Comparative Effectiveness of WestEd's *iRAISE* Professional Development

AN INTERIM REPORT OF A RANDOMIZED EXPERIMENT IN MICHIGAN AND PENNSYLVANIA

1/31/2015

Adam M. Schellinger

Megan Toby

Andrew P. Jaciw

Jenna Zacamy

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 research company that provides tools and services to help K-12 school systems make evidence-based decisions about the effectiveness of their programs, policies, and personnel. The company brings its expertise in research, data analysis, engineering, and project management to customers that include the U.S. Department of Education, educational publishers, foundations, leading research organizations, and state and local education agencies.

Reference for this report: Schellinger, A.M., Toby, M., Jaciw, A.P., & Zacamy, J. (2015). *Comparative Effectiveness of WestEd's iRAISE Professional Development: An Interim Report of a Randomized Experiment in Michigan and Pennsylvania*. (Empirical Education Rep. No. Empirical_iRAISE-7027-IR1-Y1-O.1). Palo Alto, CA: Empirical Education Inc.

Table of Contents

Introduction

Empirical Education Inc. is the independent evaluator of WestEd's 2012 i3 Development grant for Internet-based Reading Apprenticeship Improving Science Education (iRAISE). We submit this interim report to provide an update on the progress of the randomized control trial (RCT) during the 2014-2015 school year. The RCT will measure the impact of iRAISE on 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. This report provides a summary of the 2013-2014 pilot study and its relationship to the RCT, the design summary and research questions for the RCT, an updated study timeline, preliminary results for fidelity of implementation (FOI), and initial teacher impressions from the monthly teacher surveys.

iRAISE brings SLI's 65-hour face-to-face literacy professional development (PD) to an online format. It 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 extensive existing materials, protocols, and key design elements of face-to-face Reading Apprenticeship (RA) PD and leverages interactive, internet-based technologies to enhance teachers' learning. iRAISE PD begins with a 5-day (approximately 20-hour) iRAISE Foundations training during the summer prior to implementation. After the start of the school year, teachers participate in monthly follow-up meetings from September through May. The monthly follow-up meetings provide three hours of additional support per month in two different formats: whole-group meetings introducing new learning and small-group meetings intended to produce discussion and collaboration. The fundamental goals of the intervention are (a) to influence teacher instruction, including the ability to integrate disciplinary literacy practices and explicit literacy instruction into science courses; (b) to improve teacher knowledge of, and attitude towards, literacy instruction; and (c) to improve students' general reading literacy skills.

The impact study involves a single confirmatory impact analysis to assess the average impact of iRAISE on high school students' general reading literacy skills after one year. General reading literacy will be assessed using a test developed by ETS that was previously used and validated through the RAISE i3 Validation grant. The evaluators will also conduct exploratory analyses to investigate whether impact varies by socioeconomic status, ELL status, and student pretest score. Further exploration will assess impacts on teacher knowledge, attitudes, and approaches to literacy instruction.

1. The single confirmatory research question involving student outcomes is: *Is there a positive impact of iRAISE on general reading literacy outcomes, after one year, as measured through an ETS assessment of the construct?*

In addition to the confirmatory research question, the impact study examines the following exploratory research questions.

2. Is there a differential impact of iRAISE on general reading literacy, after one year, *depending on student ELL status*?

3. Is there a differential impact of iRAISE on student general reading literacy, after one year, *depending on student socioeconomic status*?
4. Is there a differential impact of iRAISE on student general reading literacy, after one year, depending *on student prior achievement*?
5. Are impacts of iRAISE on student general reading literacy, after one year, mediated through impacts on teacher literacy instructional practices?

BENEFITS OF AN RCT

For this experimental study, Empirical worked with iRAISE program managers and RAISE state coordinators to initially recruit roughly 100 teachers to participate. After establishing eligibility criteria, the randomized sample included 82 teachers. We divided these 82 teachers into two groups: a group of teachers who were to be trained on and use iRAISE (*iRAISE* group) and a group of teachers who were to continue with their existing program (control group), that is, "business as usual." We first paired teachers within each school and used a random number generator to randomly determine which teacher would join the *iRAISE* group and which teacher would be in the control group.

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 and in their ability to take advantage of the new program. Random assignment to experimental conditions does not, however, assure 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 provides 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 for the randomized control trial 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 where iRAISE was in place and teachers using the schools' current methods, where the outcomes of interest are the student test scores on the ETS literacy assessment. This section details the methods we will use 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 initial composition of the experimental teacher groups.

EXPERIMENTAL DESIGN

There is always a level of uncertainty in our estimates of the effects of a program. The uncertainty can be understood in terms of the likelihood that we would obtain a different result if we took a new sample of teachers from the same larger population. It is important to recognize that the study results could change if we were to select a new sample. Our design attempts to efficiently deploy the available resources to reduce uncertainty and improve precision; in other words, to reduce the likelihood that we would obtain a different result if we tried the experiment again. The design of the experiment is

based on our best estimation of the amount of variability in outcomes that is not attributable to the program, and we attempt to detect the stable signal (the effect) against this random variation, and dampen this noise where possible. Due to the challenges inherent in recruiting schools and the voluntary nature of any experimental study, the sample was largely one of convenience. The reader must be cautious in generalizing the results beyond the sample, taking into consideration the particular characteristics of the sample and other conditions of the study. Before beginning the experiment, we created a design or plan in which we establish the specific questions to be answered.

First, before seeing the results, we specify the research questions and identify the effects that we will analyze to address the questions. This includes average impacts, as well as differential and mediated effects of the program. In this way, we avoid 'fishing' for results in the data, a process that can lead to mistaking chance differences for differences that are probably important as a basis for decisions. Because some differences in outcomes will appear simply by chance, mining the data in this way can capitalize on chance—concluding that there is an effect when really we're just picking the outcomes that happen to be large enough to be considered significant, but are attributable only to chance variation. We can still explore the data after the fact, but this is useful mainly for generating ideas about how the new program worked; that is, as hypothesis-generating efforts for motivating future study, rather than as efforts from which we make firm conclusions from our existing study.

Second, an experimental design will include a determination of how large the study should be in terms of units—such as students, teachers, or schools—in order to get to the desired level of confidence in the results. In the planning stage of the experiment, we calculate either how many cases we need to detect an effect of a certain magnitude, or how big an effect we can detect given the sample sizes that are available. Technically, this is called a power analysis. We will explain several aspects of the design and how they influence the sample size needs for the experiment.

How the Sample was Identified

How the participants for the study are chosen largely determines how widely the results can be generalized. The *iRAISE* sample was one of convenience, chosen from school districts that responded to invitations from the RAISE 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. Eligibility criteria were established: eligible teachers would teach at least one of the 5 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.

Randomization

We would like to determine whether iRAISE caused a difference in outcomes. To do so we have to isolate its effect from all the other factors influencing performance. Randomization ensures that, on

average, characteristics other than the program that affect the outcome are evenly distributed between program and control groups. By balancing out the effects of these factors between conditions, we arrive at an unbiased estimate of the program effect. Any remaining departures from the true values of the effects are due to chance differences between conditions and not systematic differences.

There are various ways to randomize to experimental conditions. Our research works within the organization of schools. The level of randomization is generally determined on the basis of the kind of program being tested. We attempt to identify the lowest level at which the program can be implemented without unduly disrupting normal processes or inviting sharing or "contamination" between control and program units. For example, school-wide reforms call for a school-level randomization while a professional development program that can be implemented individually per teacher, like iRAISE, can use a teacher-level randomization.

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. 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. 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. Because teachers, instead of students, were assigned to the *iRAISE* group or the control group, this kind of experiment is often called a "group randomized trial."

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 selection of sections. 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 average instruction. Instruction will be followed by monthly teacher surveys throughout the school year in the target class, and student outcome data will be collected from each teacher's spring target class.

