An Evaluation of the Literacy-Infused Science Using Technology Innovation Opportunity (LISTO)
i3 Evaluation (Valid 45)
Final Report

Rebecca Wolf, PhD
Michael Cook, PhD
Alan Reid, PhD
Amanda Neitzel, PhD
Steven Ross, PhD
Kelsey Risman, MA

School of Education
Center for Research and Reform in Education

June 2021

Rafael Lara-Alecio, Ph.D., Principal Investigator
Beverly Irby, Ph.D., Co-Principal Investigator
Fuhui Tong, Ph.D., Co-Principal Investigator
Cindy Guerrero, Ph.D., Lead Coordinator

Texas A&M University
College of Education and Human Development

About the Center for Research and Reform in Education (CRRE) at the Johns Hopkins University School of Education

The Center for Research and Reform in Education (CRRE) is a research center within the Johns Hopkins University School of Education. Established in 2004, our major goal is to improve the quality of education for children in grades pre-K to 12 through high-quality research and the dissemination of evidence-based research. CRRE is committed to expanding the body of evidence on the effectiveness of educational programs and initiatives and assisting organizations and school districts to obtain the information they need to make evidence-based decisions.

Specializing in independent program evaluations, CRRE's research department evaluates the impacts of programs and services through four levels of evaluation studies: (1) design and implementation quality, (2) development, (3) efficacy, and (4) effectiveness. In terms of content areas, CRRE specializes in evaluations of educational technology and technology integration, social-emotional learning, professional development, school reform, programs for English learners, and multiple core subject curriculum areas. CRRE staff work with educators and program developers to design studies that are consistent with their organization's objectives and that meet the specific needs of clients. We evaluate programs locally, nationally, and internationally.

CRRE researchers include numerous Johns Hopkins University professors and research staff with backgrounds including quantitative, qualitative, and evaluative research. The research team has published over 200 research documents, and within the past five years alone, CRRE has conducted over 45 program evaluations nearing $10 million.

Contents

EXECUTIVE SUMMARY:
An Evaluation of the Literacy-Infused Science Using Technology Innovation Opportunity (LISTO) Validation Project

Overview

This study is an evaluation of the Literacy-Infused Science Using Technology Innovation Opportunity (LISTO) validation project (Valid 45). The LISTO project was funded by the Investing in Innovation (i3) Fund.[1] It involved a multi-year intervention that provided virtual professional development and coaching, and literacy-infused science curricula to fifth-grade science teachers who taught predominantly low-income students and in predominantly rural public schools in Texas.

Multiple professors at Texas A&M University were the recipients of the i3 grant that funded LISTO. The Center for Research and Reform in Education (CRRE) at the Johns Hopkins University School of Education as the independent, third-party evaluator of LISTO. This report describes the method and findings of the evaluation.

Program Description

The purpose of Project LISTO is to support the instructional capacity of science educators and to validate innovative practices and strategies via previously developed interventions that address literacy-infused science and technology integration with standards-aligned curriculum. Specifically, LISTO compared enhanced Literacy-Infused Science (LIS) instruction to that of typical science instruction. LISTO provided standards-aligned, literacy-infused science curricula, ongoing virtual professional development, and on-going virtual mentoring and coaching to fifth-grade science teachers.

It is important to note that Hurricane Harvey brought many changes that impacted the first year of implementation for Project LISTO, including the launching of the first year of the project, implementation of all components, and fidelity of implementation. This extreme weather event included eight days of heavy rainfall from August 25, 2017 through September 1, 2017, resulting in more than 60 inches of rain that caused catastrophic flooding. School districts across Texas were hard hit, with over 1.4 million students directly impacted by the storm, more than $970 million worth of school building damage, and an estimated $1 billion school funding gap (Morath, 2017). Even after a full year, with the state's recovery still "far from over," according to the *Texas Tribune* survey, 8% of people had not yet returned to their homes (Formby, 2018). The hurricane caused a long-term impact on schools, teachers, students, and their families in the affected areas. These impacts included students missing instructional hours before and after schools reopened, staff periodically being absent from work or unable to return to their

[1] The award number is U411B160011.

classrooms fully, and schools under high pressure of gathering resources and funding for students and staff, which drove down students' tests scores (Davis et al., 2021).

