeship	Indicator 1: Content of *iRAISE* PD is focused on science	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session	1
	Indicator 2 Teachers engaged in active learning	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session	1
	Indicator 3: *iRAISE* PD exhibited coherence	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session	1
	Indicator 4: Teachers engaged in metacognitive inquiry	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session	1
	Indicator 5: Collective participation	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session	1
	Criteria for implementing Component 3 with fidelity			**Component score ranges from 0 - 5** **0 = score of < 5 - not with fidelity** **1 = score of 5 - with fidelity**	**5 = Meets fidelity**

***iRAISE* Professional Development**

Key findings related to the *iRAISE* PD include the following.

- The *iRAISE* PD was delivered as intended: over 90% of the observed PD sessions exhibited the five key design characteristics.
- 32 out of the 33 teachers who agreed to participate in the PD attended at least four days of the Foundations training, with 26 of those 32 attending all five days. However, eight teachers randomized to the *iRAISE* group did not attend any of the PD: one teacher left his teaching position, two teachers declined to participate in the study shortly after randomization, and five teachers agreed to data collection but declined the PD because of other obligations.
- Over 80% of teachers (n = 27) who attended the *iRAISE* PD and responded to survey questions about the *iRAISE* PD after attending felt that it "moderately", "more than moderately", or "completely" prepared them to use the set of literacy practices modeled during the training.

Based on observations of the i*RAISE* Foundations sessions, the PD was delivered in a manner consistent with the theory of action. Ninety percent of the sessions observed (n = 35) exhibited content that was inquiry-based, and focused on disciplinary literacy, collective participation, active learning, and coherence.

Following the five days of the *iRAISE* Foundations PD, teachers were asked to rate their level of preparation on a set of key literacy strategies modeled during the PD.

A. Modeling/demonstrating metacognitive routines (e.g. Think Aloud, Talking to the Text)
B. Teaching students to analyze their own thinking about reading texts
C. Asking students to pose questions and problems about course readings
D. Supporting students in their attempts to understand disciplinary texts such as challenging literature, textbooks, primary documents, or scientific articles
E. Supporting students in working on reading or writing activities collaboratively by setting norms, creating safety, providing prompts that promote collaboration, and providing guidance and feedback on student participation
F. Facilitating students' active engagement in learning through the use of inquiry-based instructional methods
G. Providing students with opportunities for reading a variety of texts of different types and genres
H. Employing open-ended routines or assignments—such as group discussion or free choice in reading materials—enabling all students to feel comfortable participating and to measure their success
I. Structuring lessons that hold students accountable for reading, for example, so that students have to do the assigned reading in order to succeed

As shown in Figure 2, teachers felt most prepared to 1) model or demonstrate metacognitive routines and 2) teach students to analyze their own thinking about reading texts. The previous RAISE study also found that teachers felt relatively more prepared to model or demonstrate metacognitive routines compared to

other strategies (Fancsali et al., 2015). There, however, only 4% (n = 4) of teachers on average felt less than moderately or not at all prepared, while across the set of strategies in this study, between 6% and 18% (n = 2-6) of teachers felt less than moderately prepared or not at all prepared to implement.

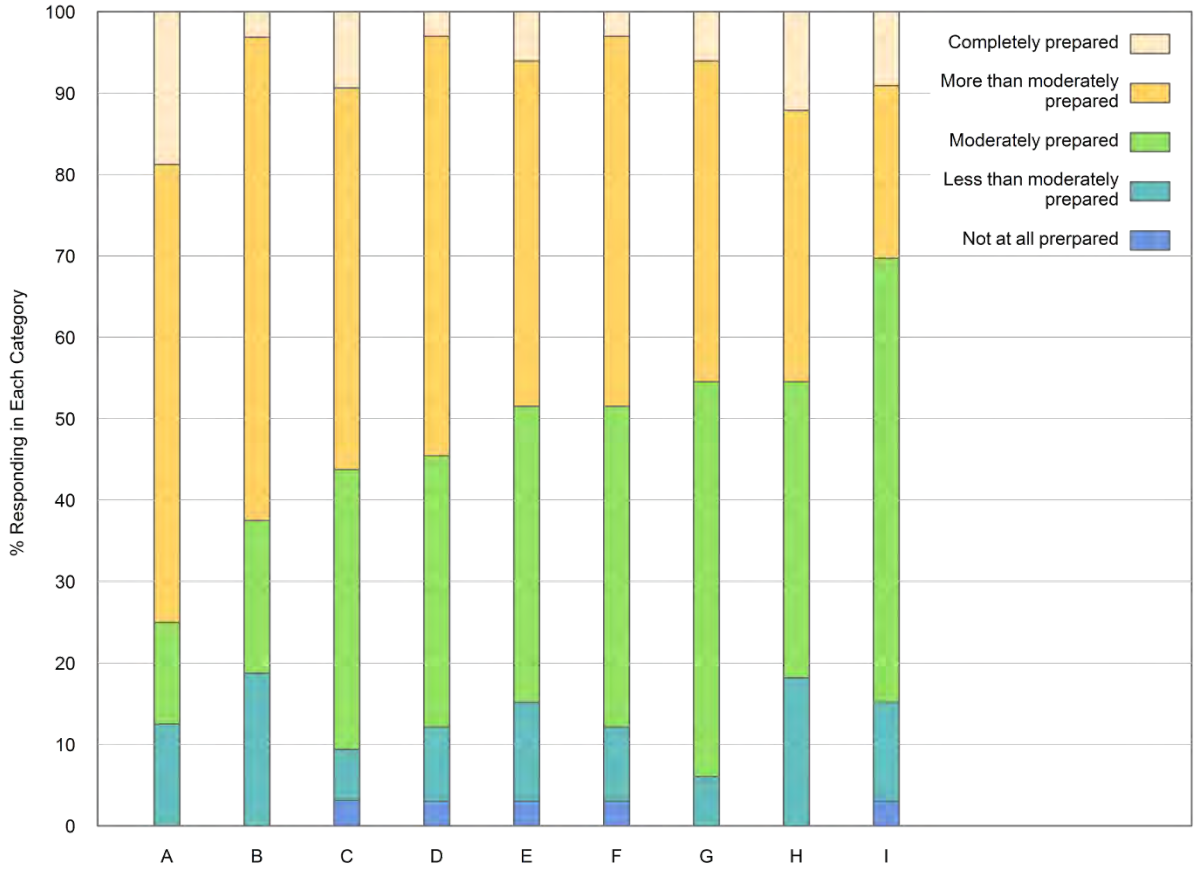

FIGURE 2. TEACHER REPORTED LEVEL OF PREPARATION AFTER *iRAISE* PD

Source. Empirical Education staff calculations based on teacher responses to study surveys

n = 32-33 for each strategy

iRAISE Ignite Sessions and PLCs

Key findings related to *iRAISE* Ignite sessions and PLCs include the following.

- Teachers averaged slightly over five meetings (both for Ignite sessions and for PLC) over the year. 25 out of 40 teachers (63%) met the teacher-level fidelity threshold for Ignite sessions, and 20 teachers (50%) met the teacher-level fidelity threshold for PLC meetings.
- While the program-level fidelity thresholds were not met for attendance at the *iRAISE* Ignite sessions, attendance varied greatly at the teacher level, with teachers who met fidelity averaging nearly eight meetings attended over the year, compared to an average of one meeting per year for teachers who did not meet fidelity. The fidelity threshold would be met for Ignite sessions if the sample was restricted to the 33 teachers actively participating in the program, but it would not be met for PLC sessions if restricted to the 33 teachers.

TABLE 11. FIDELITY OF IMPLEMENTATION AT THE TEACHER LEVEL

Event	No. of teachers meeting fidelity	% of randomized teachers meeting fidelity (*n* = 40)	% of participating teachers meeting fidelity (*n* = 33)
Foundations training (all 5 days)	26	65%	79%
Ignite sessions (at least 5)	25	63%	76%
PLC sessions (at least 7)	20	50%	61%
Monthly assignments (at least 5)	18	45%	55%

Source. Empirical Education staff calculations

The monthly Ignite sessions follow a similar structure: introduce new content with facilitators engaging and modeling with teachers as students, then dive deeper in small groups, and share back across the larger group. Overall, teachers averaged just over five Ignite sessions attended, which was the cut-point for the teacher-level fidelity threshold. If the teacher sample is restricted to the 33 teachers who actively agreed to participate in the PD, then program-level fidelity would be met for Indicator 2, with over 80% (*n* = 25) of those teachers meeting the teacher-level fidelity threshold. However, by including the eight teachers randomized to the *iRAISE* group who did not attend any of the PD, project-level fidelity was not met for this indicator. The *iRAISE* monthly PLC meetings were intended to be a key mechanism for support and collaboration among *iRAISE* teachers, allowing a more intimate, hands-on space for sharing classroom activities and experiences. Teachers were expected to attend at least seven (out of nine possible meetings, between September and May) *iRAISE* PLC meetings to meet the teacher-level fidelity threshold. To meet program-level fidelity, 67% of teachers had to have met the teacher-level threshold; 60% (*n* = 20) of teachers attended at least seven meetings, so program-level fidelity was not met. Again, among the 20 teachers who did meet the teacher-level fidelity threshold, teachers averaged eight PLC meetings attended. Those who did not meet the teacher-level fidelity threshold averaged only two meetings over the year. There are two possible hypotheses for the difference in attendance between Ignite and PLC sessions, with 80% and 60% of participating teachers meeting the respective thresholds: either overall, teachers found the Ignite sessions more valuable, or further variation in the commitment and buy-in between PLC groups caused lower attendance.

The most common reason selected for not attending monthly Ignite sessions or PLC meetings was other obligations. In open-response questions and interviews, teachers cited involvement in after-school activities (clubs, coaching, school leadership), further education (master's degree program, other PD), or family responsibilities as the main reasons for missing training. Of those teachers who attended, at least 80% (*n* = 21-24) reported that the monthly Ignite meetings were at least moderately helpful in every month except February (see Figure 3). On average, more teachers reported that the Ignite meetings were more than moderately helpful or very helpful compared to the PLC meetings, but nearly 90% (*n* = 19-26) of teachers reported that the PLC meetings were at least moderately helpful, while several of the Ignite

sessions were reported to be less than moderately helpful by between three and seven teachers (see Figure 3 and Figure 4).

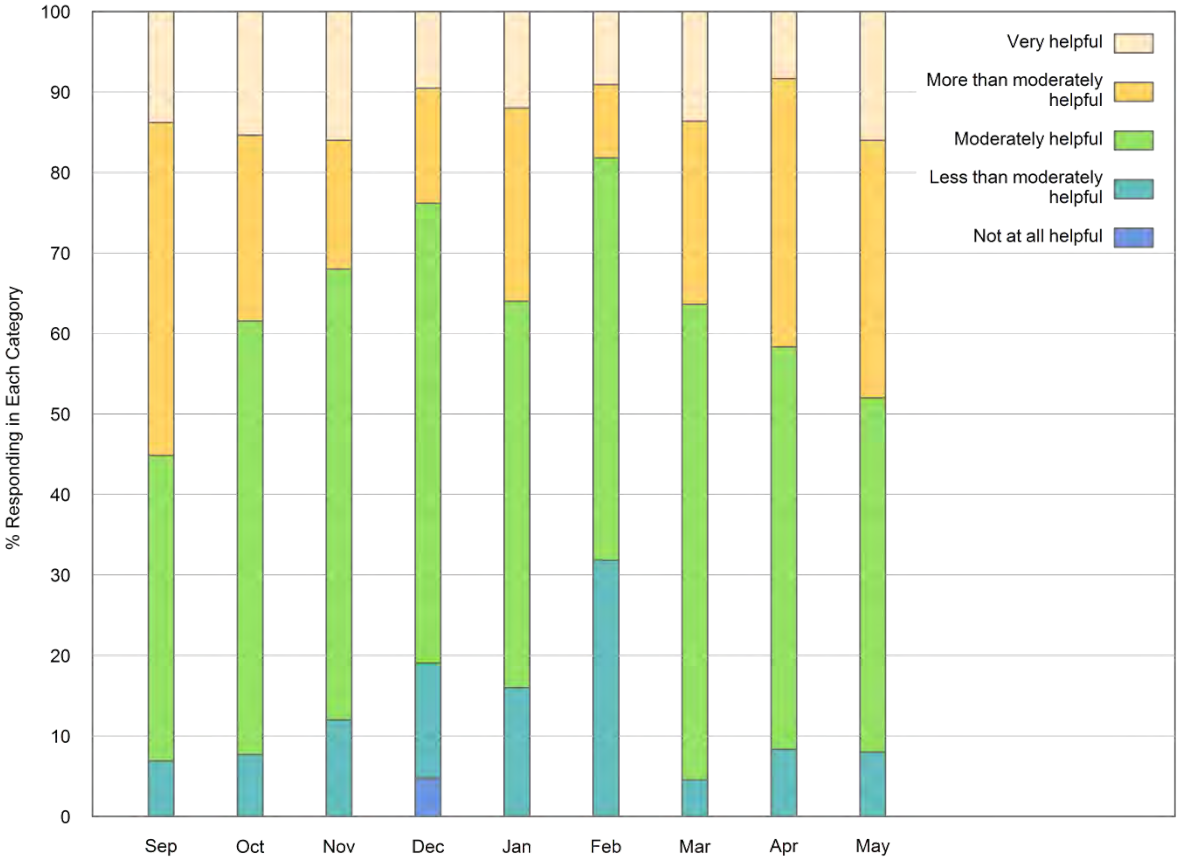

FIGURE 3. TEACHER REPORTED HELPFULNESS OF *IRAISE* MONTHLY IGNITE MEETINGS

Source. Empirical Education staff calculations based on teacher responses to study surveys