Immediately after randomization, two control teachers declined to participate in the research for reasons exogenous to the study (both left their schools). Before the start of the study, one additional control teacher and three *iRAISE* teachers also declined to participate, leaving the sample with 76 teachers, evenly balanced across *iRAISE* and control. Five of the 38 remaining *iRAISE* teachers declined

to participate in the professional development but will still participate in the data collection activities, leaving 33 teachers receiving the full program.

What Factors May Moderate the Impact of iRAISE?

The selected design allows us to measure the differential effectiveness of iRAISE for specific subgroups of students. These are variables that were measured before the experiment started and that we had reason to believe would affect the magnitude of the effect of iRAISE. Technically, these are called potential moderators because they may moderate (increase or decrease) the impact of iRAISE. We measure the effect of the interaction between each potential moderator and the variable indicating assignment (i.e., to *iRAISE* or control); that is, we measure whether the effect of iRAISE changes across levels of each moderator.

In this study, we will explore the program's effectiveness based on English language learner (ELL) status, socioeconomic status (SES), and pretest scores. We chose these particular moderators because of their prior inclusion in the RAISE Validation study.

What Factors May Mediate Between iRAISE and the Outcome?

A mediator lies along the causal path between the point where we assign cases to the iRAISE or control group, and the point when we measure student performance after the intervention is over. A mediator can either block or enhance the effect of an intervention, either entirely or in part. Because we don't assign cases to levels of the mediator, we cannot be sure whether it is a proxy for an intermediate effect that we have not identified. The mediating variable and the final outcome do not have to be at the same level: for example, the survey outcomes are measured at the class or teacher level, while the final outcome is measured at the student level.

We identified variables that we believed would facilitate the effect of iRAISE on student outcomes and that could only be measured after the experiment had started. These are called "potential mediators" and are themselves intermediate outcomes, measurable in both assignment groups, which may be impacted by iRAISE. We usually think of a mediator as a factor in *how* the program has an impact. Based on the nature of the program, we identified process variables or mediators that were likely to facilitate the overall impact of the program. To assess mediation, we will first test whether there is a difference between the *iRAISE* and control group in the anticipated mediating processes. If there is, then we will consider whether effects of the program on final outcomes are facilitated through a prior impact on the mediating processes. In this experiment, we will explore the impact of iRAISE on teacher survey outcomes that describe their use of literacy strategies and the potential mediation of these outcomes on student achievement.

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. 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 but 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 average 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). Technically, it is the ratio of the variation in the teacher averages of students' scores to the total variation in students' scores. A larger ICC means between-teacher differences in student posttest scores contribute more noise to our program effect estimate. A larger sample of teachers is then needed to dampen the noise to acceptable levels. We select a value of the ICC before the beginning of the study.

It is possible that certain design strategies lower the ICC. For example, we believe that the process of creating matched pairs of teachers balances the between-teacher differences that contribute noise to the estimate, thereby effectively lowering the ICC. Because we do not have reliable estimates of the benefits of this strategy, we do not figure them into our power calculations; therefore, in the event that matching was successful, our power calculation can be considered conservative in its determination of the number of teachers needed. (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.)

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 are likely to co-vary with the outcome. 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. (Randomization assures that the covariates, on average, take the same value in both conditions; however, in any one trial, they may be imbalanced by chance. Adjusting for the effects of this imbalance increases the precision of our estimate of the effect of *iRAISE*.) Technically, a covariate-adjusted analysis is called an analysis of covariance (ANCOVA). 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.

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

We want to be certain that if we conclude there is no impact that, in fact, there is no impact (we want to limit the possibility of drawing a false negative conclusion). Also, we want to be certain that if we conclude there is an impact that, in fact, there is an impact (we want to limit the possibility of drawing a false positive conclusion). 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 (strong, some, limited, or none) rather than as a yes-or-no declaration. As we describe later, 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 7 percentile points for the ETS 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 7 percentile points higher (i.e., at the 57th percentile) or 7 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. The sample size calculation was conducted using Optimal Design, a software program developed for this purpose (Spybrook, Raudenbush, Congdon, & Martinez, 2011). This was revised slightly upward from what was originally in the design guide: with a sample size of 80, the MDES was 0.18, while it increases to 0.19 with only 76 teachers. These calculations were done assuming an intraclass correlation coefficient 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.

SITE DESCRIPTION

Although student-level data for the confirmatory analysis will not be collected until summer 2015, school-level data is provided here to give a sense of the context for implementation. The 27 schools are spread equally across the two states, with 13 in Michigan and 14 in Pennsylvania, and nearly equally across the 4 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-provided data 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%
English Language Learners	3%
White	70%
Black	12%
Hispanic	16%
Asian	1%
American Indian/Native Alaskan	0.3%
Multi racial/No response	2%