Seven LISTO school districts (20%) are located within the declared disaster counties, inclusive of districts who applied for related Texas Education Agency accommodations that were directly impacted. Within these districts, a total of 14 LISTO campuses (17%) and 28 teachers (23%) were adversely affected by flooding and damage caused in the wake of Hurricane Harvey. A higher percentage of teachers were impacted as compared to control (29.8% treatment; 17.1% control). Teachers, students, and their families in coastal areas were displaced and some educational facilities were shuttered while others were relocated to different parts of the community and state. One treatment campus in the city of Houston, Texas in Houston ISD was damaged to the point that the building was demolished and rebuilt over the next two years. The staff and students were temporarily moved to an alternate location, which took weeks to prepare. Students missed more than four weeks of classes, and started back on September 25, 2017. These impacts included delaying the beginning of year testing, curriculum implementation, and professional development schedules for the original confirmatory group. Additionally, the observations were incomplete for the baseline collection. Two component parts of the intervention were delayed as well. The Science Role Models and Mentors did not engage until the second semester, and the Family Involvement in Science did not begin until Year 2.

Literacy-Infused Science Using Technology Innovation Opportunities (LISTO) Curricula. Teachers received LISTO curricular materials, which included 25 weeks of standards-aligned lesson plans, lesson scripts, related resources, and hands-on science activity supplies. Lessons were designed to be implemented within an 80-minute science block. Detailed, scripted lessons were organized using the 5E instructional model (in which at least three of the five E's – engage, explore, explain, elaborate, evaluate – were implemented in each lesson) and included embedded literacy-skills to facilitate listening, speaking, reading, and writing. Some of the strategies included engaging questioning, partner and group work, direct instruction of science academic vocabulary using visuals and student friendly definitions, supporting reading through pre-teaching pronunciation of vocabulary and words that are challenging to decode, strategic partner reading, leveled questioning, highlighting expository text features, sentence stems, graphic organizers, and integration of student use of technology via tablets.

LISTO included two sub-components: Family Involvement in Science (FIS) and Scientists as Role Models and Mentors (SRM^2). Although the intent was to implement both of these components starting in Year 1, they were not implemented until Year 2. Therefore, there was no influence or impact from these subcomponents on this confirmatory analysis. Family Involvement in Science (FIS) consisted of take-home booklets that included activities to engage family members in science, including vocabulary development, reading selection related to the science concept, family science activities, and science literature resources. During the spring semester of Year 2, FIS kits inclusive of FIS booklets and GoVision goggles were sent to treatment teachers to send home with consented students. During Year 2, the SRM^2 virtual mentoring component featured contributions from eight university science mentors who were strategically recruited so that their area of science field, interest, and science experiences directly aligned with LIS curriculum units. Videos of the scientists were embedded into the introductory scenarios (setting a real-life context for learning the science content), and also embedded into the closing unit activity, a science challenge that brought together the skills and content addressed in the unit. During Year 2, 19 teachers participated in SRM^2, yielding 951 student questions for scientists. The questions were synthesized, and the scientists generated responses in return. Importantly, however, this comprehensive intervention was not completely implemented fully throughout the first year.

Virtual Professional Development (VPD). During Year 1, initial onboarding VPD sessions were scheduled weekly during September, 2017. However, Hurricane Harvey adversely impacted 17 of the treatment teachers (29.8%) in six school districts. From October through the beginning of April, treatment teachers attended 90 minutes of virtual training every two weeks focused on implementation of LISTO curriculum and literacy-infused instructional strategies. On average, a total of three hours per month were reported. VPD sessions mid-April through May were related to teacher feedback, surveys, and focus group interviews. During Year 2, treatment teachers received approximately 60 minutes of virtual training every two weeks from September to April, totaling two hours per month, on average. The VPD sessions were conducted using GoToTraining, an interactive virtual platform that allows screen sharing, webcam sharing, voice chat, type chat, and breakout sessions. The VPD sessions included professional growth opportunities to develop teachers' knowledge of science content and literacy-integration, including strategies that support listening, speaking, reading and writing in science – such as vocabulary instruction, reading comprehension, oral language development, and writing in science. VPD sessions also included a preview of upcoming curriculum units, demonstrations and modeling videos, project updates, teacher feedback, and teacher spotlights.