n = 21-29 for each month

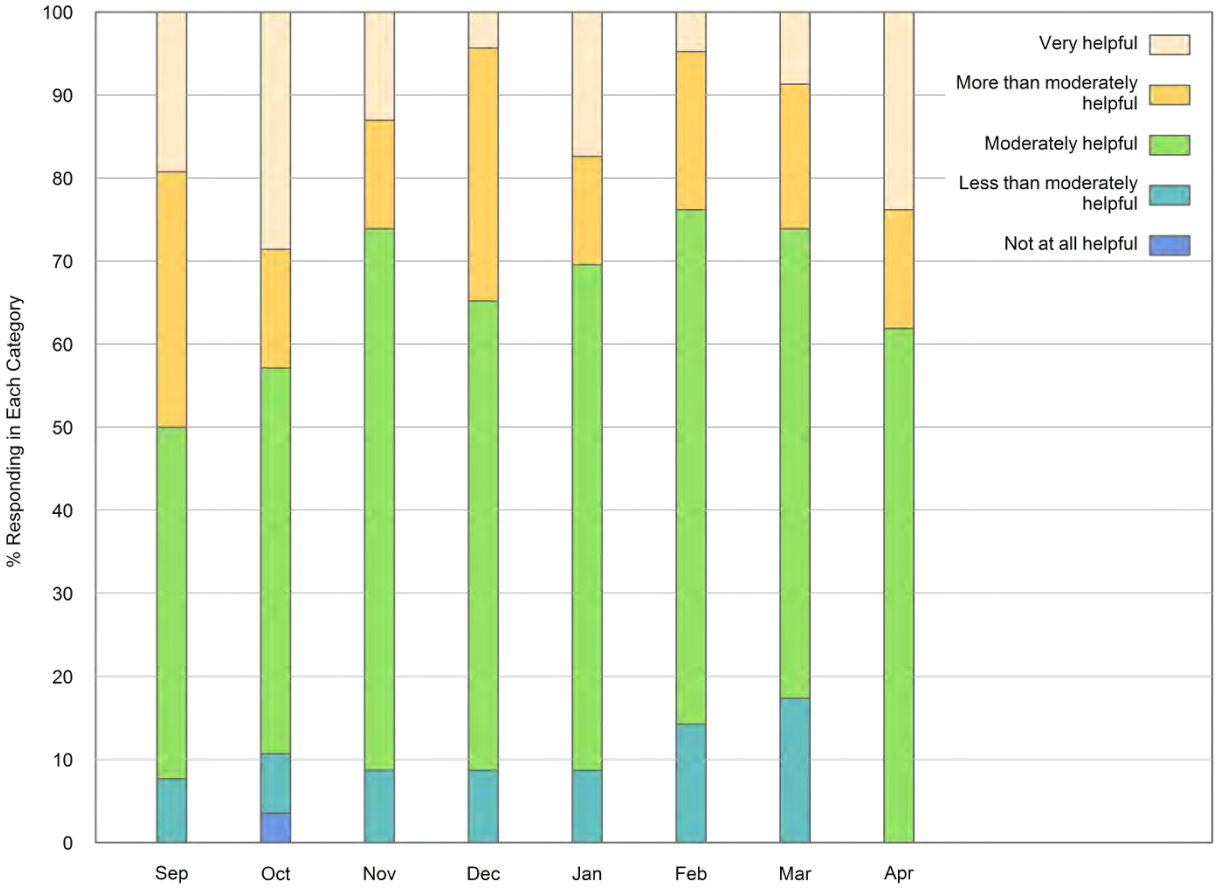

FIGURE 4. TEACHER REPORTED HELPFULNESS OF *iRAISE* PLC MEETINGS

Source. Empirical Education staff calculations based on teacher responses to study surveys

n = 21-29 for each month

The variation in implementation was also present within the monthly asynchronous work. Each month, in advance of the PLC sessions, teachers were expected to post their notetakers from the Ignite sessions, as well as create one post and respond once to a peer's post on the discussion board. We found that 18 teachers met the teacher-level fidelity threshold for Indicator 4 by completing all of the assigned work for at least 5 meetings. However, 26 teachers completed five or more of the nine monthly notetaker assignments, while 18 teachers completed five or more of the nine monthly discussion board postings. Teachers averaged six completed notetaker assignments and four completed discussion board postings.

Contextual Factors of Implementation

As shown in the *iRAISE* logic model (Figure 1), the program developers hypothesized that teachers would be supported or challenged in their implementation of *iRAISE* by contextual factors outside of the core components, such as policy or professional community. This section covers the types of support for literacy instruction that teachers reported receiving outside of *iRAISE* PD sessions and how helpful they

perceived this support to be. We have also included information on reported challenges and barriers to implementation, as well as overall impressions of *iRAISE*. Key findings related to the support and barriers to *iRAISE* implementation include the following.

- *iRAISE* teachers reported receiving support for literacy instruction (outside of official *iRAISE* PD and meetings) at a greater frequency than control teachers, and they rated this support as "very" or "more than moderately" helpful at higher levels than control teachers.
- The primary challenge to implementing Reading Apprenticeship was competing priorities, such as standardized test preparation or addressing content standards.
- 43% (n = 13) of teachers reported being fully committed to Reading Apprenticeship at the end of the year.

Support for Literacy Instruction

Four times during the year, all teachers were asked to indicate which (if any) types of support for implementing literacy instruction they received during the prior month. *iRAISE* teachers were explicitly instructed to exclude activities during monthly *iRAISE* meetings as a source of support. Teachers could select any of the following options: informal collaboration with other teachers, coaching and mentoring, model lessons, observation and feedback, resources, classroom management help, political support (for example, someone "backed them up" in a conflict over implementation of literacy instruction), a change in school or district policy that was relevant to literacy instruction, or "other." We looked at how frequently *iRAISE* and control teachers reported any of type of support across the year. *iRAISE* teachers reported receiving more frequent support for literacy instruction compared to control teachers, 52% to 25% ($p < .01$). Professional community is hypothesized in the logic model as a possible support for implementation; it is encouraging that *iRAISE* teachers reported more support for literacy instruction. Although not a focus of the implementation as it was in the previous RAISE study, *iRAISE* teachers reported interactions that fostered a community of literacy instruction. The most frequently reported types of support were informal collaboration with peers and materials/resources. The replication of the finding of the prior RAISE study —a school-level randomization that actively focused on building professional community—in this implementation is encouraging for developers of online PD.

Helpfulness of Support Received by Teachers

Teachers who reported receiving the above support were asked, in general, how helpful the support was for improving literacy instruction in their classroom. Teachers rated the support on a 5-point Likert scale. On average, *iRAISE* teachers were more likely to rate the support they received for literacy instruction (outside of the Ignite sessions or PLCs) as very helpful or more than moderately helpful compared to control teachers ($p < .01$).

Challenges to Reading Apprenticeship Implementation

Every three months, teachers were asked what challenges they faced in implementing Reading Apprenticeship. Competing priorities was the most commonly selected response, with two-thirds of teachers selecting it, on average. Many of the open-ended responses suggested that the pressures of

standardized tests created difficulty for teachers in implementing Reading Apprenticeship. The next most commonly selected responses were student behavior, lack of materials, and 'Reading Apprenticeship is too much work', selected by 36%, 36%, and 34% of teachers on average, respectively.

TABLE 12. RAISE TEACHERS REPORTING CHALLENGES IN IMPLEMENTING READING APPRENTICESHIP

	Oct (*n* = 32)	Jan (*n* = 30)	Apr (*n* = 30)
Competing priorities	62.5%	66.7%	66.7%
Lack of materials	34.4%	26.7%	46.7%
Student behavior	34.4%	36.7%	36.7%
Lack of understanding of how to implement Reading Apprenticeship	34.4%	33.3%	20.0%
Student ability	25.0%	26.7%	30.0%
Not enough training on Reading Apprenticeship	18.8%	16.7%	10.0%
Reading Apprenticeship is too much work	15.6%	36.7%	50.0%
Other	9.4%	36.7%	23.3%
Lack of parent support	9.4%	6.7%	10.0%
None	9.4%	3.3%	6.7%
Lack of administrative support	6.3%	6.7%	20.0%

Note. The three most frequently reported options in each month are shaded in blue.

Source. Empirical Education staff calculations

At the same time points, *iRAISE* teachers were asked whether or not there were any school district policy constraints that made implementing Reading Apprenticeship difficult. The responses remained fairly consistent across the school year, with just under 20% (*n* = 5-7) of teachers indicating that they believed district policy interfered with implementation of Reading Apprenticeship. In the RAISE study, which was a school-level randomization, roughly 10% of teachers reported this same interference. The teachers who reported facing district policy constraints were then given an opportunity to explain their answer, with most of these responses highlighting logistical challenges: teachers mentioned obstacles such as photocopying limits or pressure to cover material for tests/standards.

Alignment with Classroom Goals and Content Standards

In the last survey of the year, *iRAISE* teachers were asked to think back over their experience and determine how well Reading Apprenticeship aligned with the content standards and goals of their classroom (see Table 13). Overall, nearly 90% (*n* = 26) of teachers reported that Reading Apprenticeship was very well aligned or somewhat well aligned with their classroom standards and 96% felt similarly about its alignment with their classroom goals. The difference in the percentage of teachers reporting that Reading Apprenticeship aligns with their classroom standards, as compared to the percentage reporting alignment with classroom goals, echoes the above report of challenges to implementation; as teachers aspire to implement new strategies and try to shift their classroom practice, they may feel that they struggle to cover content standards.

TABLE 13. *IRAISE* TEACHERS REPORTING ALIGNMENT OF READING APPRENTICESHIP WITH CLASSROOM

	Not well aligned	Somewhat well aligned	Very well aligned
Classroom goals (*n* = 29)	3.4%	44.8%	51.7%
Classroom standards (*n* = 29)	10.3%	55.2%	34.5%

Source. Empirical Education staff calculations

At least 60% (*n* = 18) of teachers agreed with the statement that Reading Apprenticeship had improved their students' skills in each of the following five areas: subject-specific tests, reading comprehension skills, writing skills, completion of assigned reading, and engagement in class (See Figure 5). Between 18% and 36% of teachers reported in each category that their students were not at all improved as a result of Reading Apprenticeship.

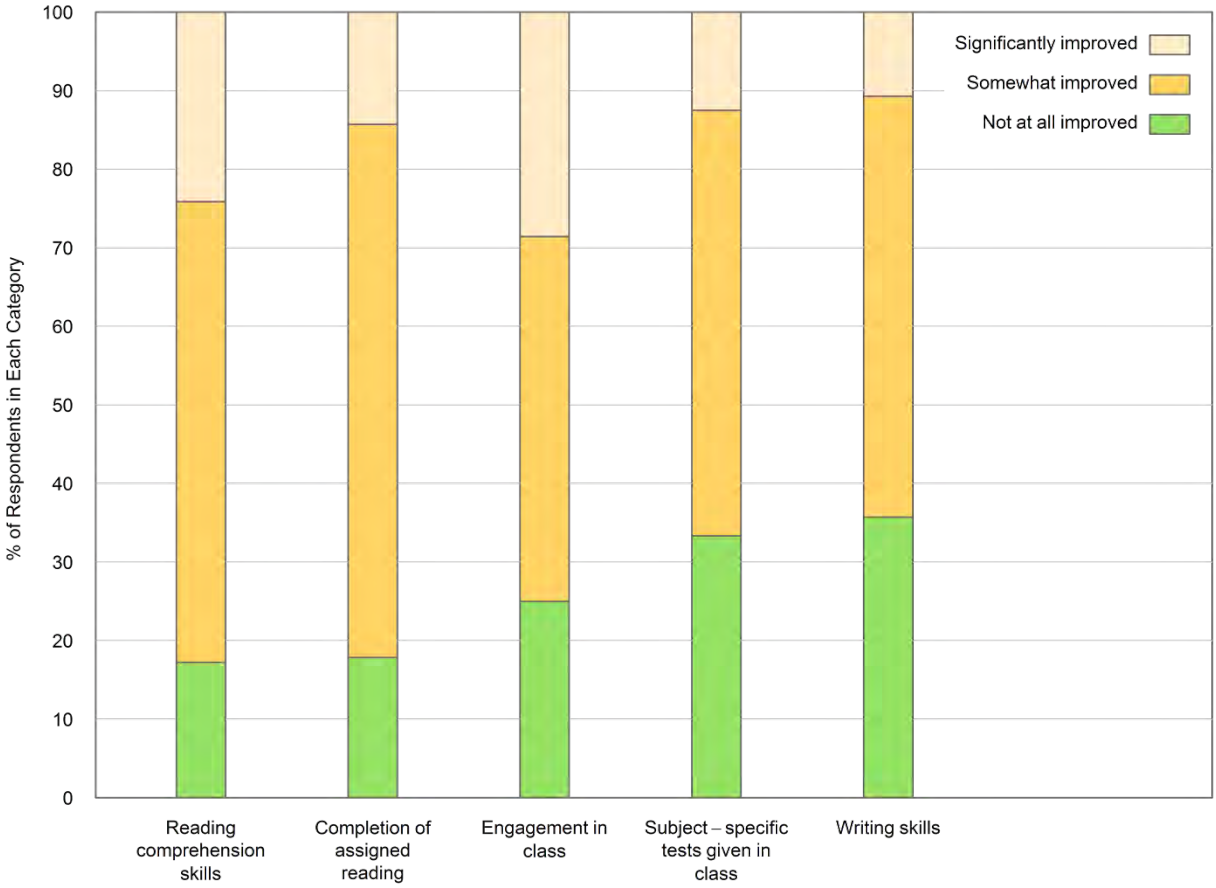

FIGURE 5. TEACHER REPORTED LEVEL OF STUDENT IMPROVEMENT

Source. Empirical Education staff calculations based on teacher responses to study surveys

n = 21-29 for each month

Commitment to Reading Apprenticeship and Overall Impressions

By the end of the year, 43% (n = 13) of *iRAISE* teachers reported being fully committed to Reading Apprenticeship work in their classroom, another 53% (n = 16) reported being willing to give it a try, and only one teacher reported that it was not a priority. Additionally, teachers were asked how well they understood the Reading Apprenticeship framework.[6] We found that 21% (n = 7) of *iRAISE* teachers reported that they "get" the Reading Apprenticeship model and use it often as they plan and reflect on their teaching, with another 56% (n = 19) reporting that it is starting to make more sense as they work to integrate it into their daily practice. These results are lower than the corresponding findings from the prior RAISE study using the same survey questions. While the findings of that study are taken from the end of the second year of implementation, we may hypothesize that the less positive levels of commitment to and understanding of Reading Apprenticeship may be due to the lower attendance at the PD sessions. Additionally, while the prior RAISE study was focused on building school-level teams, *iRAISE* is designed to allow for individual teachers within schools to participate and build networks with other teachers across schools and districts. This may have contributed to lower feelings of community, a theme that echoed in the teacher interviews - teachers reported that they saw the value of Reading Apprenticeship and how it could benefit students, but struggled with some implementation components during the school year. While some PLC groups gelled quickly, others wished they could be placed with teachers in their own school or district, who understood the context around their teaching or could comment specifically on the challenges of their subject area. Many teachers expressed the desire to plan lessons with their peers, believing that implementing what they were learning through *iRAISE* in their local network would be more useful. All of the teachers interviewed agreed that *iRAISE* would help teachers and students across science subjects, but several mentioned that biology teachers may have had a "leg-up", owing to the alignment of program materials.