Source. NCES 2012-2013 school year

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

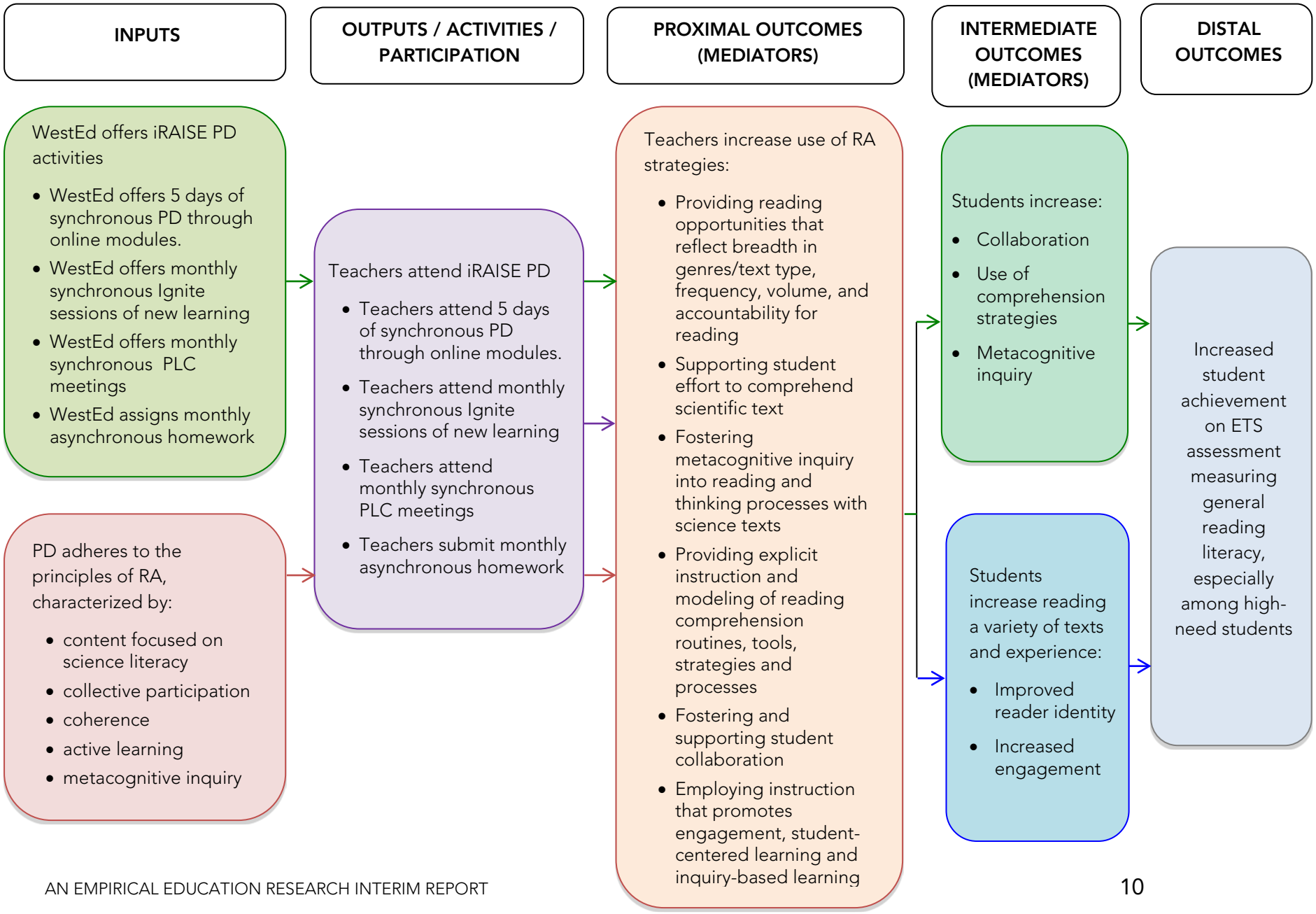
Fidelity of Implementation

As a requirement of the NEi3, Empirical will be reporting of the fidelity of implementation of the iRAISE program. The implementation study applies mixed methods to assess the key components of the logic model (Figure 1 below), including: presence of inputs, activities and outputs, such as the presence of course modules and course facilitator interactions with teachers and teacher online interactions in the use of the program.

The implementation study involves data collection from multiple sources, including observation of professional development, monthly teacher surveys, and attendance log records. Professional development sessions for teachers in the *iRAISE* condition are being observed. Monthly surveys of *iRAISE* and control teachers are providing information about classroom practices, leadership, and ongoing support. Additionally, data from the surveys on teacher practices will be used to describe the *iRAISE* /control contrast.

During the pilot study in the 2013-2014 school year, Empirical worked with the developers of iRAISE to identify the key components of the intervention and the indicators for each component. We will assess implementation fidelity in terms of the following components: (1) WestEd delivers professional development, (2) teachers attend professional development, and (3) professional development adheres to the principles of Reading Apprenticeship (i.e., the extent to which content is focused on science literacy, level of collective participation, level of active learning, level of coherence, level of engagement in metacognitive inquiry). Program components and fidelity indicators are shown in the fidelity matrix (Table 2) below, and a longer description can be found in the October internal feedback report from the pilot study (Toby, Schellinger, & Jaciw, 2013).

FIGURE 1: FACTORS THAT FACILITATE AND INHIBIT IMPLEMENTATION: POLICIES, PROFESSIONAL COMMUNITY, SUPPORT FOR IMPLEMENTATION, EXTERNAL CONTEXT



FIDELITY MATRIX

TABLE 2. FIDELITY OF MATRIX

Key component	Operational definition	Source of information/ schedule of data collection	Individual-level threshold	Sample-level threshold
Component 1: WestEd 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
	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: WestEd assigns monthly asynchronous activities	Observations, program log data, and teacher surveys	Not applicable	0: WestEd assigns at least one asynchronous activity per month 1: WestEd 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 2. FIDELITY OF MATRIX

Key component	Operational definition	Source of information/ schedule of data collection	Individual-level threshold	Sample-level threshold
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)
	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 2. FIDELITY OF 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**

SCHEDULE OF MAJOR MILESTONES

Table 3 lists the major milestones in this study and associated dates.

TABLE 3. TIMELINE OF KEY *iRAISE* DATA COLLECTION AND ACTIVITIES FOR YEAR 2 RCT

Target Date	Data Collection Event / Planning Activity	Completed
January 2014	State coordinators send recruitment materials and answer initial questions about iRAISE RCT	X
February 28, 2014	Empirical submits iRAISE Second Internal Feedback Report to WestEd	X
March 28, 2014	Final deadline for iRAISE RCT district applications	X
March 31, 2014	Empirical invites selected/eligible districts to participate in the research study and sends districts agreements	X
April 18, 2014	Final deadline for district agreements	X
April-May 9, 2014	Once district agreement is signed, interested teachers/principals participate in a webinar for the study, consent to participate, and complete initial survey	X
May 16, 2014	Final deadline for teacher consent forms	X
May 16-30, 2014	Empirical forms matched pairs of teachers within schools, solicits principal feedback, and finalizes matched pairs	X
May 31, 2015	Empirical randomizes teachers to iRAISE or control group	X
June 2, 2014	Empirical provides results of randomization to WestEd and participating teachers and schools	X
June 2014	WestEd contacts teachers assigned to the iRAISE group about the summer PD and scheduling	X
July 31, 2014	Empirical submits iRAISE Final Internal Feedback Report to WestEd	X
August 11-15, 2014	Reading Apprenticeship Science Foundations Training	X
August 18-22, 2014	Reading Apprenticeship Science Foundations Training	X
August 2014-May 2015	Empirical deploys monthly teacher surveys during the school year	In progress
September 2014-May 2015	Monthly Ignite Sessions and Professional Learning Communities (PLCs)	In progress
Spring 2015	Empirical coordinates ETS assessment administration and obtains student posttests	In progress
Summer-Fall 2015	Data analysis and report writing	
December 2015	Delivery of Final Report	

DATA SOURCES AND COLLECTION

The data for this study are primarily provided by the school districts and collected by Empirical Education. In addition to achievement and demographic data, we also collect implementation data over the entire period of the experiment, beginning with the teacher trainings in August 2014 and ending with the academic calendars of the schools in June 2015. Data collected through teacher background forms, training observations, multiple teacher surveys, principal surveys, *iRAISE* log data, and teacher interviews are used to provide evidence of the implementation. In addition, we have reviewed various program documents and materials. Table 4 outlines the timeline of the major data collection phases for the RCT.

TABLE 4. IMPLEMENTATION DATA COLLECTION SCHEDULE FOR THE iRAISE STUDY

Data collection elements	2014-2015 school year									
	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]	[X]	
District/school data request				[X]			[X]			
ETS assessment								[X]	[X]	

Teacher Training Observations

We observed the initial *iRAISE* Foundations training and asked additional questions about the initial training through the teacher online surveys. We continue to observe the monthly Ignite sessions and code the PD content for our implementation evaluation.

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 survey. This packet provided general information about the research study, data collection activities, and participant responsibilities, in addition to the teacher consent form. The survey also included teacher background questions for teachers to complete, 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 to participating teachers beginning in August 2014 and will continue on a monthly basis through May 2015. Table 5 outlines the survey schedule and response rates for the control and *iRAISE* teachers participating in the study.

TABLE 5. SURVEY PARTICIPATION RATES

		Response rates		
Survey	**Date**	**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	89%	87%	88%
Monthly Survey 4	December 2014	87%	87%	87%
Total		94%	92%	94%

The teacher surveys have been extensively piloted, both through the prior RAISE RCT and the prior year's iRAISE pilot study, and revisions have been made to capture more detail on the variation in time spent in classroom activities and the level of student engagement.

The repeated target class questions on the monthly surveys will not be presented in this report. This bank of questions, asked of both control and *iRAISE* teachers, reflects a spectrum of teaching practices and literacy strategies that are expected to be impacted by the implementation of the iRAISE program over the school year, and we are waiting to process additional rounds of survey responses before we summarize the differences between iRAISE and control in teacher reported practices. Typical questions ask 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. Along with the target class questions, teachers report in each monthly survey on the context for literacy instruction, including support from other teachers and administrators.

Principal Survey

Consented principals are asked to participate in a one-time survey in May 2014. This survey gathers 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 will be asked to participate in brief interviews in May 2015. These interviews will gather valuable information on the context for implementation, including challenges and supports, and provide 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 the fall to familiarize teachers with the data collection process and to allow us to track attrition at the student level. We will request updated spring target class rosters in late winter. For the ETS assessment, we will request target class rosters and student IDs to manage the data collection. We will also request student-level demographic data, including the standardized assessment pre-test scores, in the summer of 2015. These data are required to conduct equivalence tests, and conduct moderator analyses. Specifically, we will ask 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
- Pre-test scores

All student and teacher data having individually identifying characteristics will be stripped of such identifiers for analysis, and the data will be stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA). This experiment falls within the protocol approved by Ethical and Independent Review Services, Empirical Education's Institutional Review Board (IRB). Under this protocol and following FERPA guidelines, student or parental permission was not necessary, nor was it was required by the school district.