Virtual Mentoring and Coaching (VMC). As part of the technology innovations, participating fifth grade teachers received the Applied Pedagogical Education Xtra Imaging System (APEXIS) hardware and access to the Hoot Education platform, through which VMC was conducted. Teachers participated in virtual coaching sessions in which coaches provided real-time feedback to teachers as they implemented the LISTO curriculum. Due to delays caused by Hurricane Harvey, additional time was necessary to get observation equipment in place and to provide training and ongoing supports for teachers to utilize the online platform and classroom technology. As a result, VMC was delayed until spring 2018, and monitoring fidelity of teacher implementation of the LISTO lessons did not occur during the first semester of the project. During the second semester, coaches conducted two live, real-time coaching sessions and provided written feedback to identify what went well during the lesson and areas of improvement related to lesson plan and instructional strategy implementation. Teachers were asked to reflect on the feedback. Coaches met to discuss trends observed during VMC sessions and strategically incorporated supports within the ongoing VPD sessions.

During Year 2, teachers participated in five VMC sessions including an initial goal-setting session and four real-time coaching sessions. In addition to written feedback, teachers also participated in a virtual reflection session each semester in which the teacher and coach met synchronously online to review selected time stamps of a recorded classroom observation and reflect on teacher LISTO lesson implementation and teacher-selected instructional goals.

Research Design

The evaluation of LISTO involved a multisite cluster randomized trial (CRT) designed to meet the Every Student Succeeds Act (ESSA) Tier 2 standards for “moderate” evidence, as well as the What Works Clearinghouse (WWC) standards "with reservations." The study estimated program impacts on both student and teacher outcomes and documented the fidelity of implementation and educators’ perceptions of program quality.

Schools with participating fifth-grade science teachers were randomly assigned to either the treatment or control condition. Schools were randomly assigned within district blocks when more than one school in a district chose to participate in the study. Fifth-grade science teachers may have participated in the intervention for either one or two years over the 2017–18 and 2018–19 school years, and some teachers were allowed to join the study after the random assignment of schools. Students were exposed to the intervention only in their fifth-grade year, either in the 2017–18 or 2018–19 school year. Again, data for the year 2017-2018 reflected a low fidelity of implementation for the entire first semester, due to the reasons previously discussed. The resulting impacts included delaying the beginning of year testing, curriculum implementation, baseline observations, and professional development schedules for the original confirmatory group.

Research Questions

1. What is the impact of LISTO on fifth-grade students’ science and reading achievement after one year of treatment compared with the business-as-usual condition?
2. What is the impact of LISTO on fifth-grade science teachers' instructional delivery after one or two years of treatment compared with the business-as-usual condition?
3. Was each key component of LISTO implemented with fidelity?
4. How do teachers perceive the effectiveness of the VPD, and do they perceive their practice to improve with reflections included in training?
5. How do teachers and coaches perceive the ease of use and quality of VMC using Hoot Education and APEXIS software and hardware?

Sample

Prior to the 2017–18 school year, LISTO personnel recruited 71 Texas schools in 37 school districts in which low-income students comprised more than 50% of the student population. Schools were randomized to either the treatment or control condition within district, whenever possible. Fifth-grade science teachers in participating schools were then recruited to participate in the study. For each school, up to four classes or rotations were selected to participate in the study. Students were included in the study if they were in the sampled classes, and if their parents provided consent for them to participate in the study. A total of 5,180 students were included across both years of the study.

One hundred twenty-one teachers participated in the study for 2017–18. Teachers were allowed to join the study through the beginning of the 2018–19 school year. Thirty-one teachers joined the study in 2018–19, and 69 participated for two consecutive years. Students were exposed to the program in their fifth-grade year only. This count reflects teachers who had non-missing student outcomes in either of the 2017-18 or 2018-19 school years, or had at least one observation submitted in the 2018-19 school year.

Measures and Instruments

The evaluation examined the impact of LISTO on the following student outcomes:

- State of Texas Assessments of Academic Readiness (STAAR) science
- STAAR reading
- Iowa Test of Basic Skills (ITBS) science
- Big Ideas in Science Assessment (BISA)
- Science Interest Survey

Program impacts were also estimated for researcher-made teacher outcomes including:

- Focus on academic tasks and/or student feedback while presenting new science content
- Focus on oral language while presenting new science content
- Use of research-based instructional practices while teaching science

Fidelity of program-level implementation was measured by attendance of virtual professional development and coaching sessions, and by receipt of the program's curricular materials. Perceived program quality was captured by teacher responses collected via surveys and in focus groups and interviews.