While several teachers enjoyed the opportunity to implement their learning from the PD immediately as they went along and to learn from teachers of different backgrounds, some struggled to plan lessons or adjust schedules as they implemented new techniques. One teacher, who attended all required meetings despite reporting not enjoying the different PD components, said that the time commitment "soured her on the experience", adding that "she felt like a student, doing only the minimum to get by." Another teacher interviewee, who reported wholly positive experiences and began to see the implementation of Reading Apprenticeship "empower her students as readers", likened the course to "jumping into a pool and learning to swim at the same time." When first implementing, several teachers saw a tension between integrating Reading Apprenticeship into their instruction and covering their existing content standards. One teacher, who touched on this theme during his interview, took this thought a bit further. He felt pressure from school leadership to raise standardized test scores and thus felt "limited in [his] ability to experiment," but in fact realized at the end of the year that many standardized tests are now asking "thinking questions, questions that really ask students to do something critical," and that

[6] The full-text of the three response options: 1) "I get it and am referring to it often as I plan and reflect on my teaching"; 2) "It is starting to make more sense to me as I work with the approach to integrate it into my daily practice"; 3) "I understand some aspects of it, but I do not understand how it would translate into daily practice."

finding a balance between what he was learning through *iRAISE* and his traditional teaching would best serve his students. "What's more important than reading?" another teacher added. Those teachers who began to see changes in their students spoke about a feedback loop – when they saw students actually try something that they had been modeling, it encouraged them to continue implementing new strategies. One teacher, when asked about student motivation, offered this thought on his lower-performing students: "When you give them time to work and struggle with something, without expecting them to read quickly and digest it all right away, it gives them the idea that they can try. Sometimes that's all that it takes." Despite concerns about the time commitment to learn and integrate Reading Apprenticeship into their instruction, all of the teachers interviewed said they would continue to try using Reading Apprenticeship, expressing sentiments similar to one teacher who "[felt] ready to dive in deep next year with a fresh start."

IMPACT RESULTS

Overview

This section addresses the following research questions.

1. Is there a positive impact of *iRAISE* on classroom instructional practices, as measured through teacher surveys?
2. Is there a positive impact of *iRAISE* on general reading literacy outcomes, after one year, as measured through an ETS assessment of the construct?
3. Is there a differential impact of *iRAISE* on general reading literacy, after one year, depending on student English Language Learner (ELL) status, socioeconomic status (SES), prior achievement, or science subject area?
4. Are impacts of *iRAISE* on student general reading literacy, after one year, mediated through impacts on teacher literacy instructional practices?

This section also addresses several additional exploratory questions used to provide context for the results.

Impacts on Classroom Instructional Outcomes

In this section we address the impact of *iRAISE* on classroom instructional practices. Key findings include that *iRAISE* produced positive and statistically significant impacts[7] on 11 of the 12 instructional practices.

- We have a high level of confidence in there being positive impact on: variety of text types used (effect size = .393), teachers' fostering of student independence (effect size = .382), students

[7] As described in Appendix C, we interpret results based on p values as follows:

1. We have a high level of confidence when $p \leq .05$. (This is the level of confidence conventionally referred to as statistical significance.)
2. We have moderate/some confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.
4. We have no confidence when $p > .20$.

We categorize a result as statistically significant if the p value warrants limited or more confidence of there being a real effect.

practicing metacognitive inquiry (effect size = .457), teachers instructing comprehension strategies (effect size = .316), students practicing comprehension strategies (effect size = .516) and teacher self confidence in literacy instruction (effect size = .619).

- We have some confidence in there being a positive impact on teachers' use of traditional instructional strategies (effect size = .329), teachers instructing metacognitive inquiry (effect size = .236), teachers modeling metacognitive inquiry (effect size = .250), teachers modeling comprehension strategies (effect size = .243) and student collaboration (effect size = .285).

Table 14 provides a summary of the 12 constructs used in the analysis and the results for the comparison with the RAISE study.

TABLE 14. THE IMPACT OF *iRAISE* ON CLASSROOM INSTRUCTIONAL PRACTICES

Construct	Description	*iRAISE*			RAISE		
		N	Effect size	*p* value	*N*	Effect size	*p* value
1	Variety of Text Types Total number of text types that a teacher asked students to work with over a week, in or outside of class (e.g., newspapers, textbooks, historical documents)	67	.393	.033	206	0.04	.798
2	Fostering Student Independence Total number of minutes over a week that a teacher uses practices to foster independence, such as providing guided practice of reading comprehension strategies and having students teach other students	68	.382	.034	206	0.51	< .001
3	Traditional Instructional Strategies Total number of minutes over a week that a teacher employs traditional strategies, such as direct instruction and giving quizzes to assess comprehension	68	.329	.066	206	0.09	.562
4	Teachers Instructing Metacognitive Inquiry Total number of metacognitive inquiry strategies in which teachers provided instruction over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	52	.236	.095	206	- 0.09	.528
5	Teachers Modeling Metacognitive Inquiry Total number of metacognitive inquiry strategies that teachers modeled during their class over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	56	.250	.086	206	0.11	.422
6	Students Practicing Metacognitive Inquiry Total number of metacognitive inquiry strategies that students practiced during class over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	59	.457	.003	206	0.46	.001
7	Teachers Instructing Comprehension Strategies Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) in which teachers provided instruction over a week	59	.316	.040	206	- 0.06	.719

TABLE 14. THE IMPACT OF *iRAISE* ON CLASSROOM INSTRUCTIONAL PRACTICES

Construct	Description	*iRAISE*			RAISE		
		N	Effect size	*p* value	*N*	Effect size	*p* value
	Teachers Modeling Comprehension Strategies						
8	Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) that teachers modeled during the class over a week	61	.243	.100	206	0.23	.096
	Students Practicing Comprehension Strategies						
9	Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) that students practiced during class over a week	62	.516	.001	206	0.62	< .001
	Student Collaboration						
10	Total number of minutes over a week that teachers had students work on reading and writing activities in pairs, in small groups, and as a whole class	67	.285	.129	206	0.47	.008
	Student Engagement						
11	Total of teachers' ratings on the proportion of students in their class completing homework, paying attention in class, and participating in class activities	67	-.136	.487	206	0.05	.760
	Teacher Self-Confidence in Literacy Instruction						
12	Total of teachers' ratings on their confidence in their ability to provide literacy instruction, such as providing opportunities for reading a variety of texts of different genres and teaching students to analyze their own thinking about texts	65	.619	.004	206	0.41	.014

Note. Effect sizes are the effect estimates from the impact models divided by the respective pooled standard deviation of the outcome variable distribution.

Source. Empirical Education staff calculations

Looking across the two projects, we observe that several impacts of Reading Apprenticeship are replicated: we have high or some confidence on the same 6 dimensions in both projects. The correlation between the effect sizes for the twelve constructs across the two projects is a moderate .54. While the outcomes for the RAISE project was measured after the second year of implementation, the teacher impact findings here are based on survey responses reported in the course of the one-year implementation of *iRAISE.* To estimate the average impact, we averaged the responses in each condition across survey occasions. However, the trends in both conditions were not always parallel and linear. Sometimes they converged or diverged over time. To capture these trends, we report average responses in each condition across the surveys in Appendix G. We assessed also whether the impacts on the instructional variables converged or diverged and whether there was an overall (across conditions) downward trend in responses (see Table G2 in Appendix G.)

Impact on Students

ETS Literacy Assessment

In this section, we address the impact of *iRAISE* on performance on the ETS Assessment. We found no impact on reading literacy after one year, with an effect size of 0.002 (p = .96). Table 15 provides a summary of the samples used in the analysis and the results for the comparison of scores for students in *iRAISE* and control groups. The 'Unadjusted' row includes the raw means and standard deviations, as well as counts for students, teachers, and assignment pairs for the analytical sample. The last two columns provide the effect size, that is, the size of the difference between the means of the outcomes for *iRAISE* and control groups reported in standard deviation units and as a change in percentile ranking.[8] We also provide the p value, which indicates the probability of arriving at a difference with a magnitude as large as—or larger than—the magnitude of the one observed when there truly is no difference. The 'Adjusted' row is based on the same sample of students. The mean difference estimate—and therefore the effect size—is adjusted for the effects of a series of covariates, which means that the effects of chance differences between conditions on the covariates are factored out of the result. This adjustment also increases the precision of the program effect estimate by accounting for the effect of the covariates on the outcome variable. We see that the size of the difference—an effect size of 0.002—is small, and the p value of .96 gives us no confidence that there is a difference between *iRAISE* and control groups in the outcome.

[8] This metric tells where a student who performs at the 50th percentile of the distribution of posttest performance for the *iRAISE* group falls relative to the 50th percentile of the control distribution.

TABLE 15. EFFECT SIZES FOR STUDENTS WITH POSTTESTS

	Condition	Means	Standard deviations[a]	No. of students	No. of teachers	No. of assignment pairs	Effect size	*p* value	Change in percentile ranking
Unadjusted effect size[a]	**Control**	0.013	0.249	751	34	31	- 0.034	.79	- 1.34%
	iRAISE	0.009	0.242	717	35	34			
Adjusted effect size[b]	**Control**	0.013					0.002	.96	0.04%
	iRAISE	0.013							

[a] The unadjusted effect size includes as the numerator the impact estimate from a model without covariates (the *p* value corresponds to this effect.)

[b] The adjusted effect size includes the regression-adjusted benchmark impact estimate in the numerator (the *p* value corresponds to this effect.) Both effect sizes include the pooled standard deviation of the posttest scores in the denominator. Between-grade differences in the control posttest were factored out of the standard deviation in the denominator of the effect sizes. The *iRAISE* mean was obtained by adding the regression-adjusted estimate of the average one-year effect of *iRAISE* to the unadjusted control mean. The means under the adjusted effect size both show .013, due to rounding. The mean difference between them was .0005 units.

Source. Empirical Education staff calculations

Moderation of the Impact

Next, we addressed the question of whether the impact of *iRAISE* on the ETS assessment varies across student subgroups. We examined if the impact was different across levels of incoming achievement, English speaker status, and SES. We also considered the questions of whether impact varies across the science subject areas, specifically: physics, chemistry, biology and other subjects (which consisted largely of earth, environmental, and general science), as well as whether higher values of teacher-level implementation fidelity, as predicted from baseline characteristics, are associated with greater impact.

Key findings include the following.

- We have some confidence in there being a positive differential impact depending on levels of incoming reading achievement: students with lower reading pretests benefited more from the program.
- We have limited confidence in a differential impact depending on science subject, with greater impacts in biology and physics classes relative to earth and general science.
- No differential impacts were associated with English speaker status, SES, or with level of implementation fidelity.

Analyses of differential impact may have limited statistical power. This is the case with comparisons of impact across science subject areas, where we look at differences in impact across subgroups of the cases randomized--in this case, teachers. This is also the case for analysis of differences in impact for ELL students. There were only 13 students, within classes of seven teachers, designated as "Limited English Speakers", greatly limiting the sensitivity to detect differential impacts by level of English proficiency. On

the other hand, most teachers' classes in the experiment included students of high and low socioeconomic status, allowing a greater sensitivity to detect differential impact. The sample sizes for the subgroups across which we assess differential impact are summarized in Table 16.

TABLE 16. SAMPLE SIZES OF TEACHERS AND STUDENTS USED IN THE MODERATOR ANALYSES

Sample	Status	*iRAISE*	Control
English Proficiency	Not English Proficient	*J* = 5 *n* = 10	J = 2 *n* = 3
	English Proficient	*J* = 34 *n* = 700	J = 34 *n* = 746
FRPL	Not eligible for free or reduced price lunch	*J* = 31 *n* = 336	*J* = 33 *n* = 364
	Eligible for free or reduced price lunch	*J* = 34 *n* = 374	*J* = 34 *n* = 385
Subject	Biology	*J* = 17 *n* = 346	*J* = 11 *n* = 237
	Chemistry	*J* = 8 *n* = 179	*J* = 10 *n* = 215
	Physics	*J* = 7 *n* = 125	*J* = 7 *n* = 175
	Other	*J* = 3 *n* = 67	*J* = 6 *n* = 124
Prior science performance	Bottom 1/3 range of pre-test	*J* = 33 *n* = 272	*J* = 34 *n* = 277
	Middle 1/3 range of pre-test	*J* = 34 *n* = 352	*J* = 34 *n* = 374
	Top 1/3 range of pre-test	*J* = 25 *n* = 93	*J* = 26 *n* = 100
Prior reading performance	Bottom 1/3 range of pre-test	*J* = 34 *n* = 302	*J* = 34 *n* = 307
	Middle 1/3 range of pre-test	*J* = 34 *n* = 307	*J* = 34 *n* = 320
	Top 1/3 range of pre-test	*J* = 26 *n* = 108	*J* = 28 *n* = 124

Note. *J* is the number of teachers; *n* is the number of students. Student numbers may not sum to the analytic sample size of students used to assess overall impact, as some students have missing values for the moderating characteristics

Source. Empirical Education staff calculations

Including Pretests as a Moderator

We first show whether the impact of *iRAISE* varies for students at different levels of prior achievement. We had available both a science pretest and a reading pretest.[9] The tests were highly correlated with each other (r = .973, $p < .001$). We analyzed them separately.

The Science Pretest

The 'Fixed effects' in Table 17 provide the estimates of primary interest, including an estimate of the change in the impact of *iRAISE* for a 1-unit increase on the science pretest. At the bottom of the table, we give results for technical review—these consist of what are called random effects estimates.[10]

TABLE 17. MODERATING EFFECT OF THE SCIENCE PRETEST ON THE IMPACT OF *IRAISE* ON READING LITERACY ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the control student with a zero value for the pretest and with remaining covariates set to zero	- .040	.041	34	- .97	.338
Change in outcome for the control student for each one unit increase on the science pretest	.064	.004	1246	16.70	< .01
Effect of *iRAISE* for a student with an average science pretest	.006	.005	1246	1.32	.187
Change in the effect of *iRAISE* for each unit-increase on the science pretest	.- .004	.003	1246	- 1.38	.168
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Pair mean achievement	.0001	.0001		.97	.17
Teacher mean achievement	.0001	.0001		1.31	.09
Within-teacher variation	.0036	.0001		24.74	< .01

[a] We do not display the fixed effect estimates for covariates used to improve precision. The intercept value represents performance for cases with zero values for the covariates.

Source. Empirical Education staff calculations

The moderating effect of the science pretest score on the impact of *iRAISE*, that is, whether the program was differentially effective for students depending on their past science achievement, is shown in the fourth row. The *p* value of .168 indicates that we have limited confidence that the true differential impact

[10] Random effects are added to the statistical equation to account for dependencies in observed scores that happen because students come from the same teachers and because teachers are grouped in matched pairs.

is different from zero. That is, we have limited confidence that the impact of *iRAISE* changes depending on previous science achievement.