ETS Assessment

For the ETS literacy assessment, testing will occur in a window from the beginning of April through the end of May. Empirical Education will coordinate with ETS and designated point-of-contacts in each school to ensure the appropriate technology is available for the computer-based test and sufficient training and documentation is provided to all teachers. During the testing window, Empirical will monitor the assessment response rate and remain in frequent communication with the point of contacts to prevent attrition.

Teacher Background

This section describes the sample of teachers that was randomized. We start with the baseline sample, which consists of the participating teachers that were randomly assigned to the *iRAISE* or control groups. The sample for which outcomes are analyzed may be modified somewhat from baseline through attrition or for other reasons that data become unavailable. We expect similar outcomes in both conditions given that random assignment would, on average, achieve a balanced distribution, and none of the differences by condition were statistically significant at the 5% level.

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 RA exposure	5 (13%)	3 (8%)	No

INITIAL IMPLEMENTATION FINDINGS

This section presents findings from the background survey, training survey, and first four monthly surveys. The data presented are meant to provide context for the teacher sample and give a picture of implementation over the first semester of the school year. We include results on teacher attitudes toward literacy instruction, context for instruction, and impressions of the training and support. Differences presented here between the *iRAISE* and control groups are not tested for statistical significance, and the target class questions will not be analyzed until we have the full year of data.

Attitudes toward Literacy Instruction at Baseline

On the initial background survey, deployed in June 2014, teachers were asked about their attitudes surrounding literacy in the science classroom. Although we present their responses by randomized group, their responses were given prior to randomization. They were asked the extent to which they agreed with the following statements.

- My role in teaching literacy in my science content area is essential for students to succeed in reading
- My role in teaching literacy in my science content area is essential for students to succeed in science
- Being an effective reader is an essential precondition for grasping the science content areas that I teach
- My job is to teach science content and let the ELA department worry about teaching reading

Attitudes were similar among the *iRAISE* and control groups, with nearly 80% of teachers in both groups either somewhat or strongly agreeing that literacy and effective reading in the science classroom are essential to student success in both reading and science (Figures 2 and 3). Similarly, nearly 80% of teachers disagreed or strongly disagreed that reading was only the domain of ELA departments. We expect similar outcomes in both conditions given that random assignment would, on average, achieve a balanced distribution. Still, we see that by chance alone, there are some differences in responses between conditions.