Analytic Approach

The impact of LISTO on student and teacher outcomes was estimated using hierarchical linear modeling. Propensity score weighting was also used to estimate program impacts on teacher outcomes due to large differences on the pretest measure because pretest data were collected after program implementation had begun.

Findings

Outcomes collected in the 2017–18 school year were considered to be exploratory, given the timing of Hurricane Harvey, which hit Texas in August of 2017, as mentioned earlier. Outcomes in the 2018–19 school year served as the confirmatory contrasts. In both school years, students were exposed to the program through their teachers in only their fifth-grade year. One year of exposure for students may have been insufficient to increase student achievement in science or reading, yet some impacts were observed.

Program impacts. The following program impacts should be cautiously interpreted due to limitations of delayed and incomplete implementation in the first year of the project as previously described. LISTO resulted in increased teacher capacity to implement research-based strategies while teaching science content, yet this improvement did not necessarily translate into improved student achievement in science or reading. The LISTO professional development and coaching covered pedagogical strategies for teaching science, including those that have been shown to improve literacy and be particularly effective for ELs. Findings showed that LISTO teachers implemented these research-based pedagogical strategies to a greater extent than did control teachers. The research team believes that due to impacts of Hurricane Harvey and issues with teachers submitting the first round of classroom observation recordings, there was a low return on the first round of classroom observations during Year 1. Specifically, reviewers rated eight LISTO teachers' instruction statistically significantly higher on a rubric than 22 comparison teachers' instruction after two years of treatment. At the same time, there were no observed differences between LISTO and control teachers on two other outcomes dealing with the share of instructional time spent teaching new science content while performing various activities.

In 2017–18, after the first year of program implementation, there was a statistically significant difference in science achievement for students in LISTO versus control classrooms. LISTO students scored approximately 48 points lower than did control students on the STAAR science assessment. Students in LISTO classrooms expressed slightly lower average interest in science than students in control classrooms by 0.07 points on a 5-point survey scale, or -0.14 standard deviations ($p<.05$).

In 2018–19, or after the second year of program implementation, students in LISTO classrooms had lower average science achievement on the state test than did students in control classrooms, but there were no statistically significant differences in student performance on formative science assessments. LISTO students underperformed control students on the STAAR science assessment in 2018–19 by roughly 73 points or -0.13 standard deviations ($p<.05$). There were also no differences in science interest between LISTO and control students in 2018–19. However, qualitative data collected from teachers suggested that students had improved science vocabulary as a result of LISTO participation, which led to improvements in student engagement and self-efficacy. Student interaction and engagement are higher when students interpret the activities and content to be relevant and challenging (Nguyen et al., 2018; Davis & McPartland, 2012).

There were no statistical differences in reading achievement for LISTO and control students in either study year. However, treatment teachers indicated a marked improvement in student writing, particularly with regard to scientific vocabulary. LISTO teachers reported that their students began to articulate naturally-occurring, everyday scientific processes (such as rain and the water cycle) while using the correct scientific terminology. Teachers attributed this shift directly to the expository readings in the LISTO curriculum.

Fidelity of program implementation. Fidelity of program-level implementation was measured using teacher attendance for VPD and VMC sessions, as well as evidence that curricula materials were mailed to teachers. The fidelity of implementation for each program component was analyzed separately for the 2017–18 and 2018–19 school years. Teachers were excluded from the fidelity sample if (a) they did not attend any of the VPD training sessions; (b) they (or their schools) withdrew from the study; or (c) they left their schools. In both study years, LISTO failed to meet the criterion for high fidelity as determined by teacher participation in the VPD and VMC sessions, with 80% and 74% of teachers attending 90% of VPD and VMC sessions, respectively. The 90% threshold equates to attending 15 of 17 VPD sessions that were offered and attending all four VMC sessions that were offered. Because not all LISTO teachers attended the VMD and VPD sessions with the regularity that was required for a high level of fidelity, this could have contributed to a lack of positive effects on student and teacher outcomes. Curricula materials, on the other hand, did meet its intended level of fidelity, as 100% of schools with participating LISTO teachers received the materials.