The Reading Pretest

The 'Fixed effects' in Table 18 provide the estimates of primary interest, including for the change in the impact of *iRAISE* for a 1-unit increase on the reading pretest.

TABLE 18. MODERATING EFFECT OF THE READING PRETEST ON THE IMPACT OF *IRAISE* ON READING LITERACY ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the control student with a zero value for the pretest and with remaining covariates set to zero	- .019	.039	34	- .49	.627
Change in outcome for the control student for each one unit increase on the reading pretest	.168	.004	1252	47.10	< .01
Effect of *iRAISE* for a student with an average pretest	.003	.004	1252	0.75	.453
Change in the effect of *iRAISE* for each unit-increase on the pretest	- .006	.003	1252	- 1.77	.076
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Pair mean achievement	.0001	.0001		1.29	.10
Teacher mean achievement	.0001	.0001		1.09	.14
Within-teacher variation	.0003	.0001		24.79	< .01

[a] We do not display the fixed effect estimates for covariates used to improve precision. The intercept value represents performance for cases with zero values for the covariates

Source. Empirical Education staff calculations

The moderating effect of the reading pretest score on the impact of *iRAISE*, that is, whether the program is differentially effective for students at different points along the pretest scale, is shown in the fourth row. The *p* value of .076 indicates that we have some confidence that the true differential impact is different from zero. That is, we have some confidence that the impact of *iRAISE* increases as the level of prior reading achievement decreases: the program has more benefit for students with lower incoming reading achievement.

Including ELL Status as a Moderator

We are also interested in the moderating effect of student English proficiency; that is, whether *iRAISE* is differentially effective for English proficient students compared to English learners. As noted earlier, only 13 students in the analysis sample were designated as Limited English Proficient, which severely limits the power of this test.

The 'Fixed effects' in Table 19 provide the estimates of primary interest. The estimate of the difference between English proficient and non-proficient students in the impact of *iRAISE* is shown in the fourth row. The coefficient is -.060. The *p* value of .311 indicates that we can have no confidence that the true differential impact is different from zero.

TABLE 19. THE MODERATING EFFECT OF ENGLISH PROFICIENCY ON THE IMPACT OF *iRAISE* ON READING LITERACY ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the Non-English proficient control with a zero value for the pretest and with remaining covariates set to zero	- .086	.074	34	-1.16	.253
Control group difference (English proficient minus not proficient) in the outcome	.081	.051	1395	1.58	.115
Effect of *iRAISE* for Non-English proficient student	.059	.059	1395	1.00	.316
Average difference (English proficient minus not proficient) in the effect of *iRAISE*	- .060	.059	1395	- 1.01	.311
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Pair mean achievement	.0003	.0003		1.07	.143
Teacher mean achievement	.0007	.0003		2.33	.010
Within-teacher variation	.0073	.0002		26.13	< .01

[a] We do not display the fixed effect estimates for covariates used to improve precision. The intercept value represents performance for cases with zero values for the covariates.

Source. Empirical Education staff calculations

Including Socioeconomic Status as a Moderator

We explored also the moderating effect of student SES; that is, whether *iRAISE* is differentially effective for students who qualify for free or reduced price lunch (lower SES) and those who do not (higher SES).

The 'Fixed effects' in Table 20 provide the estimates of primary interest, including an estimate of the difference between students at different levels of SES on the impact of *iRAISE* on reading literacy.

TABLE 20. THE MODERATING EFFECT OF SOCIOECONOMIC STATUS (SES) ON THE IMPACT OF *iRAISE* ON READING LITERACY ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the high-SES control with a zero value for the pretest and with zero values for the covariates	- .042	.059	34	- 0.71	.480
Control group difference (low-SES minus high-SES) in the outcome	.004	.007	1395	0.62	.534
Effect of *iRAISE* for high-SES student	.007	.010	1395	0.67	.503
Average difference (low-SES minus high-SES) in the effect of *iRAISE*	- .013	.010	1395	- 1.34	.180
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Pair mean achievement	.0003	.0003		1.13	.130
Teacher mean achievement	.0007	.0003		2.30	.011
Within-teacher variation	.0007	.0003		26.13	< .01

[a] We do not display the fixed effect estimates for covariates used to improve precision. The intercept value represents performance for cases with zero values for the covariates. The results in this table do not depend on the other fixed effects (not shown) in the model.

Source. Empirical Education staff calculations

We observe a small differential impact of *iRAISE* depending on level of SES with greater impact for higher-SES students. The coefficient of interest is - .013 in the fourth row. With a *p* value of .180, we have limited confidence that the true differential impact is different from zero.

Including Science Subject Area as a Moderator

While not a part of the original study plan, we consider also whether impact varies across the four science subjects in which *iRAISE* was implemented. This analysis involves making comparisons across mutually exclusive subgroups of the units randomized (teachers) which limits the statistical power to detect differential effects, if they exist. A test of whether there is a difference across the four subject categories yielded a *p* value of .17, giving us limited confidence of there being a differential impact. The results are summarized in Table 21 below. The impact for earth science[11] is shown in the 5th row with a point estimate of - 0.032, and a *p* value of .21, giving us no confidence in there being an impact for that subject area. The next three rows show the additional impact (i.e., in addition to the impact in earth science) associated with the other three subject areas expressed in the metric of the ETS assessment; they are: .050 (p = .103) for biology, .013 (p = .682) for chemistry, and .054 (p = .103) for physics. This gives us some confidence that the impact in biology and physics was greater than in earth science, but no confidence of a difference between chemistry and earth science in the impact of the program.

[11] This category is predominantly earth science but also includes students from environmental and general science classes.

TABLE 21. THE MODERATING EFFECT OF SUBJECT ON THE IMPACT OF *iRAISE* ON READING LITERACY ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for earth/general science in the control condition with zero values for the covariates.	- .024	.060	35	- 0.40	.688
Control group difference (biology minus earth/general science) in the outcome	- .007	.021	1400	- 0.35	.724
Control group difference (chemistry minus earth/general science) in the outcome	.023	.021	1400	1.08	.280
Control group difference (physics minus earth/general science) in the outcome	- .025	.022	1400	- 1.14	.253
Effect of *iRAISE* for earth/general science	- .032	.025	1400	- 1.26	.210
Average difference (biology minus earth/general science) in the effect of *iRAISE*	.050	.030	1400	1.63	.103
Average difference (chemistry minus earth/general science) in the effect of *iRAISE*	.013	.031	1400	.41	.682
Average difference (physics minus earth/general science) in the effect of *iRAISE*	.054	.033	1400	1.63	.103
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Pair mean achievement	.0003	.0003		1.05	.147
Teacher mean achievement	.0007	.0003		2.34	.010
Within-teacher variation	.0007	.0003		26.23	< .01

[a] We do not display the fixed effect estimates for covariates used to improve precision. The intercept value represents performance for cases with zero values for the covariates. The results in this table do not depend on the other fixed effects (not shown) in the model.

Source. Empirical Education staff calculations

Levels of Predicted Implementation as Moderators of Impact

A further question is whether impact on student achievement increases as the level of Fidelity of Implementation (FOI) increases. We described in an earlier section that while implementation of system-wide components by the developer was achieved, implementation involving teachers' participation was lower than the threshold set by the program developers. Also, there were individual teacher differences in attained levels of FOI. This motivates the question of whether impact increases as FOI increases.

The simplest way to analyze the relationship between teacher levels of FOI and student outcomes is to correlate FOI with student achievement among the *iRAISE* teachers only. This would give some indication of whether teachers' adherence to the developer-recommended guidelines for reaching fidelity thresholds is associated with greater achievement. However, a simple correlation of this sort has serious limitations,

because we cannot tell from it whether stronger implementation leads to higher achievement, or if some other factors, possibly characteristics of the teachers themselves, lead to both stronger implementation and higher achievement. To work around this problem of the potential confounding of strength of implementation with attributes of teachers; that is, to estimate the relationship between level of impact and FOI independently of other teacher-level factors that may influence achievement, we apply methods by Unlu et al. (2010). This exploratory analysis—which we describe more fully in Appendix D—uses teacher baseline characteristics to infer what the levels of FOI would have been for the control teachers, had they been randomly assigned to *iRAISE*. We then conduct a moderator analysis to see if the impact of *iRAISE* varies depending on the model-predicted levels of FOI. The results are summarized in Table 22.

Calculating a FOI Value for each *iRAISE* Teacher

Before applying the method described above, it is necessary to summarize FOI for each *iRAISE* teacher in terms of a single numeric index. We considered just the indicators used to calculate the FOI results in the implementation study that involved teachers active participation in the program (key Component 2 in the Fidelity Matrix). The formula for calculating the individual FOI level involves averaging values of the four indicators of fidelity (completion of asynchronous assignments and levels of participation in preliminary training, in Ignite sessions, and in PLCs). Their contributions are weighted to reflect the priorities set in the fidelity matrix, with attendance at initial training receiving the most weight. Specifically, we used the following formula:

$$FOI = (1/11) \times (\text{Indicator 1} + \text{Indicator2} + \text{Indicator3} + \text{Indicator4})$$

The formula provides a summary FOI score for each teacher in *iRAISE*. The formula has a direct correspondence to the approach used to calculate the level of fidelity for *Component 2: Teachers Attend Professional Development* (Table 22). Indicator 1 (Participation in 5-day *iRAISE* synchronous Foundational training) ranges between 0 and 5, while Indicators 2 (participation in Ignite sessions), 3 (participation in PLC meetings), and 4 (completion of asynchronous assignments) each range between 0 and 2. The sum of the indicator scores ranges between 0 and 11. In the formula above, we divide the sum by 11 so that the score ranges between 0 and 1, with 1 indicating maximum possible FOI and 0 indicating a complete lack of FOI.

Calculating Levels of FOI in Terms of Baseline Characteristics

After arriving at a value of FOI for each *iRAISE* teacher, we expressed the FOI scores in terms of teacher covariates from the baseline survey. That is, we used a statistical equation to relate attributes of *iRAISE* teachers, measured through the baseline survey, to their achieved levels of FOI. We then used this result to identify, for each teacher, what his/her expected fidelity of implementation level is, given his/her baseline characteristics. Importantly, this was done for teachers in both conditions, giving each a model-predicted level of fidelity of implementation (denoted FOI*).

Finally, we assessed whether FOI*—the model-determined fidelity level—moderates the impact of *iRAISE* on student achievement on the ETS assessment; that is, whether greater impact occurs as the level of FOI* increases. The results are displayed in Table 22.

TABLE 22. MODERATING EFFECT OF THE MODEL-DETERMINED LEVEL OF FIDELITY OF IMPLEMENTATION (FOI*) ON THE IMPACT OF *iRAISE* ON READING LITERACY ACHIEVEMENT

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Intercept: Outcome for the control student with zero values for covariates including FOI*.	- .085	.078	34	- 1.09	.282
Change in outcome for the control student for each 1 unit increase in FOI*	- .011	.025	1396	- .43	.669
Effect of *iRAISE* for a student with a teacher who has a zero score for FOI*	.021	.032	1396	0.65	.516
Change in the effect of *iRAISE* for each unit-increase in FOI*	- .018	.041	1396	- 0.45	.653
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Pair mean achievement	.0009	.0006		1.50	.07
Teacher mean achievement	.0012	.0005		2.27	.01
Within-teacher variation	.01165	.0004		26.20	< .01

[a] We do not display the fixed effect estimates for covariates used to improve precision. The intercept value represents performance for cases with zero values for the covariates. The results in this table do not depend on the other fixed effects (not shown) in the model.

Source. Empirical Education staff calculations

The moderating effect of the model-determined FOI score on the impact of *iRAISE*, that is, whether the program is differentially effective for students at different points along the fidelity of implementation scale, is shown in the fourth row. The *p* value of .653 gives us no confidence that the true differential impact is different from zero. In other words, higher levels of FOI* are not associated with greater impacts.

Teacher Mediating Outcomes

We investigated further the role of instructional processes as possible mediators of impact on student achievement. We found that *iRAISE* has a positive impact on 11 out of 12 instructional processes considered critical to the program intervention model. We have limited to no confidence of a positive association between any of the 12 instructional practices and student achievement.

iRAISE theory posits that impact on student achievement is at least partially mediated through prior impacts on 11 of the 12 instructional practices (excluding Construct 3, as described in the Data Collection and Sources section of this report). As described earlier, a mediator is an intermediate outcome, itself impacted by the program, that facilitates impact on a more-distal outcome; in this case, on student reading literacy. To understand the mediating role of the 12 instructional practices, we analyze two relationships involving each posited mediator variable. First, we examine the impact of *iRAISE* on the mediator. (These were discussed above with results displayed in *Table 14: The Impact of iRAISE on Classroom Instructional Practices*.) If there is no impact on the teacher practice, then it cannot induce subsequent impact on student

achievement; in other words, it is not a mediator. Second, we examine if the mediator is related to student achievement, independently of the treatment effect and after accounting for effects of baseline covariates. If there is an impact on a teacher practice, but the practice is not related to student achievement, then the teacher practice is not a mediator. The results of these analyses are displayed in Table 23. For an instructional practice to be considered a mediator of impact on achievement, there must be impact on the practice (these results are displayed under "Stage 1" in Table 23), and the practice must be related to student achievement (these results are displayed under "Stage 2").