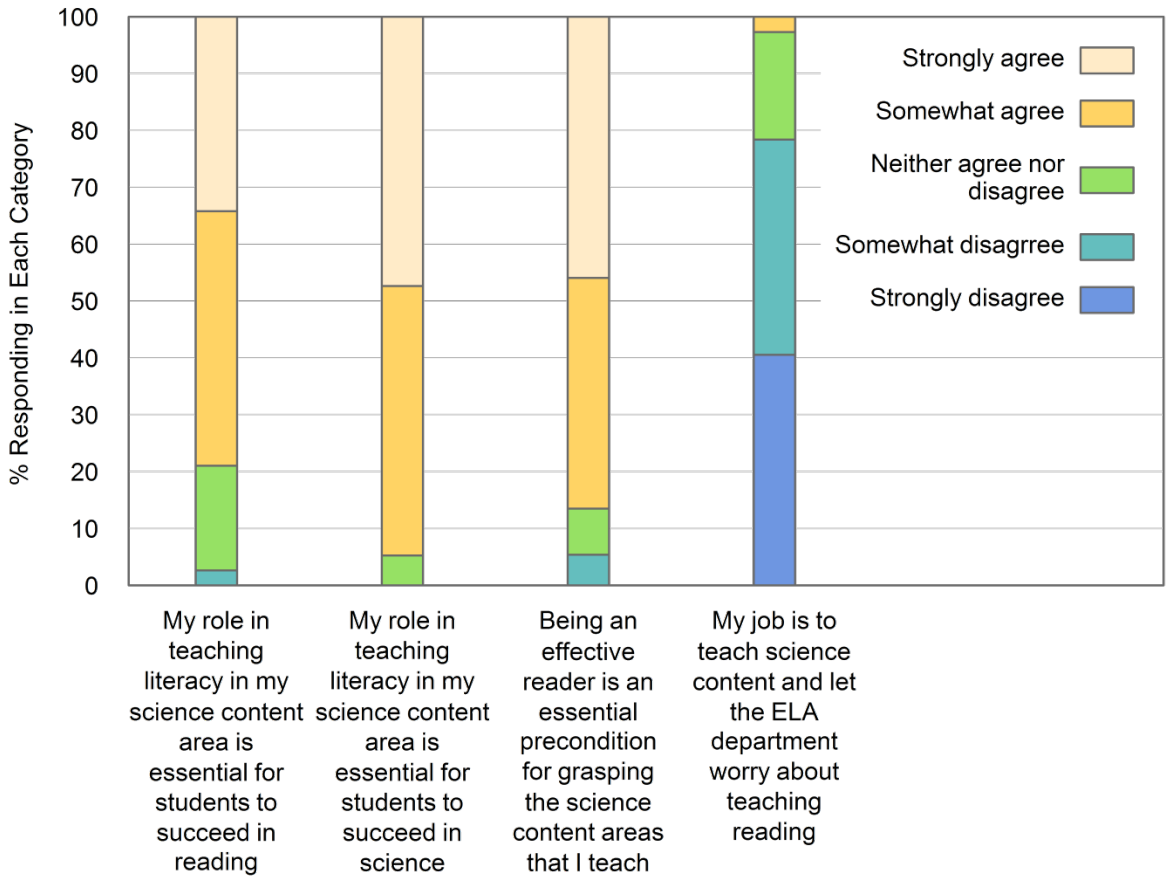


FIGURE 2. *iRAISE* TEACHERS' ATTITUDES TOWARDS LITERACY INSTRUCTION

Note. *n* = 38 for each statement except the third, where *n* = 37

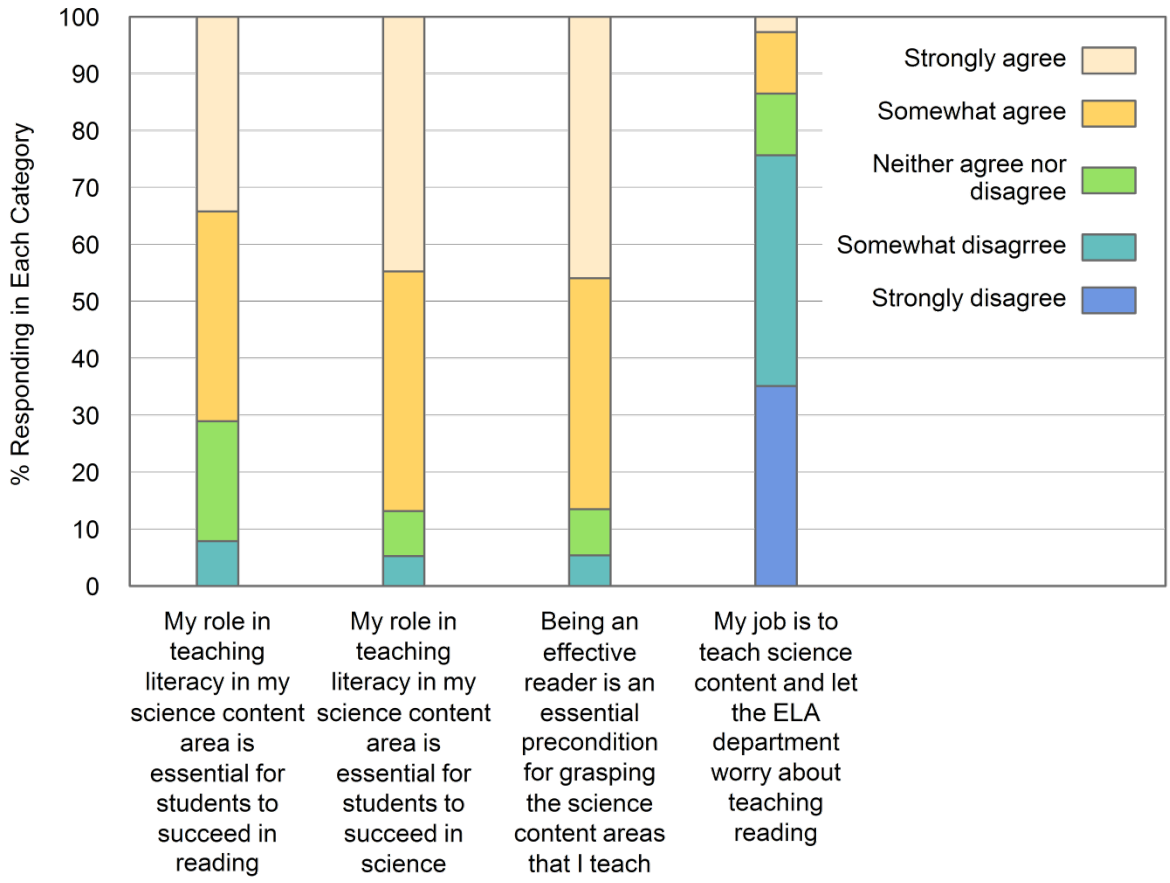


FIGURE 3. CONTROL TEACHERS' ATTITUDES TOWARDS LITERACY INSTRUCTION

Note. *n* = 38 for the first two statements, *n* = 37 for the second two statements

Teacher Impressions of Foundations Training

Of the 33 teachers attending the professional development, 28 attended the full 5 days (20 hours) of the Foundations training, while 4 teachers missed 1 day (or 4 hours), and 1 teacher missed three days (12 hours).

iRAISE teachers were sent a short survey exploring their opinions of the Foundations training one week after training. The survey asked about their levels of preparation for teaching literacy strategies after the training, their formative opinions of the benefits or hindrances of online trainings, any technical difficulties they experienced, any comments on the Reading Apprenticeship approach, and any recommendations for improving the training.

Teachers were asked whether they would prefer iRAISE or a face-to-face training, with roughly one-third of teachers selecting that they would prefer iRAISE, one-third a face-to-face training, and one-third not sure. Teachers who responded positively to iRAISE valued the flexibility that not traveling

afforded them, while teachers who preferred a face-to-face training mentioned the difficulty in bonding and social interaction that typically accompany in-person trainings. Roughly half of the teachers reported some level of technical difficulty, with the main problems being slow connection speed and audio difficulties. Almost all of the audio difficulties were solved over the course of the Foundations training as participants became familiar with each other and adapted to taking turns speaking or muting or unmuting microphones as needed.

iRAISE teachers were then asked about the extent to which they felt the training prepared them to implement a range of literacy strategies referenced on the background survey. Teachers felt most prepared to model/demonstrate metacognitive routines such as Think Aloud and Talking to the Text, with 78% of teachers more than moderately or completely prepared to do so. Teachers felt least positively about structuring lessons so that students would be held accountable for reading, with only 27% of teachers reporting more than moderate or complete levels of preparation. Figure 4 shows the extent to which teachers felt the training prepared them to implement the following reading strategies (bolded text matches the response labels in Figure 4).

- **Modeling**/demonstrating metacognitive routines (e.g. Think Aloud, Talking to the Text)
- Teaching **students** to **analyze** their **own thinking** about reading texts
- Supporting students in their attempts to understand **disciplinary text** (e.g. challenging literature, textbooks, primary documents, scientific articles)
- Asking **students** to **pose questions** and problems about course readings
- Supporting students in **working** on reading or writing activities **collaboratively**, (i.e. setting norms, creating safety, providing prompts that promote collaboration, and providing guidance/feedback on student participation)
- Facilitating students' active engagement in learning through the use of **inquiry-based instruction**al methods
- Providing students opportunities for reading a **variety of texts** of different types/genres
- Employing **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.
- Structuring lessons so that **students** are held **accountable** for **reading** (e.g., students have to do the assigned reading in order to be successful)

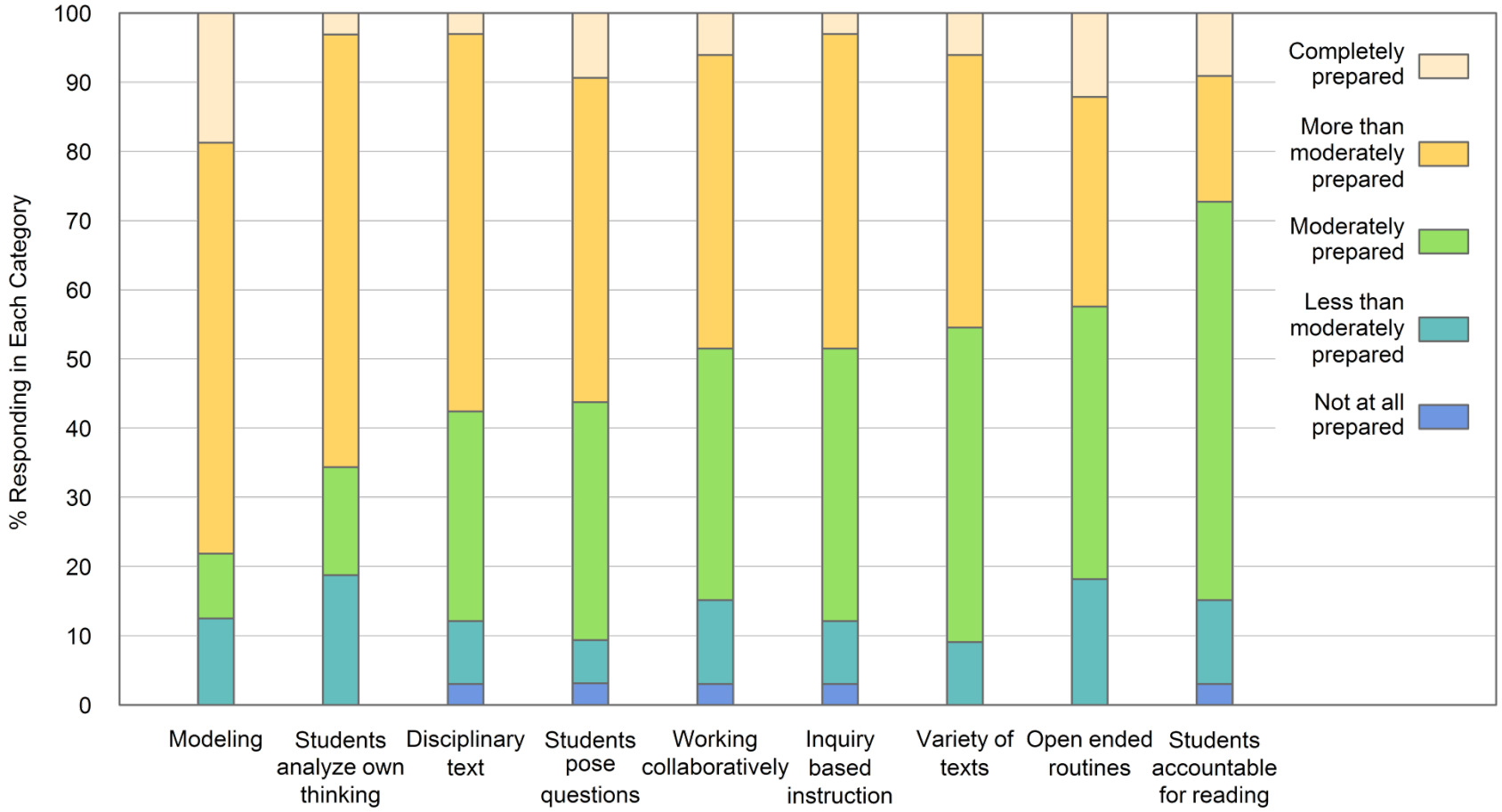


FIGURE 4. *iRAISE* TEACHERS' FEELINGS OF PREPAREDNESS AFTER FOUNDATIONS TRAINING

Note. n = 33 for each response except the fourth, sixth, and seventh, where n = 32