Perceived program quality. Perceived quality of the program was captured by teacher surveys and focus groups, which gathered teacher perceptions about the VPD, VMC, and curricula components. Teacher perceptions of the program were overwhelmingly positive. Responses collected from surveys and focus groups indicated that the VPD and VMC sessions were extremely useful and beneficial for teachers of all backgrounds and years of experience. The LISTO-provided curricula was particularly appreciated by first-year teachers because it provided a clear structure and pacing guide for the class. Although some teachers reported issues with the pacing and technology, participants agreed that the trainings were of high quality.

With regard to observed program effects on students, LISTO teachers reported an increase in student engagement and confidence in science-based content. Anecdotally, teachers felt that LISTO made a noticeable impact on struggling readers. The integration of technology and the literacy-infused instructional strategies fostered a more inclusive and participatory learning environment where learners interacted more with the teacher and with one another than they previously had, which empowered students in their own learning. Although the quantitative data did not show improvements on student outcomes, teachers endorsed LISTO for its ancillary benefits.

Conclusion

As previously mentioned, the first year of implementation encountered a number of delays and set-backs in full implementation for the original confirmatory group. LISTO (Valid 45), and the corresponding VPD, VMC, and curricula resources did not lead to improved student achievement in science or reading for students consented to participate in the study. There was a negative impact on students' science achievement in both 2017–18 (ES = -0.10) and in 2018–19 (ES = -0.13). There was a negative program impact on students' science interest (ES = -0.14), as measured by a survey, in 2017–18, and no impact in 2018–19. These quantitative findings were in conflict with qualitative data collected from LISTO teachers, who indicated that the program led to improvements in both science vocabulary and engagement and self-efficacy in science for students. LISTO teachers also indicated that the program had benefited their struggling readers, but there was no observed program impact on student reading achievement in either 2017–18 or 2018–19. While LISTO may have yielded some benefits for students, these benefits were not well captured on the standardized tests or survey instruments employed.

LISTO had positive effects on teacher practices for a subsample of teachers, specifically on increased delivery of research-based instruction to teach science content as rated on a rubric by external reviewers (ES = +1.12). There were no differences in two other teacher outcomes, however, focused on the share of instructional time spent teaching new science content while performing various activities.

One potential reason for the lack of observed positive effects on student outcomes was that the VPD and VMC components of the program did not meet the programmatic level of fidelity, as measured by teacher participation. Although the sessions were offered, teachers attended 90% of the VPD sessions in only 62–72% of schools and 90% of the VMC sessions in only 54–73% of schools, depending on the implementation year. Therefore, LISTO teachers may not have participated in the program to the extent needed to observe program impacts on student and teacher outcomes.

The LISTO teachers who participated in the program reported that the VPD and VMC were well-received by teachers. At times, teachers found the VPD and VMC sessions lengthy, yet the VPD allowed for greater teacher collaboration, and overall, teachers found the VPD and VMC to be very helpful and useful. The curricula were also appreciated by the teachers, with first-year teachers in particular benefitting from the pacing guides. Teachers also reported some barriers to implementation, including technological issues with the hardware and software and inadequate instructional time to fully engage in the implementation of the program.

In sum, LISTO appeared to improve instructional practices for a sample of teachers who implemented the program for two years with complete data (including the first round of classroom observation recordings that were missing among other teachers who participated for two years), but did not positively impact student or teacher outcomes more broadly. One likely reason for the lackluster effects was the relatively low levels of teacher participation in all VPD and VMC sessions that were offered, exacerbated by the disruption from the impacts from Hurricane Harvey, causing late starts in many districts during the first year. Arguably, having limited years (and here, less total program time than originally planned) to learn and implement a new curriculum reduces the capacity of teachers to perfect instructional strategies and consequently impact student achievement relative to control-group colleagues, who may employ less innovative but more familiar curricula. Likewise, the research team believes that only one year's exposure by students to novel ways of learning science in fifth grade without intervention in early grades to build the foundation could limit the development of positive attitudes or translate increases in learning quality from LISTO to higher achievement on standardized science and reading assessments.

Encouragingly, treatment teachers' overall positive reactions to the program suggest its potential to improve student affect and learning, but more extensive implementation experience by teachers and multi-year exposure by students starting early grades may be needed to yield measurable benefits. Clearly, such focuses emerge as a highly recommended topic for future research. Again, we remind that these conclusions should be interpreted with caution given the challenges presented by Hurricane Harvey described earlier in this document.