TABLE 23. IMPACTS ON CLASSROOM INSTRUCTIONAL PRACTICES AND THEIR RELATIONSHIP TO READING LITERACY

Construct	Description	N	Stage 1 Effect size	p value	Effect associated with mediator	p value
	Variety of Text Types					
1	Total number of text types that a teacher asked students to work with over a week, in or outside of class (e.g., newspapers, textbooks, historical documents)	67	.393	.033	.008	.416
	Fostering Student Independence					
2	Total number of minutes over a week that a teacher uses practices to foster independence, such as providing guided practice of reading comprehension strategies and having students teach other students	68	.382	.034	.003	.310
	Traditional Instructional Strategies					
3	Total number of minutes over a week that a teacher employs traditional strategies, such as direct instruction and giving quizzes to assess comprehension	68	.329	.066	.001	.644
	Teachers Instructing Metacognitive Inquiry					
4	Total number of metacognitive inquiry strategies in which teachers provided instruction over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	52	.236	.095	-.002	.898
	Teachers Modeling Metacognitive Inquiry					
5	Total number of metacognitive inquiry strategies that teachers modeled during their class over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	56	.250	.086	-.018	.163
	Students Practicing Metacognitive Inquiry					
6	Total number of metacognitive inquiry strategies that students practiced during class over a week (e.g., asking questions about the text, writing to clarify understanding, discussing meaning of texts)	59	.457	.003	-.007	.605
	Teachers Instructing Comprehension Strategies					
7	Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) in which teachers provided instruction over a week	59	.316	.040	.014	.151

TABLE 23. IMPACTS ON CLASSROOM INSTRUCTIONAL PRACTICES AND THEIR RELATIONSHIP TO READING LITERACY

Construct	Description	N	Stage 1 Effect size	p value	Effect associated with mediator	p value
	Teachers Modeling Comprehension Strategies					
8	Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) that teachers modeled during the class over a week	61	.243	.100	.009	.379
	Students Practicing Comprehension Strategies					
9	Total number of comprehension strategies (e.g., setting a reading purpose, annotating text, choosing a reading approach that fits the purpose) that students practiced during class over a week	62	.516	.001	-.004	.615
	Student Collaboration					
10	Total number of minutes over a week that teachers had students work on reading and writing activities in pairs, in small groups, and as a whole class	67	.285	.129	-.002	.411
	Student Engagement					
11	Total of teachers' ratings on the proportion of students in their class completing homework, paying attention in class, and participating in class activities	67	-.136	.487	.003	.348
	Teacher Self-Confidence in Literacy Instruction					
12	Total of teachers' ratings on their confidence in their ability to provide literacy instruction, such as providing opportunities for reading a variety of texts of different genres and teaching students to analyze their own thinking about texts	65	.619	.004	.001	.562

Source. Empirical Education staff calculations

To help with overall interpretation, we summarized the results of both stages of the mediating processes in Table 24, showing just the direction of the effect (+, 0 or -) and with shading representing the level of confidence in the result being different from zero.

TABLE 24. SUMMARY OF POTENTIAL MEDIATING PROCESSES

		iRAISE	
Construct	**Description**	**Stage 1**	**Stage 2**
1	Variety of Text Types	+	+
2	Fostering Student Independence	+	+
3	Traditional Instructional Strategies	+	+
4	Teachers Instructing Metacognitive Inquiry	+	-
5	Teachers Modeling Metacognitive Inquiry	+	-
6	Students Practicing Metacognitive Inquiry	+	-
7	Teachers Instructing Comprehension Strategies	+	+
8	Teachers Modeling Comprehension Strategies	+	+
9	Students Practicing Comprehension Strategies	+	-
10	Student Collaboration	+	-
11	Student Engagement	+	+
12	Teacher Self-Confidence in Literacy Instruction	+	+

[a] We did not convert these estimates into effect sizes, given the outcome distributions were highly skewed, however, given the *p* values, we have no confidence in there being an impact on these practices.

Source. Empirical Education staff calculations

Legend: +, 0, or – represents the direction of the effect; shading represents the level of confidence in the result being real.

high = dark gray	some = mid-grey	limited = light grey	none = white

Under 'Stage 1', we found that *iRAISE* had a positive impact on several instructional processes considered critical to the program intervention model. Under 'Stage 2', we had limited confidence of a positive association between "teachers instructing comprehension strategies" and student achievement, and limited confidence of a negative association between "teachers modeling metacognitive inquiry" and student achievement. Only for the former do we observe significant positive effects along the two stages of the mediating processes: a positive impact of the program on the mediator and a positive association (albeit with limited confidence) between the mediator and student achievement controlling for the effects of other variables.

Discussion

OVERVIEW

This report presents the findings of a one-year RCT investigating the effectiveness of *iRAISE,* a fully online version of the developer's existing face-to-face Reading Apprenticeship PD. As a part of their i3 development grant, SLI created and piloted this approach in high school science classrooms (including biology, chemistry, physics, and earth/environmental science) during the 2014-2015 school year in 27 schools in Michigan and Pennsylvania. Our evaluation of the project consisted of a teacher-level randomized experiment including 82 teachers. Teachers participated in a 20-hour Foundations training in the summer (of the year before implementation), which was followed by monthly sessions of whole-group training in two hour Ignite sessions and small-group training in one hour PLC sessions, accompanied by monthly asynchronous assignments. Teachers reported on their classroom practice through monthly surveys, and student achievement was measured on a reading literacy assessment developed by ETS. The study was designed to answer the following research questions.

1. Is there a positive impact of *iRAISE* on classroom instructional practices, after one year, as measured by teacher surveys?
2. Is there a positive impact of *iRAISE* on general reading literacy outcomes, after one year, as measured through an ETS assessment of the construct?
3. Is there a differential impact of *iRAISE* on general reading literacy, after one year, depending on student English Language Learner (ELL) status, socioeconomic status (SES), prior achievement, or subject area?
4. Are impacts of *iRAISE* on student general reading literacy, after one year, mediated through impacts on teacher literacy instructional practices?

IMPACTS ON CLASSROOM INSTRUCTIONAL PRACTICES

We found significant impacts of *iRAISE* on eleven of twelve outcome measures developed from the surveys measuring classroom practice. These results were consistent with the prior RAISE research study. The results are an important replication of the previous findings, as they substantiate the success of SLI's development of a lower cost and more accessible online version of their training. With similar impacts and effect sizes to the previous RAISE study, the *iRAISE* program changed teacher classroom practices with online PD over the course of one school year.

IMPACTS ON STUDENTS

We found no impact of *iRAISE* on general reading literacy achievement, as measured by the ETS literacy assessment, compared to the control condition. There was a significant moderating effect of pretest on the ETS literacy assessment with increased impact observed for students with lower incoming reading achievement. The differential impact favoring lower achieving students may be considered consistent with the result of a study in 2008 (Kemple et al, 2008), which found an impact of Reading Apprenticeship on students who were two or more grade levels below average in reading. Exploratory analyses revealed

few substantial associations between the mediators and student achievement on the ETS assessment of reading literacy.

IMPLEMENTATION RESULTS

While the PD was delivered in a manner consistent with the program logic model, teachers did not meet the attendance standards expected by the developers over the course of the school year. The FOI results may explain the measured teacher impressions of the *iRAISE* program on the final monthly surveys: 43% of respondents reported feeling fully committed to Reading Apprenticeship at the end of implementation. Compared to the prior RAISE study, where teachers reported high levels of commitment in their second year of implementation, *iRAISE* teachers may have struggled to build the same type of ownership. Teacher interviewees reported that they would benefit more from *iRAISE* if they were implementing alongside other teachers from their school or district, enabling collaborative lesson planning and a deeper understanding on the context around their teaching.

CONCLUSION

After a one-year implementation with *iRAISE,* we do not find an overall effect of the program on student achievement. However, we do find that levels of incoming reading achievement moderate the impact of *iRAISE* on general reading literacy such that lower scoring students benefit more from the program. Additionally, we found a positive effect on several classroom instructional practice outcomes. The effect sizes for the 12 constructs had a moderate correlation (r=.54) between the *iRAISE* and RAISE studies, supporting replication of a common process and resulting effects. This finding supports the basic goal of *iRAISE*: to provide the same PD value in a less expensive, and (for some) more accessible modality.

Despite levels of implementation that did not meet the expectations of the program developers, teachers self-reported that they did change their classroom practice as a result of the *iRAISE* program, and impacts of *iRAISE* were greater for students who were performing at lower levels of incoming achievement. We may hypothesize that the pedagogical shift evinced by *iRAISE,* as measured by the teacher survey outcomes, is translated most efficiently to students who are struggling with literacy. Given that *iRAISE* had an impact on teacher practices in literacy instruction and increased benefits for low-achieving students—consistent with positive findings from prior studies—we may express confidence in the promise of low-cost, accessible, and high-quality online-only PD, addressing the needs of schools struggling to meet the demands of literacy for college and career readiness.

References

ACT, Inc. (2012). *Catching Up to College and Career Readiness*. Iowa City, IA: ACT, Inc.

Fancsali, C., Abe, Y., Pyatigorsky, M., Ortiz, L., Hunt, A., Chan, V., Saltares, E., Toby, M., Schellinger, A., & Jaciw, A. P. (2015). *The Impact of the Reading Apprenticeship Improving Secondary Education (RAISE) Project on Academic Literacy in High School: A Report of a Randomized Experiment in Pennsylvania and California Schools*. (Empirical Education Rep. No. Empirical_ RAISE-7019-FR1-O.2). Palo Alto, CA: Empirical Education Inc. Retrievable from http://empiricaleducation.com/pdfs/RAISEfr.pdf

Greenleaf, C. L., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., & Jones, B. (2011a). Integrating literacy and science in science: Teaching and learning impacts of Reading Apprenticeship professional development. *American Educational Research Journal*, *48*(3), 647–717.

Greenleaf, C.L., Hanson, T., Herman, J., Litman, C., Rosen, R., Schneider, S., & Silver, D. (2011b). *A Study of the Efficacy of Reading Apprenticeship Professional Development for High School History and Science Teaching and Learning*. Final report to Institute for Education Sciences, National Center for Education Research, Teacher Quality/Reading and Writing, Grant # R305M050031

Institute of Education Sciences (IES). (2010). *Request for applications: Reading for Understanding Research Initiative*. CFDA Number: 84.305F. Retrieved from http://ies.ed.gov/funding/pdf/2010_84305F.pdf

Jaciw, A. P., Newman, D., Lazarev, V., Lin, L., & Ma, B. (2016, February). *Does "What Works" Work for Me? Translating Causal Impact Findings from Multiple RCTs of a Program to Support Decision-Making*. Presented at the spring conference of The Society for Research on Educational Effectiveness.

Kemple, J., Corrin, W., Nelson, E., Salinger, T., Herrmann, S., & Drummond, K. (2008). *The Enhanced Reading Opportunities Study: Early Impact and Implementation Findings* (NCEE 2008-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36*, 249-277.

National Council of Chief State School Officers & National Governors Association (NCCSSO). (2010). *Common Core Standards for English Language Arts and Literacy in History/Social Studies & Science*. Retrieved from http://www.corestandards.org/the-standards

Ness, M. K. (2008). Supporting secondary readers: When teachers provide the "what," not the "how." *American Secondary Education*, *37*(1), 80–95.

Ness, M. K. (2009). Reading comprehension strategies in secondary content area classrooms: Teacher use of and attitudes towards reading comprehension instruction. *Reading Horizons*, *49*(2), 143–166.

Next Generation Science Standards (NGSS) Lead States. (2013). *Next Generation Science Standards: For States, by states*. Washington D.C.: National Academies Press.

New America Foundation. (n.d.). *Federal Education Budget Project.* Retrieved from http://febp.newamerica.net/

O'Reilly T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing Reading Comprehension Assessments for Reading Interventions: How a Theoretically Motivated Assessment Can Serve as an Outcome Measure. *Educational Psychology Review, 26*(3), 403-424.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group (cluster) randomized controlled trials.* (NCEE 2009-0049). Washington, DC: U.S. Department of Education.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition. Newbury Park, CA: Sage.

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2006). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software* (through version 1.56). Version 2.0, available at http://www.wtgrantfoundation.org/resources/overview/research_tools

SAS Institute. (2006). SAS/STAT Software: Changes and Enhancements through Release 9.1. Cary, NC: SAS Institute Inc.

Somers, M. A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). *The Enhanced Reading Opportunities Study Final Report: The Impact of Supplemental Literacy Courses for Struggling Ninth-Grade Readers*. NCEE 2010-4021. National Center for Education Evaluation and Regional Assistance.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2011). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software* (through version 2.0). Version 2.0, available at http://www.wtgrantfoundation.org/resources/overview/research_tools

Unlu, F., Bozzi L., Layzer C., Smith, A., Price, C., & Hurtig, R. (2010). *Linking implementation* fidelity *to impacts in an RCT: A matching approach.* Paper presented in symposium: Using Matching Methods to Analyze RCT Impacts on Program-Related Subgroups at the annual fall conference of APPAM, Boston, MA.

U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). (2013). *Reading Assessment, Data Explorer Tool*. Retrievable from http://nces.ed.gov/nationsreportcard/naepdata

U. S. Department of Education, Institute of Education Sciences, What Works Clearinghouse (WWC). (2013, March). *What Works Clearinghouse: Procedures and Standards Handbook (Version 3.0).* Retrieved from http://whatworks.ed.gov

Vaughn, S., Swanson, E. A., Roberts, G., Wanzek, J., Stillman-Spisak, S. J., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly*, *48*(1), 77–93.

Xu, Z., & Nichols, A. (2010). *New estimates of design parameters for clustered randomization studies: Findings from North Carolina and Florida* (Working Paper No. 43). National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Appendix A. Considerations for Statistical Power

How Large a Sample Do We Need?

We conducted a power analysis to determine the number of teachers that the experiment would need in order to say, with specific levels of confidence, that the program has an impact. This is an important part of experimental design, and here we walk through the factors considered.

How Small an Impact Do We Need?

The size of the sample required for a study depends on how small an effect we need to detect. Experiments require a larger sample to detect a smaller impact, other things being equal. It is important to know the smallest potential impact that would be considered educationally useful in the study's particular setting. As a hypothetical example, using percentile ranks as the measure of impact, we may predict that a program of this type can often move an average student 15 percentile points. As a practical matter for educators, however, an improvement as small as 10 percentile points may have value. The researcher may then set the smallest effect of interest to be 10 points or better. Thus, if the program makes less than a 10-point difference, the practical value will be no different from zero. It is necessary to decide in advance on this value as part of the power analysis because it determines the sample size. Conversely, if we had a fixed number of cases to work with, we would want to know how small an effect we could detect—the so-called "minimum detectable effect size" (MDES)—given the available sample. Whatever the MDES for a study, it remains possible that effects exist that are smaller than the MDES that we are unlikely to detect with the sample size available.

How Much Variation is there between Teachers?

When we randomize at the teacher level but the outcome of interest is a test score of students associated with those teachers, we pay special attention to the differences among teachers in student scores. The greater the variation in the teacher averages of student scores, the more teachers we need in the experiment to detect the impact of the program. This is because the extra variation among teachers adds noise to our measurement, which makes the effect of the program, the signal, harder to detect. A summary statistic that is important for the statistical power calculation is the intraclass correlation coefficient (ICC). In technical terms, it is the ratio of the variance among teachers in student scores to the total variation in those scores. A larger ICC means between-teacher differences in student posttest scores contribute more uncertainty to our program effect estimate. A larger sample of teachers is then needed to dampen the noise to acceptable levels. We assume a value of the ICC before the beginning of the study, when conducting the power analysis. (The ICC, like other parameters in the power calculation, reflects our best estimate of what the value is, largely based on compilations of results from other studies. It is not possible to get estimates of these parameters using data from the study at hand until after the study is over.) Certain design strategies are also applied to increase statistical power essentially by accounting for between-teacher differences that contribute to the ICC. For example, randomizing similar teachers within matched pairs removes between-pair differences that contribute noise to the estimate of program impact.