iRAISE teachers also responded extremely positively to the overall training, with 91% of teachers agreeing or strongly agreeing that it would help them collaborate with colleagues to respond to student needs, 79% agreeing or strongly agreeing that they were provided with adequate resources to implement what they learned, and 97% of teachers agreeing or strongly agreeing that it would lead to a change in their teaching practices (Figure 5).

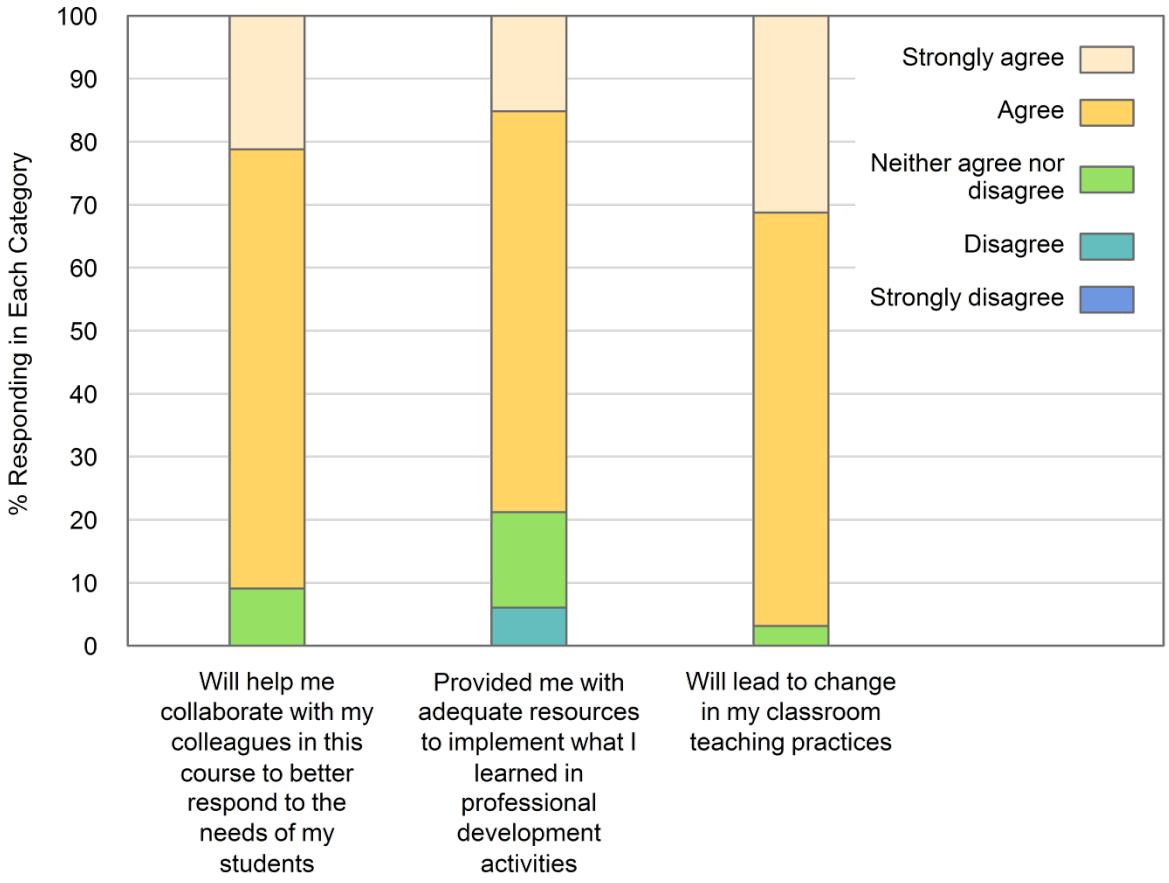


FIGURE 5. *iRAISE* TEACHERS' IMPRESSIONS OF FOUNDATIONS TRAINING

Note. *n* = 33 for all statements

Teachers who responded with recommendations about the training (*n* = 33) most frequently referenced the timing and structure. Several teachers recommended having smaller groups or asynchronous content to accommodate different schedules, and six teachers reported feeling like the four-hour sessions were too long. Four teachers suggested having clearer goals and outlines for each individual session. Full text responses for this and other free-response questions can be found in Appendix A.

Feedback on Monthly Ignite and PLC Sessions

Table 7 shows the self-reported training attendance at monthly Ignite and PLC sessions over the fall months. Teachers were asked if they attended each session. If they responded yes, they were also asked how many hours they attended, how helpful the session was, and for any comments regarding

the training (reported below). If they responded no, they were asked for reasons for not attending. The most common responses were "other obligations" or "other", usually referencing other scheduled events or family/personal matters.

TABLE 7. TEACHERS REPORT TRAINING ATTENDANCE

	September	October	November	December
Ignite	29/32 (91%)	26/30 (87%)	24/30 (80%)	20/28 (71%)
PLC	26/30 (87%)	26/30 (87%)	22/28 (79%)	N/A[a]

[a] Due to the timing of the monthly surveys, questions regarding the Ignite sessions are asked in their respective months, while questions regarding the PLC sessions are not asked until the following month's survey (e.g., questions about the December PLC sessions are on the January teacher survey)

Teachers began to report lower levels of satisfaction with the ongoing monthly support over the course of the semester. The percentage of teachers deeming the monthly Ignite session as "more than moderately helpful" or "very helpful" fell from 55% in September to just 25% in December, with a similar drop in the level of helpfulness in the PLC sessions, 50% to 28% (see Figures 6 and 7).