How Much Value Do We Gain From a Pretest and other Covariates?

In order to estimate effects of interest with additional precision, we make use of other variables likely to be associated with performance. These are called covariates because they co-vary with performance on the outcome measure. By including covariates in the analysis, we increase the precision of our effect estimates by accounting for some of the variation in the outcome; that is, by effectively dampening some of the noise so that the signal—the effect of *iRAISE*—becomes easier to detect. In technical terms, a covariate-adjusted analysis is called an Analysis of Covariance. In our experiments, a student's score on a pretest is almost always the covariate most closely associated with the outcome. Where possible, we adjust for the effect of the pretest. The proportion of variance in the outcome accounted for through modeling covariates is called the "Coefficient of Determination" or R-squared value.

How Much Confidence Do We Want to Have in our Results?

We want to be certain that we do not incorrectly conclude (1) that there is no impact when there is one (we want to avoid drawing false negative conclusions), and (2) that there is impact when there is not one (we want to avoid drawing false positive conclusions). Conventionally, researchers have given priority to avoiding false positive conclusions, requiring differences large enough that they would be seen 5% of the time in the absence of an effect before concluding that there is an effect, while at the same time, allowing a conclusion of no effect when in fact there is an effect 20% of the time. For the power analysis, we adhere to these criteria. However, our conclusions reached about the presence of an effect are expressed in terms of levels of confidence rather than as a yes-or-no declaration. As we described earlier in the report, we interpret results in terms of whether they give a lot, some, limited, or no confidence that there is a true impact.

Sample Size Calculation for This Experiment

Taking all the above factors into consideration, and with the number of teachers that were available for this study, we estimated that the smallest effect size that we can detect is an absolute difference of seven percentile points for the ETS literacy assessment for a student who performs at the median of the distribution. This effect size is what we would see if we took a student who performs at the 50th percentile of the distribution of posttest performance for the *iRAISE* group and found that student's score to be seven percentile points higher (i.e., at the 57th percentile) or seven percentile points lower (i.e., at the 43rd percentile) than the median score for the control distribution. We can also express this difference as a standardized effect size, which is the proportion of the standard deviation of posttest performance. In terms of that metric, the MDES for the ETS assessment is 0.19 for the analysis sample available to estimate impact. The sample size calculation was conducted using Optimal Design, a software program developed for this purpose (Spybrook, Raudenbush, Congdon, & Martinez, 2011). These calculations were done assuming an ICC of .15, a randomization level R-squared of .70 (which accounts for effects of both blocking and modeling covariates [Xu & Nichols, 2010]), a student-level R-squared of .50, and 25 students per teacher).

Appendix B. Details of the Approach to Estimating Impacts

Program Impact

The primary question for the experiment was whether, following the intervention, students in *iRAISE* classrooms had higher scores on the ETS literacy assessment than students in control classrooms. To answer this question, we analyzed outcomes for the randomized groups. The randomization resulted in two groups that at the outset are statistically equivalent. One receives *iRAISE* and the other one does not. As a result, the average difference between the randomized groups on the posttest is an accurate measure of the program effect plus random error.

We put our data for students, teachers, and classes into a system of statistical equations that allow us to obtain estimates of the effects of interest. The primary relationship of interest is the causal effect of *iRAISE* on achievement as measured by the ETS literacy assessment. We use SAS PROC MIXED and PROC GLIMMIX (SAS Institute Inc., 2006) as the primary software tools for these computations. The output of the analysis process consists of estimates of effects, as well as *p* values that tell us how much confidence we should have that the estimates are different from zero.

We can increase the precision of our effect estimates by accounting for the effects of covariates in the analysis. Therefore, our statistical equations included a series of covariates. We also had to account for the fact that students are clustered by teacher. We expect outcomes for students who are grouped together to be dependent as a result of shared experiences. We had to add this dependency to our equation in order to prevent artificially high confidence levels about the results. To do this, we modeled a teacher-level random effect as we describe further in the section below *Fixed and Random Effects*.

Handling Missing Data

To control for potential bias in the effect estimate arising from the covariates having missing values, we used a dummy variable method. With this approach, for each of the covariates that is included in the model, a dummy variable was created. This variable was assigned a value of one if the value of the variable was missing for a given student, and zero otherwise. The missing values from the original variable were replaced with zero. The dummy method yields effect estimates with less bias than the tolerance threshold set by the What Works Clearinghouse with levels of attrition such as those observed here (this finding is obtained through a simulation study described in Puma, Olsen, Bell, & Price, 2009). Specifically, the method fares no worse and, in some cases, performs better when compared to other standard approaches, including case deletion and non-stochastic and several stochastic regression imputation methods.

When student achievement outcomes (posttests) were missing, we used listwise deletion and simply dropped the observation from the analysis. This approach to handling missing data is one of several recommended by Puma et al. (2009). In their simulation work, they found that this method produced impact estimates with bias that was smaller than 0.05 standard deviations of the outcome measure (they considered bias in both the estimated impact and its associated standard error).

Potential Mediators

The objective of a mediation analysis is to examine whether an impact of the program on student achievement happens through prior impact on an intermediate outcome such as the use of one or more instructional practices. If an impact is demonstrated on the intermediate variable, and we can also establish an association between the intermediate variable and student achievement independent of the effect of the program, then the intermediate variable may be a mediator of the impact on achievement.[12] Because we are not randomly assigning cases to levels of the mediator variable, we leave open the possibility that the mediating variables we are examining are proxies for other variables that are the true mediators of the process, but that we have not observed. That is, we cannot be sure of the causal status of the mediator.

We assess mediation whether or not there is an overall impact on student achievement because the mediating path that we are investigating may be one of several, and their effects may cancel when combined, leading to zero overall effect. However, impact on a mediator is necessary (though not sufficient) for that that variable to play a mediating role in the impact of *iRAISE* on student achievement[13].

Fixed and Random Effects

The covariates in our equations measure either (1) fixed characteristics that take on a finite set of values (e.g., there are only two levels of gender) or (2) a set of characteristics that is assumed to have a distribution over a population and where we treat the values that we measure as though they were a random sample from that larger population. The former are called fixed effects; the latter, random effects. Random effects add uncertainty to our estimates because they account for sampling variation, or the changes we would observe in the outcomes if we re-sampled units from the same population. Fixed effects produce less uncertainty but also limit the extent to which we can generalize our results.

We usually treat the effects of units that were randomized as random effects, so that in the statistical equations, our estimates reflect the degree of uncertainty that comes if we were to draw a different sample of such units from the same population.[14] This allows us to argue for the generalizability of our findings from a sampling perspective. Treating the effects of units that were randomized as fixed forces us to use other arguments if our goal is to generalize.

---

[12] In technical terms, the estimate of a given mediated effect is the product of the effect of treatment on the mediator, times the effect of the mediator on the final response variable, normally student achievement, holding constant the treatment effect (Krull & MacKinnon, 2001). In a mediation model with a single mediator, this is equivalent to (or for multilevel models, approximate to) the difference between (1) the effect of treatment on the final outcome before adjusting for the effect of the mediator, and (2) the effect of treatment on the final outcome after adjusting for the effect of the mediator (Krull & MacKinnon, 2001).

[13] We offer two caveats for interpreting the results of moderator and mediator analyses in this report. First, we have not adjusted the results for multiple comparisons. With many results, some will reach statistical significance by chance alone. A multiple comparison adjustment is sometimes used, but given the exploratory nature of these analyses we do not apply this adjustment here. Second, the mediators are based on teacher self-report of activities, which may be less accurate than other more objective measures.

[14] Although we seldom randomly sample cases from a broader population, and in some situations we use the entire population of cases that is available, we believe that it is still correct to estimate sampling variation (i.e., model random effects). It is entirely conceivable that some part or the whole set of participants at a level end up being replaced by another group (for whatever reason) and it's fair to ask how much change in outcomes we can expect from this substitution.

Using random or fixed effects for participating units serves a second function: it allows us to more accurately represent the dependencies among cases that are clustered together, especially for the clusters randomly assigned to conditions. All the cases that belong to a cluster share an increment in the outcome—either positive or negative—that expresses the dependencies among them. An appropriate measure of uncertainty in our estimate of the program's effectiveness takes into consideration the relative levels of variation *within* and *between* the clusters randomized. All of our statistical equations include a student-level error term and a randomization-level error term. The variation in these terms reflect the differences we see (1) among students within clusters, and (2) across randomized clusters, that are not accounted for by all the other effects in our statistical equation.

The choice of terms for each statistical equation is not rigid but depends on the context and the importance of the factors for the question being addressed. The tables reporting the estimates resulting from the computation will provide an explanation of these choices in table notes where necessary for technical review.

Appendix C. Reporting the Results

When we run the computations on the data, we produce several results: among them are effect sizes, the estimates for fixed effects, and *p* values.

Effect Sizes

We translate the difference between program and control groups into a standardized effect size by dividing the average group difference by a measure of the variability in the outcome. This measure of variability is also called the standard deviation and can be thought of as the average distance of all the individual scores from the average score (more precisely, it is the square root of the average of squared distances). Dividing the difference by the standard deviation gives us a measure of the impact in units of standard deviation, rather than units of the scale used by the particular test. This standardized effect size allows us to compare the results we find with results from other studies that use different measurement scales. In studies involving student achievement, effect sizes as small as 0.1 (one tenth of a standard deviation) are sometimes found to be important educationally. We also report the effect size where we divide the average difference, adjusted for the effects of pretest score and other covariates, by the standard deviation. This is called the 'adjusted effect size'. This adjustment will often provide a more precise estimate of the impact.

Estimates

We provide estimates to approximate the actual effect size. Any experiment is limited to the small sample of students, teachers, and schools that represent a larger population in a real world (or hypothetical) setting. Essentially we are estimating the population value. When we report an estimate in a table, the value refers to the change in outcome for a one-unit increase in the associated variable. For example, since we code participation in the control group as 0, and participation in the program group as 1, the estimate is essentially the average difference in the outcome that we expect in going from the control to the program group while holding other variables constant.

***p* values**

The *p* value is very important, because it gives us a gauge of how confident we can be that the result we are seeing is not due simply to chance. Specifically, it tells us what the probability is that we would obtain a result with a magnitude as large as—or larger than—the magnitude of the one observed when in fact there is no effect. Roughly speaking, it tells us the risk of concluding that the program has had an effect when in fact it hasn't. This mistake is also known as a false-positive conclusion. Thus a *p* value of .1 gives us a 10% probability of drawing a false-positive conclusion if in fact there is no impact of the program. This is not to be confused with a common misconception that *p* values tell us the probability of our result being true.

We can also think of the *p* value as the level of confidence, or the level of belief we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting *p* values.

1. We have a high level of confidence when $p \leq .05$. (This is the level of confidence conventionally referred to as statistical significance.)
2. We have moderate confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.
4. We have no confidence when $p > .20$.

In reporting results with p values higher than conventional statistical significance, our goal is to inform the local decision makers with useful information and provide other researchers with data points that can be synthesized into more general evidence.

Appendix D. A Post-Experimental Method to Assessing Impact under Strong Implementation

The approach by Unlu et al. (2010) that we adopt here follows three steps (illustrated in Figure D1 below). Step (1) involves modeling the relationships between baseline covariates (BC) and observed levels of Fidelity of Implementation (FOI) in the *iRAISE* group. The goal is to identify baseline characteristics of teachers and classes that are predictive of FOI, and express FOI in terms of those baseline variables. Step (2) involves applying the modeled relationship between baseline characteristics and FOI from the first step to obtain *model-determined* levels of FOI for both treatment and control teachers. Step (3) involves assessing whether level of predicted FOI moderates the impact of the program on the ETS assessment. That is, it involves assessing whether impact of the program changes (in particular, whether it increases) as the model-determined levels of FOI increase; in other words, whether we can expect impact of *iRAISE* on performance on the ETS assessment to increase as teachers adhere more to the implementation requirements. Further information about the benefits of the method and its potential to produce accurate results can be found in Unlu et al. (2010).

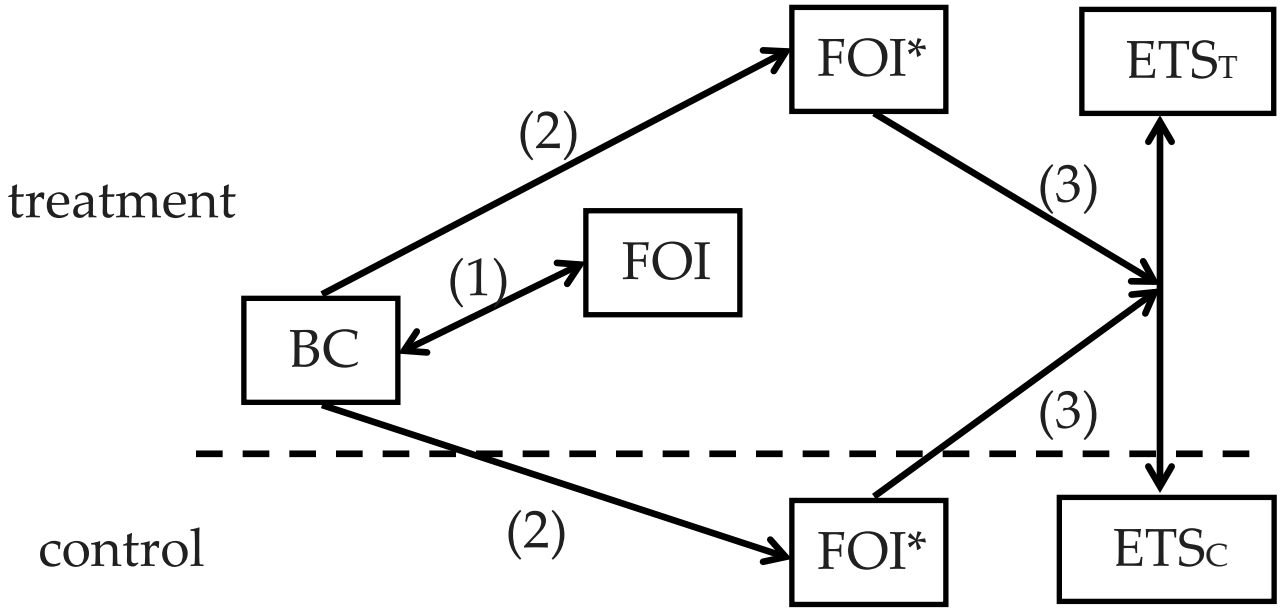

FIGURE D1. THE STEPS IN THE PROCESS OF ESTABLISHING WHETHER IMPACT INCREASES WITH INCREASING LEVELS OF MODEL-PREDICTED LEVELS OF IMPLEMENTATION

Appendix E. Fidelity of Implementation

TABLE E1. FIDELITY OF IMPLEMENTATION MATRIX

Key component	Operational definition	Source of information/ schedule of data collection	Individual-level threshold	Sample-level threshold
Component 1: SLI delivers PD	Indicator 1: 5 days of PD are offered to teachers through online modules	Observations, program log data, and teacher surveys	Not applicable	0: < 5 days of PD offered to teachers 1: 5 days offered to teachers
	Indicator 2: Delivery of monthly whole group synchronous Ignite meetings (2 hours each)	Observations, program log data, and teacher surveys	Not applicable	0: < 95% of monthly meetings 1: 95% or more of monthly meetings occur
	Indicator 3: Delivery of monthly small-group synchronous PLC meetings (1 hour each)	Observations, program log data, and teacher surveys	Not applicable	0: < 95% of monthly meetings 1: 95% or more of monthly meetings occur
	Indicator 4: SLI assigns monthly asynchronous activities	Observations, program log data, and teacher surveys	Not applicable	0: SLI assigns at least one asynchronous activity per month 1: SLI assigns one or more asynchronous activities per month
	Criteria for implementing Component 1 with fidelity			**Component score ranges from 0-4.** **Score of 0-3 = not with fidelity** **Score of 4 = with fidelity**

TABLE E1. FIDELITY OF IMPLEMENTATION MATRIX

Key component	Operational definition	Source of information/ schedule of data collection	Individual-level threshold	Sample-level threshold
Component 2: Teachers attend PD	Indicator 1: Participation in 5-day *iRAISE* synchronous Foundational training	Observations, program log data, and teacher surveys	Individual score ranges from 0-5, based on number of days teachers attended at least 80% of the session. (Example: 2 = Teacher participated in ≥ 80% of 2 sessions)	Sample-level score ranges from 0-5. (Examples: 2 = 80% or more teachers attend at least two days, 5 = 80% or more teachers attend all five days)
	Indicator 2: Teachers participation in monthly whole group synchronous Ignite meetings	Observations, program log data, and teacher surveys	0: Teacher participated in < 5 monthly meetings 1: Teacher participated in ≥ 5 monthly meetings	0: (0%≤ teachers with a score of 1 <33%) 1: (33%≤ teachers with a score of 1 < 67%) 2: (67% ≤ teachers with a score of 1<=100%)
	Indicator 3: Teachers participation in once-monthly small-group synchronous PLC meetings	Observations, program log data, and teacher surveys	0: Teacher participated in < 75% of PLC meetings 1: Teacher participated in ≥75% of PLC meetings	0: (0%≤ teachers with a score of 1 <33%) 1: (33%≤ teachers with a score of 1 < 67%) 2: (67% ≤ teachers with a score of 1<=100%)
	Indicator 4: Teachers complete asynchronous assignments	Program log data, access to 'Canvas' platform of work submitted	0: Teacher posted work for 0 – 4 meetings 1: Teacher posted work for 5-9 meetings	0: (0%≤ teachers with a score of 1 <33%) 1: (33%≤ teachers with a score of 1 < 67%) 2: (67% ≤ teachers with a score of 1<=100%)
	Criteria for implementing Component 2 with fidelity			**Component score ranges from 0-11. Score of < 9 = not with fidelity Score of ≥ 9 = with fidelity**

TABLE E1. FIDELITY OF IMPLEMENTATION MATRIX

Key component	Operational definition	Source of information/ schedule of data collection	Individual-level threshold	Sample-level threshold
Component 3: Adherence of PD to the principles of RA	Indicator 1: Content of *iRAISE* PD is focused on science	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session
	Indicator 2 Teachers engaged in active learning	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session
	Indicator 3: *iRAISE* PD exhibited coherence	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session
	Indicator 4: Teachers engaged in metacognitive inquiry	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session
	Indicator 5: Collective participation	Observations	0: indicator not observed during session 1: indicator observed during session	0: indicator observed in < 76% of sessions 1: indicator observed in ≥ 76% of session
	Criteria for implementing Component 3 with fidelity			**Component score ranges from 0 - 5** **0 = score of < 5 - not with fidelity** **1 = score of 5 - with fidelity**

Source. Empirical Education staff calculations

Appendix F. Teacher Survey Constructs

Table F1 presents the list of items comprising each instructional practice construct, as developed in the RAISE RCT (see Fancsali et al., 2015).

TABLE F1. FULL TEXT OF SURVEY QUESTIONS INCLUDED IN CONSTRUCTS

Construct Name	Survey Question
Variety of Text Types	Please select the types of texts that students in your target class worked with during the week (0 = No; 1 = Yes): • Newspaper/ magazine articles (including articles on-line) Textbook • Graphs/ charts/ images/ diagrams • Historical documents • Literature • Illustrations • Reference text • Lab procedures
Fostering Student Independence	Over the entire week, how many minutes did you spend using each of the following approaches to help your students understand text (# minutes): • Guided practice of reading comprehension strategies • Students teach other students • During the week, which of the following strategies did students learn to help them understand text (0 = No; 1 = Yes): • Discussing confusing parts of text- Teacher instructs • Discussing confusing parts of text- Teacher models • Discussing confusing parts of text- Student practice
Traditional Reading Strategies	Over the entire week, how many minutes did you spend using each of the following approaches to help your students understand text (# minutes): • Direct instruction (e.g. presentation, summary, background info on topic, mini-lecture) • Video • Quizzes • Asked oral questions about details of the text to check student understanding
Teachers Instructing Metacognitive Inquiry	During the week, which of the following strategies did students learn to help them understand text? Please indicated whether you provided instruction, modeled (i.e., presented an example of a behavior that students can emulate or learn from), and /or asked students to practice while you monitored progress (0= No; 1 = Yes): • Working in groups to discuss meaning of texts • Asking questions about the text • Writing to clarify understanding • Previewing long or challenging texts to identify strategies for dealing with them (Selected Teacher instructs)

TABLE F1. FULL TEXT OF SURVEY QUESTIONS INCLUDED IN CONSTRUCTS

Construct Name	Survey Question
Teachers Modeling Metacognitive Inquiry	During the week, which of the following strategies did students learn to help them understand text? Please indicated whether you provided instruction, modeled (i.e., presented an example of a behavior that students can emulate or learn from), and /or asked students to practice while you monitored progress (0 = No; 1 = Yes): • Working in groups to discuss meaning of texts • Asking questions about the text • Writing to clarify understanding • Previewing long or challenging texts to identify strategies for dealing with them
Students Practicing of Metacognitive Inquiry	During the week, which of the following strategies did students learn to help them understand text? Please indicated whether you provided instruction, modeled (i.e., presented an example of a behavior that students can emulate or learn from), and /or asked students to practice while you monitored progress (0 = No; 1 = Yes): • Working in groups to discuss meaning of texts • Asking questions about the text • Writing to clarify understanding • Previewing long or challenging texts to identify strategies for dealing with them
Teachers Instructing Comprehension Strategies	During the week, which of the following strategies did students learn to help them understand text? Please indicated whether you provided instruction, modeled (i.e., presented an example of a behavior that students can emulate or learn from), and /or asked students to practice while you monitored progress (0 = No; 1 = Yes): • Setting a reading purpose • Choosing a reading approach that fits the reading purpose • Visualizing what the author is describing or representing content in drawings • Making sense of graphs and other visuals • Predicting • Annotating text (e.g. making notes in the margins of text) • Re-reading • Taking on different roles to make sense of the text (e.g. presenter, note taker)

TABLE F1. FULL TEXT OF SURVEY QUESTIONS INCLUDED IN CONSTRUCTS

Construct Name	Survey Question
Teachers Modeling Comprehension Strategies	During the week, which of the following strategies did students learn to help them understand text? Please indicated whether you provided instruction, modeled (i.e., presented an example of a behavior that students can emulate or learn from), and /or asked students to practice while you monitored progress (0 = No; 1 = Yes): • Setting a reading purpose • Choosing a reading approach that fits the reading purpose • Visualizing what the author is describing or representing content in drawings • Making sense of graphs and other visuals • Predicting • Annotating text (e.g. making notes in the margins of text) • Re-reading • Taking on different roles to make sense of the text (e.g. presenter, note taker)
Students Practicing Comprehension Strategies	During the week, which of the following strategies did students learn to help them understand text? Please indicated whether you provided instruction, modeled (i.e., presented an example of a behavior that students can emulate or learn from), and /or asked students to practice while you monitored progress (0 = No; 1 = Yes): • Setting a reading purpose • Choosing a reading approach that fits the reading purpose • Visualizing what the author is describing or representing content in drawings • Making sense of graphs and other visuals • Predicting • Annotating text (e.g. making notes in the margins of text) • Re-reading • Taking on different roles to make sense of text (e.g. presenter, note taker)
Student Collaboration (Survey 2, 4, 6, 8 only)	During the week, how many minutes did your target class students spend working in class on reading activities and writing activities in the following situations (# minutes): • Reading in pairs • Reading in small groups • Writing in pairs • Writing in small groups • Writing as a class

TABLE F1. FULL TEXT OF SURVEY QUESTIONS INCLUDED IN CONSTRUCTS

Construct Name	Survey Question
Student Engagement	What portion of students in the target class did the following occur (1 = none; 2 = some; 3 = about half; 4 = most; 5 = Nearly all): • Completed their homework • Paid attention in class • Actively participated in class activities
Teacher Self-Confidence in Literacy Instruction (Survey 3, 8 only)	• Please rate your level of confidence in your ability to do the following (classroom instruction, 1 = very low; 2 = low; 3 = moderate; 4 = high; 5 = very high): • Provide opportunities for reading a variety of texts of different types/genres • Teach students to analyze their own thinking about texts • Structure lessons so that students have to do the assigned reading in order to be successful • Support students in their attempts to understand disciplinary text (e.g. challenging literature, textbooks, primary documents, scientific articles) • Provide explicit instruction around reading comprehension strategies (e.g. setting a reading purpose, previewing text, chunking, visualizing) • Model/demonstrate reading comprehension strategies (e.g. setting a reading purpose, previewing text, chunking, visualizing) • Support students in working on reading and writing activities in groups (small groups or whole class), (i.e. setting norms, creating safety, providing prompts that promote collaboration, and providing guidance/feedback) • Give students roles that make them responsible for making sense of texts (e.g. asking students to lead discussions or make arguments based on their interpretations of texts) • Facilitate students' active engagement in learning through the use of inquiry-based instructional methods (i.e., where students learn by questioning and problem-solving) • Ask students to pose questions and problems about course readings • Employ routines or assignments that are open-ended (e.g. group discussion; free choice in reading materials) so that all students feel comfortable participating and can have some measure of success

Appendix G. Teacher Survey Construct Trends

Figures G1 – G12 display responses over time to the 12 constructs measured through the teacher surveys. The sample sizes corresponding to analysis of each construct at each time point are displayed in Table G1. Table G2 describes the results of inferential tests of impacts and trends over time. For each of the constructs we address these questions.

- Is there a difference between conditions in responses averaged over time?
- Is there an overall upward or downward trend over time, with both conditions considered together?
- Is there a difference between conditions in the trends in response over time?

A summary response to the questions above follows.

- We observe a positive overall impact of *iRAISE* for 11 of 12 of the constructs.
- We observe a downward trend in responses across both conditions for 11 of the 12 constructs.
- We observe a difference between conditions in trends over time for 7 of the 12 constructs.

Concerning the third point, the trends show a steeper average linear decline in outcomes in the *iRAISE* condition; however, this has to be interpreted in terms of the relative performance of the two groups, especially during the initial phase of the experiment. The *iRAISE* group often starts higher, and maintains an advantage across the survey occasions. This allows more room for a steeper decline, which we observe in some cases.

What are the Trends of Classroom Instructional Practices Across the School Year?