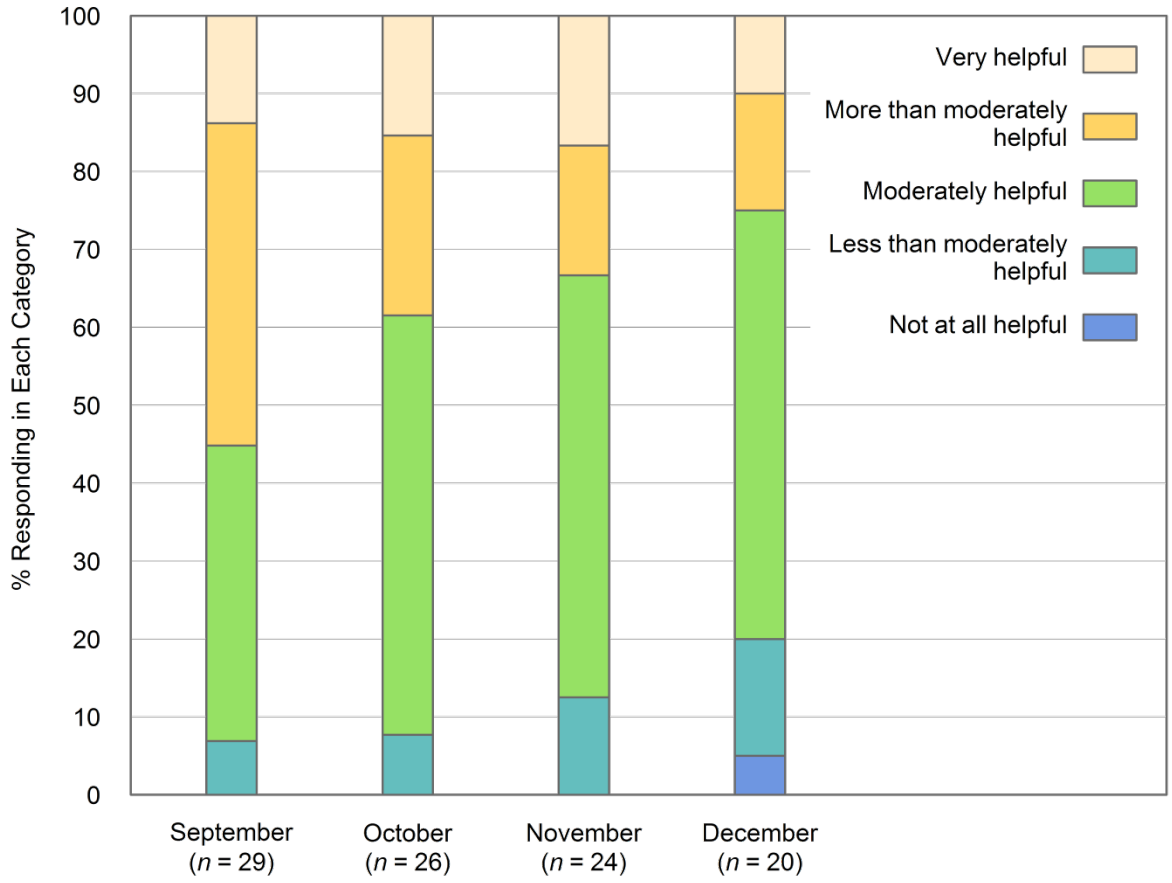


FIGURE 6. *iRAISE* TEACHERS' REPORT OF HELPFULNESS OF IGNITE SESSIONS

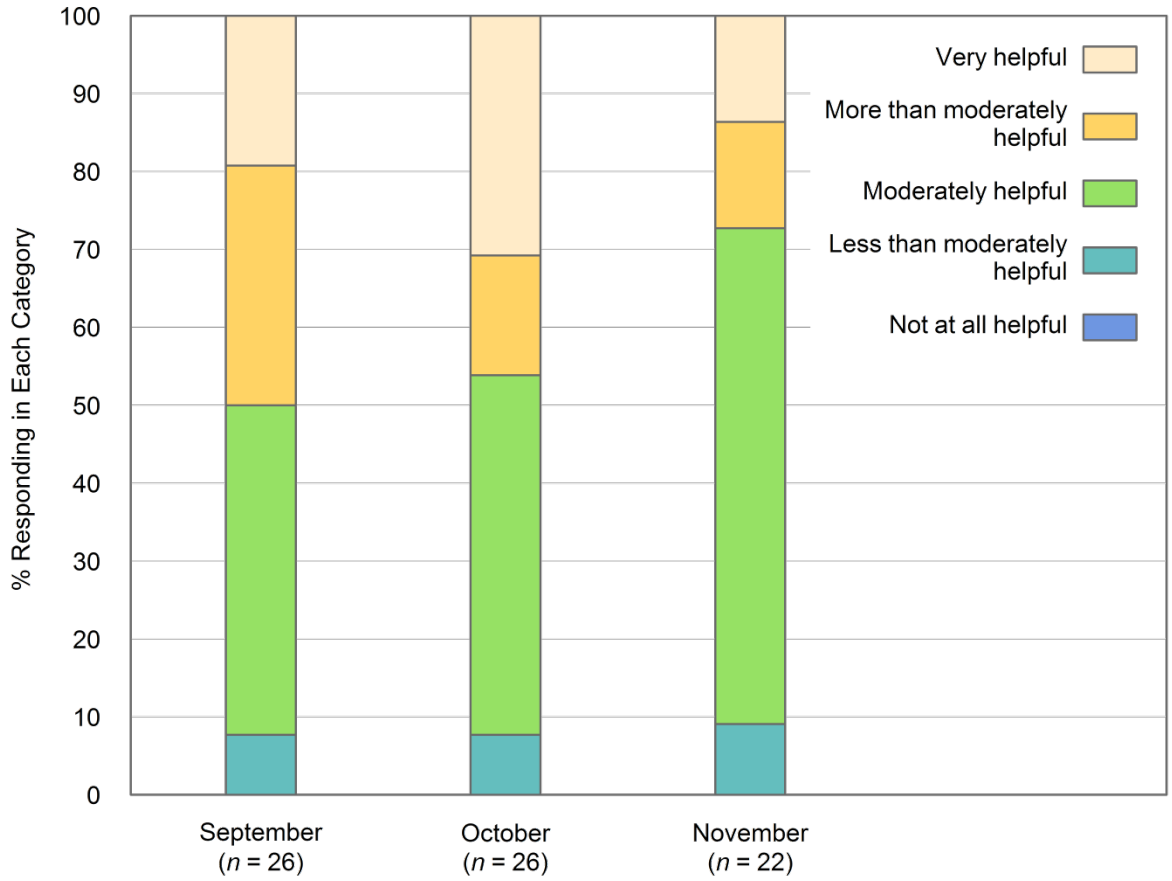


FIGURE 7. TEACHERS' REPORT OF HELPFULNESS OF PLC SESSIONS

Impediments to Implementation

Researchers also surveyed teachers regarding any challenges with implementing iRAISE. Teachers could select "None" or check all that apply from the list in Table 8. In response to the October survey, the most selected challenges were competing priorities and the pace of instruction. Only 11% of teachers selected facing no challenges.

TABLE 8. CHALLENGES IN IMPLEMENTING RA

	Number of teachers (*n* = 28)
Competing priorities	20 (71%)
RA slowed the pace of instruction	16 (57%)
Lack of materials	12 (44%)
Lack of understanding of how to implement RA in my class	11 (39%)
Student behavior	11 (39%)
RA takes too much time to incorporate	9 (32%)
Student ability	9 (32%)
Not enough training on RA	7 (25%)
Lack of parent support	4 (14%)
Too much work to implement	4 (14%)
Lack of administrative support	3 (11%)
None	3 (11%)
Other	2 (7%)

Teachers were also asked in November to rate their understanding of Reading Apprenticeship on the following scale.

- I get it and am referring to it often as I plan and reflect on my teaching
- It makes sense to me as I work with the approach to integrate it into my daily practice
- I understand some aspects of it, but I do not understand how it would translate into daily practice
- I do not get it
- Other (please explain)

Over two-thirds of respondents (n = 30) answered positively, with 10% of respondents selecting the most positive option, "I get it and am referring to it often as I plan and reflect on my teaching," and 60% of teachers selecting "It makes sense to me as I work with the approach to integrate it into my daily practice." Only 1 teacher answered "I do not get it."

Implementation of Reading Apprenticeship in Different Classes

The target class for each teacher was chosen in such a way from their eligible sections as to try to maintain similarity between classes within pairs and to have an adequate representation of class types across the sample. In the initial monthly surveys, several teachers commented on characteristics of their target classes that might affect implementation. In response, a question on the December survey asked the following.

How do you incorporate Reading Apprenticeship into instruction in your target class compared to your other science classes?

- I incorporate Reading Apprenticeship about the same in my target class as in my other science classes
- I incorporate Reading Apprenticeship less in my target class than in my other science classes
- I incorporate Reading Apprenticeship more in my target class than in my other science classes
- In some science classes I incorporate Reading Apprenticeship more than in my target class; in other science classes I incorporate Reading Apprenticeship less than in my target class

As shown in Figure 8, two-thirds of teachers reported that they implement Reading Apprenticeship about the same in their target classes as in their other science classes.