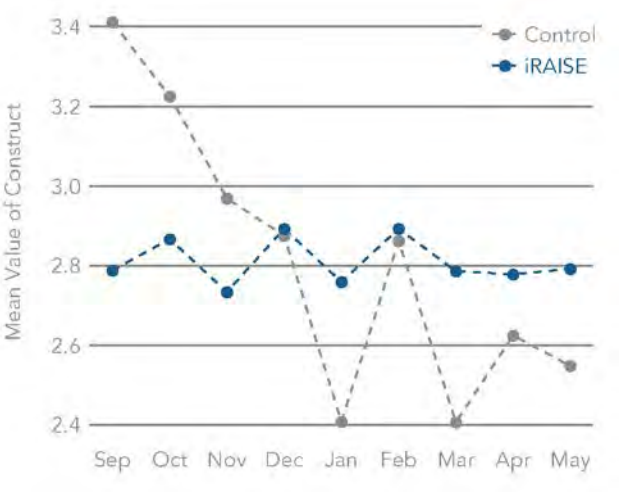

FIGURE G1. AVERAGE SCORES FOR CONSTRUCT 1

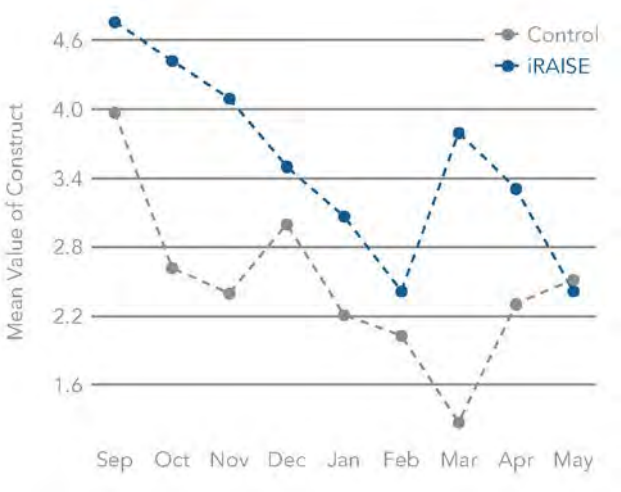

FIGURE G2. AVERAGE SCORES FOR CONSTRUCT 2

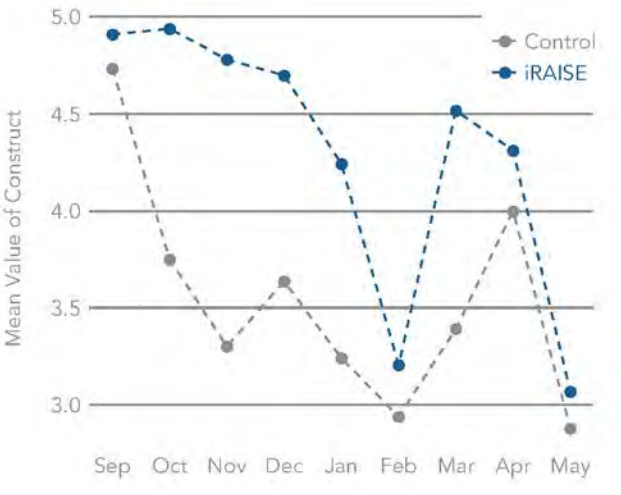

FIGURE G3. AVERAGE SCORES FOR CONSTRUCT 3

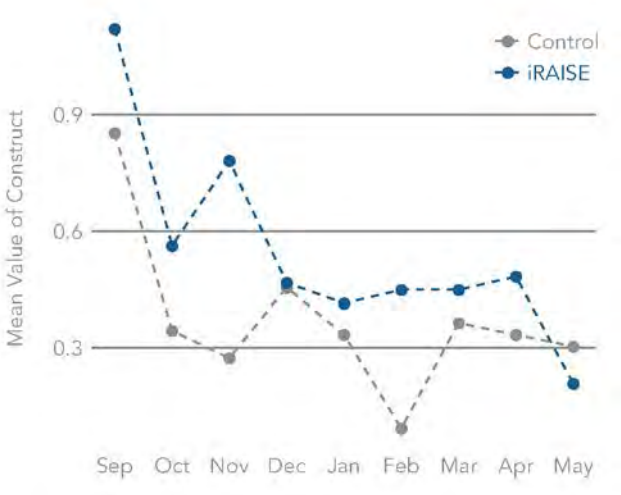

FIGURE G4. AVERAGE SCORES FOR CONSTRUCT 4

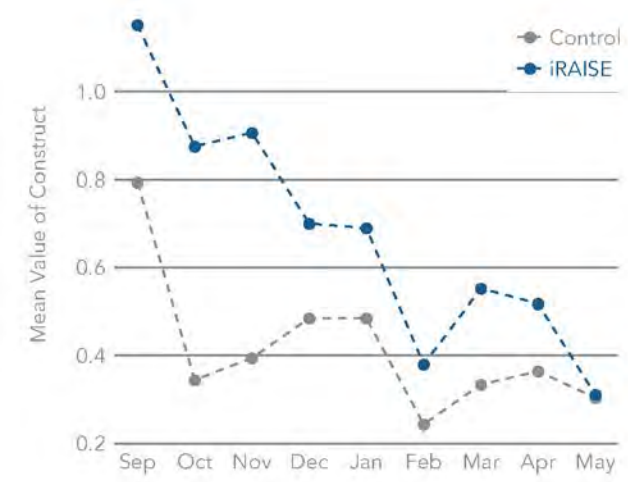

FIGURE G5. AVERAGE SCORES FOR CONSTRUCT 5

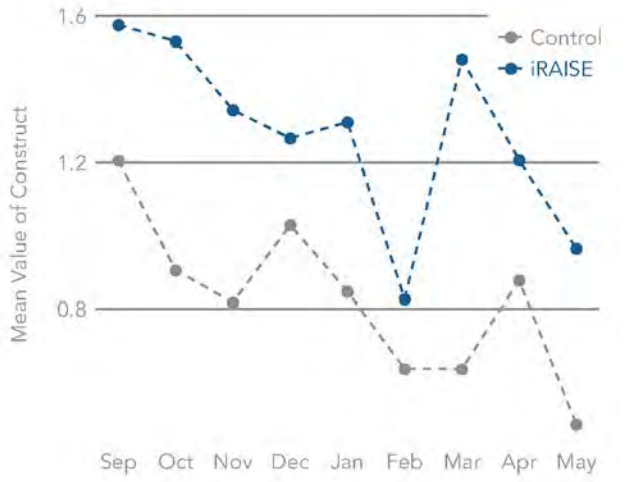

FIGURE G6. AVERAGE SCORES FOR CONSTRUCT 6

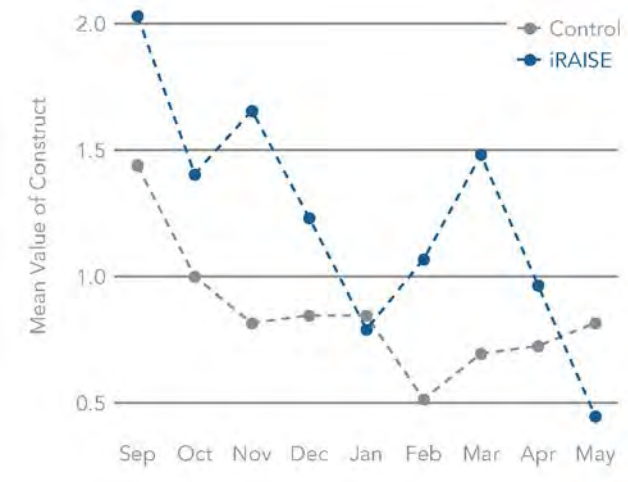

FIGURE G7. AVERAGE SCORES FOR CONSTRUCT 7

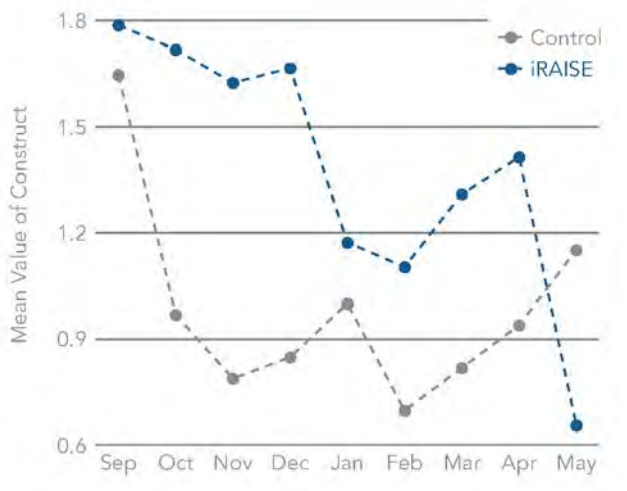

FIGURE G8. AVERAGE SCORES FOR CONSTRUCT 8

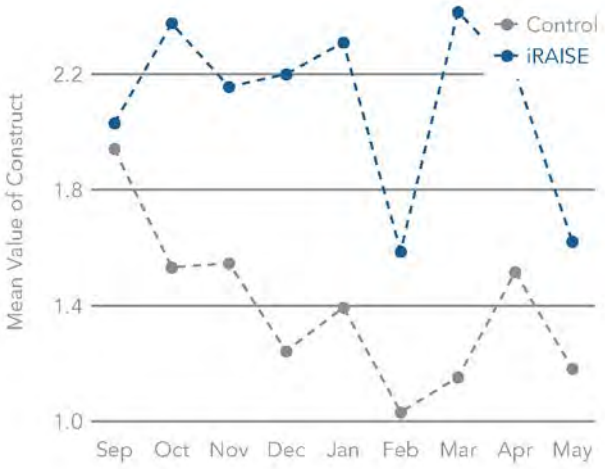

FIGURE G9. AVERAGE SCORES FOR CONSTRUCT 9

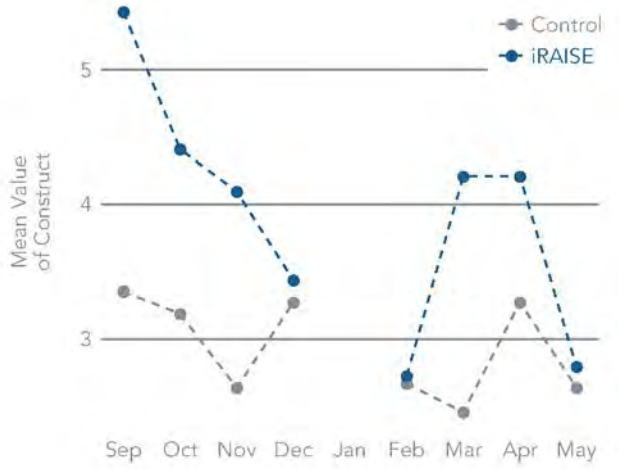

FIGURE G10. AVERAGE SCORES FOR CONSTRUCT 10

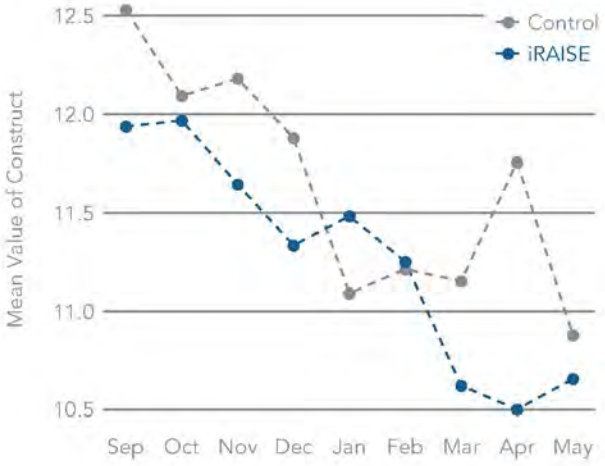

FIGURE G11. AVERAGE SCORES FOR CONSTRUCT 11

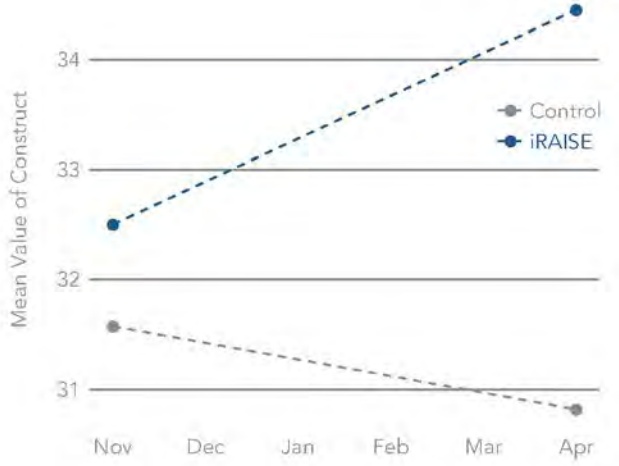

FIGURE G12. AVERAGE SCORES FOR CONSTRUCT 12

TABLE G1. SAMPLE SIZES USED FOR ANALYSIS PERTAINING TO TRENDS OVER TIME

Construct	Group	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
C1	*iRAISE*	34	31	33	32	27	29	32	32	31
	Control	33	30	30	28	29	28	28	27	24
C2	*iRAISE*	34	34	33	33	33	33	33	33	33
	Control	33	33	32	30	29	29	29	29	29
C3	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	33	32	32	30	29	29	29	29	29
C4	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	33	32	32	30	29	29	29	29	29
C5	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	33	32	32	30	29	29	29	29	29
C6	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	33	32	32	30	29	29	29	29	29
C7	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	33	32	32	30	29	29	29	29	29
C8	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	33	32	32	30	29	29	29	29	29
C9	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	33	32	32	30	29	29	29	29	29
C10	*iRAISE*	34	32	33	33	0	33	33	33	33
	Control	33	32	32	30	0	29	29	29	29
C11	*iRAISE*	34	32	33	33	33	33	33	33	33
	Control	32	31	31	30	29	28	29	28	29
C12	*iRAISE*	0	0	33	0	0	0	0	33	0
	Control	0	0	32	0	0	0	0	29	0

Source. Empirical Education staff calculations

TABLE G2. RESULTS OF INFERENTIAL TESTS OF IMPACTS AND TRENDS OVER TIME

Construct	Question 1: Is there a difference between conditions in responses averaged over time?	Question 2: Is there an overall upward or downward trend over time, considering responses in both conditions?	Question 3: Is there a difference between conditions in the trends in response over time?
1. Variety of Text Types	Yes: *iRAISE* is higher (High confidence)	Yes: downward (High confidence)	Yes: (High confidence) Steeper average linear decline for controls
2. Fostering Student Independence	Yes: *iRAISE* is higher (High confidence)	Yes: downward (High confidence)	No
3. Traditional Reading Strategies	Yes: *iRAISE* is higher (Some confidence)	Yes: downward (High confidence)	No
4. Teachers Instructing Metacognitive Inquiry	Yes: *iRAISE* is higher (Some confidence)	Yes: downward (High confidence)	Yes: (Some confidence) Steeper average linear decline for treatment
5. Teachers modeling Metacognitive Inquiry	Yes: *iRAISE* is higher (Some confidence)	Yes: downward (Some confidence)	Yes (High confidence) Steeper average linear decline for treatment
6. Students practicing Metacognitive Inquiry	Yes: *iRAISE* is higher (High confidence)	Yes: downward (High confidence)	No
7. Teachers Instructing Reading Comprehension	Yes: *iRAISE* is higher (High confidence)	Yes: downward (High confidence)	Yes (High confidence) Steeper average linear decline for treatment
8. Teachers modeling Reading Comprehension	Yes: *iRAISE* is higher (Some confidence)	Yes: downward (High confidence)	Yes (High confidence) Steeper average linear decline for treatment

TABLE G2. RESULTS OF INFERENTIAL TESTS OF IMPACTS AND TRENDS OVER TIME

Construct	Question 1: Is there a difference between conditions in responses averaged over time?	Question 2: Is there an overall upward or downward trend over time, considering responses in both conditions?	Question 3: Is there a difference between conditions in the trends in response over time?
9. Students practicing Reading Comprehension	Yes: *iRAISE* is higher (High confidence)	Yes: downward (High confidence)	No
10. Student Collaboration	Yes: *iRAISE* is higher (Limited confidence)	Yes: downward (High confidence)	Yes: (Some confidence) Steeper average linear decline for treatment
11. Student Engagement	No	Yes: downward (High confidence)	No
12. Teacher confidence in literacy instruction	Yes: *iRAISE* is higher (High confidence)	No	Yes (High confidence) Steeper average linear positive incline for treatment

Note. We do not report *p* values for these results. Instead, we report levels of confidence that estimated values do not reflect just chance differences. Levels of confidence are determined as follows.

A high level of confidence $p <= .05$

Some confidence: $.05 < p <= .15$

Limited confidence: $.15 < p <= .20$

The *p* values correspond to probabilities of observing estimated effects with magnitudes as large or larger than those observed under the following null hypotheses.

1) Question 1: there is no average impact.
2) Question 2: the linear trend over time is flat.
3) Question 3: the time trends have the same angle in both conditions.

Source. Empirical Education staff calculations