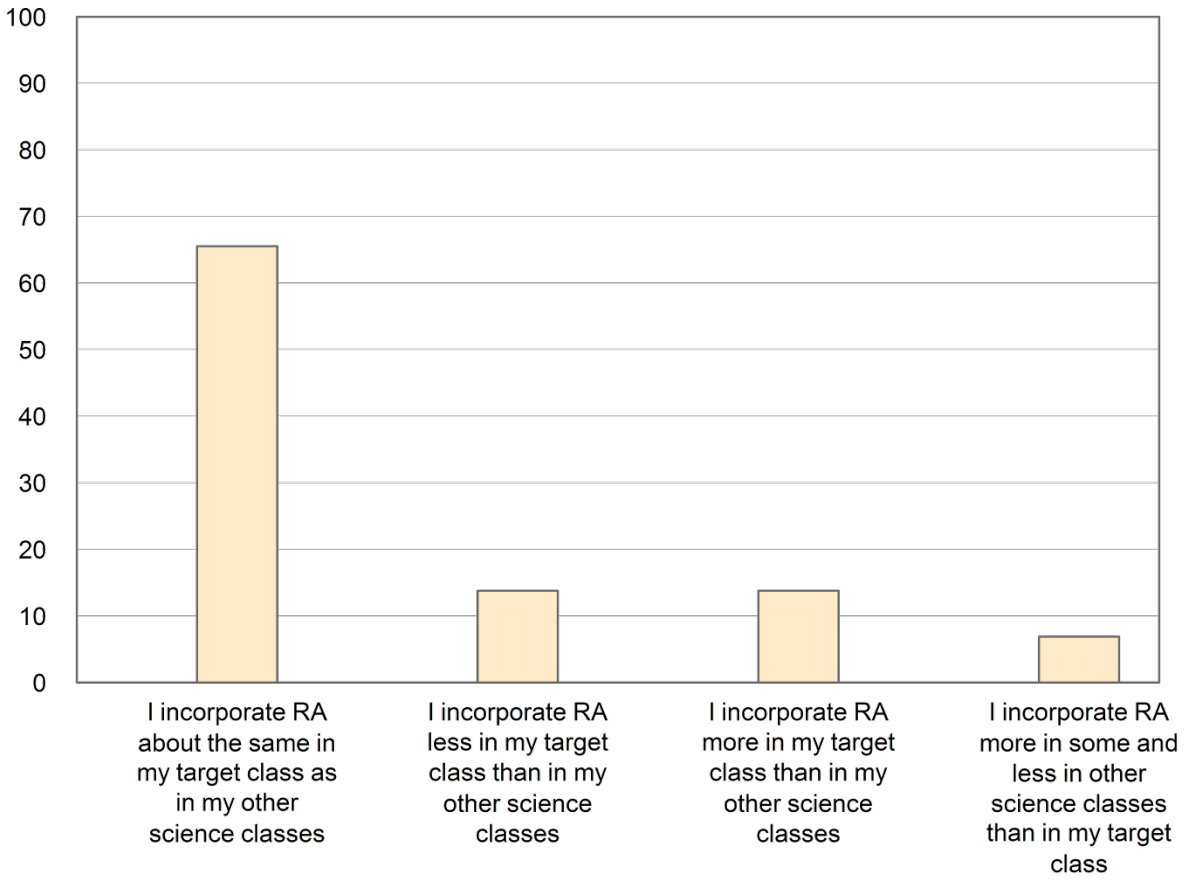


FIGURE 8. *iRAISE* TEACHERS' REPORT OF RA IMPLEMENTATION ACROSS CLASSES

Note. *n* = 29

Support and Discussions for Literacy Instruction

On every other monthly survey, all teachers are asked whether or not they received support for literacy instruction or had discussions with other teachers regarding literacy instruction. *iRAISE* teachers are asked to exclude official iRAISE activities. Teachers may select "I did not receive support for literacy instruction" or they can check all that apply from the types of support listed. As seen in Tables 9 and 10, the majority of teachers selected that they did not receive support for literacy instruction, while the most common response among teachers who did receive support was "resources," selected by 14% of teachers in October and 19% in December.

TABLE 9. SUPPORT FOR LITERACY INSTRUCTION, OCTOBER

	Control (*n* = 36)	*iRAISE* (*n* = 33)	Total (*n* = 69)
I did not receive support for literacy instruction	24 (67%)	15 (45%)	39 (56%)
Resources	4 (11%)	6 (18%)	10 (14%)
Coaching/mentoring	1 (3%)	7 (21%)	8 (12%)
Model lessons	1 (3%)	6 (18%)	7 (10%)
Other (please explain)	5 (14%)	2 (6%)	7 (10%)
Observation/feedback	0 (0%)	5 (15%)	5 (7%)
Classroom management help	0 (0%)	3 (9%)	3 (4%)
Political support (e.g. someone "backed you up" in a conflict over your implementation of literacy instruction)	1 (3%)	0 (0%)	1 (0%)
A change in school/district policy that was relevant to literacy instruction	0 (0%)	1 (3%)	1 (1%)

TABLE 10. SUPPORT FOR LITERACY INSTRUCTION, DECEMBER

	Control (*n* = 34)	*iRAISE* (*n* = 29)	Total (*n* = 63)
I did not receive support for literacy instruction	26 (76%)	13 (45%)	39 (62%)
Resources	2 (6%)	10 (34%)	12 (19%)
Observation/feedback	3 (9%)	9 (31%)	12 (19%)
Coaching/mentoring	2 (6%)	6 (21%)	8 (13%)
Other (please explain)	2 (0%)	3 (11%)	5 (8%)
Model lessons	0 (0%)	3 (11%)	3 (5%)
Classroom management help	0 (0%)	0 (0%)	0 (0%)
Political support (e.g. someone "backed you up" in a conflict over your implementation of literacy instruction)	0 (0%)	0 (0%)	0 (0%)
A change in school/district policy that was relevant to literacy instruction	0 (0%)	0 (0%)	0 (0%)

On the November survey, teachers were asked about the type of discussions they may have had in the past four weeks with other teachers regarding literacy instruction. As shown in Table 11, 65% of control teachers and 85% of *iRAISE* teachers reported some type of discussion, with most reporting either "discussing what helps students learn the best" or "sharing successful lessons."

TABLE 11. DISCUSSIONS WITH OTHER TEACHERS REGARDING LITERACY INSTRUCTION, NOVEMBER

	Control (*n* = 34)	*iRAISE* (*n* = 33)	Total (*n* = 67)
Discussing what helps students learn the best	15 (44%)	21 (64%)	36 (54%)
Sharing successful lessons	11 (32%)	18 (55%)	29 (43%)
Reviewing student work	9 (26%)	12 (36%)	21 (31%)
Discussing problematic lessons	10 (29%)	10 (30%)	20 (29%)
I did not have these types of discussions with other teachers about literacy instruction	12 (35%)	5 (15%)	17 (25%)
Discussing resources for literacy instruction	7 (21%)	5 (15%)	12 (18%)
Other (please explain)	1 (3%)	0 (0%)	1 (1%)

Summary of Implementation

Implementation appears to be progressing as expected, with preliminary data indicating that fidelity of implementation is being met for the delivery, attendance, and coherence of the professional development sessions.

Upcoming Steps

In the coming months, researchers will administer the final four monthly surveys and observe a sample of *iRAISE* training sessions. We will also work with ETS and the districts' schools and teachers to collect the literacy assessment during April and May. The final district data (i.e., student demographic and state assessment data) will be collected in the summer. We will present the final report, which will include the results of the RCT and fidelity of implementation findings in December 2015.

References

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

Toby, M., Schellinger, A.M., & Jaciw, A.P. (2013). *iRAISE Second Internal Feedback Report*. Unpublished report.

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).