An Evaluation of the Literacy-Infused Science Using Technology Innovation Opportunity (LISTO) Validation Project

This study is an evaluation of the Literacy-Infused Science Using Technology Innovation Opportunity (LISTO) validation project (Valid 45). The LISTO project was funded by the Investing in Innovation (i3) Fund.[2] It involved a multi-year intervention that provided virtual professional development and coaching, and literacy-infused science curricula to fifth-grade science teachers who taught predominantly low-income students and in predominantly rural public schools in Texas.

Multiple professors at Texas A&M University were the recipients of the i3 grant that funded LISTO. The Center for Research and Reform in Education (CRRE) at Johns Hopkins University School of Education was the independent, third-party evaluator of LISTO. This report describes the method and findings of the evaluation.

Background

Rural school districts comprise more than 50% of all school districts in Texas.[3] In fact, Texas has more schools in rural areas (over 2,000 in SY 2013–14[4]) than any other state. Rural school districts face unique challenges, including in the recruitment and retention of highly qualified teachers (Webb, 2006). Recruitment and retention of teachers in science, technology, engineering, and mathematics (STEM) subjects may be particularly difficult (Pickrom, 2015; Monk, 2007). As a result, students in rural school districts may be less likely to receive high-quality instruction in content-areas such as science and mathematics. Rural schools face additional challenges related to professional development of current teachers, due to geographic location and limited resources (Beesley, 2011; Friedrichsen et al., 2007; Glover et al., 2016; Monk, 2007).

Scientific literacy is particularly difficult for students regardless of school location (Gee, 2005), but there is evidence that low-income students, English learners (EL), and non-White/non-Asian students face particular challenges in science; just 40% of low-income students and 35% of ELs met grade-level expectations in the 2018 State of Texas Assessments of Academic Readiness (STAAR), compared with 51% of all students in Texas.[5] Additionally, low-income students and ELs were among the lowest-achieving subgroups on Texas reading assessments. In reading, 36% of low-income students and 32% of ELs met grade-level expectations in the 2018 STAAR, compared with 46% of all students in Texas. And these populations are becoming increasingly prevalent throughout Texas.

---

[2] The award number is U411B160011.

[3] https://tea.texas.gov/sites/default/files/Texas_Rural_Schools_Spotlight_Report_2016-17%201.pdf

[4] https://nces.ed.gov/surveys/ruraled/tables/a.1.a.-2.asp

[5] https://rptsvr1.tea.texas.gov/cgi/sas/broker?_service=marykay&year4=2018&year2=18&_debug=0&single=N&batch=N&app=PUBLIC&title=2018+Texas+Academic+Performance+Reports&_program=perfrept.perfmast.sas&ptype=H&level=state&search=campname&namenum=&prgopt=2018%2Ftapr%2Fpaper_tapr.sas

Over the past decade, the percentages of low-income and EL students in Texas schools have grown steadily. The percent of low-income students increased from 56.5% of all students in 2008–09 to 60.6% of all students in 2018–19. In 2018–19, ELs accounted for approximately 19% of the K–12 student population in Texas, a 32% increase from the 2008–09 school year. Many students who are ELs are also from low-income households, which can lead to academic vulnerability.

In addition to the population growth of low-income and EL students, schools often face difficulty in the recruitment and retention of skilled teachers in rural districts. These challenges suggest teachers in rural schools may be particularly in need of additional training and resources related to teaching science, technology, engineering, and mathematics (STEM) subjects in meeting the specific academic needs of EL students and students from low-income households (Samson & Collins, 2012). By some estimates, only 30% of teachers of EL students have had the necessary training to provide effective teaching (Ballantyne, Sanderman, & Levy, 2008). A particular concern is the ability of teachers to teach subject-specific content and English language acquisition simultaneously (Correll, 2016; Lee et al., 2004; Tong et al., 2017b).

Teachers of low-income students also may need additional training in teaching subject-specific content. Students from low-income households experience an achievement gap relative to their middle- and high-income peers (Reardon, 2011; 2013), in part because they are disproportionately taught by inexperienced, out-of-field, or uncertified teachers (Peske & Haycock, 2006). Inexperienced and uncertified teachers may have less content-specific skills and knowledge than seasoned teachers who are certified in a specific content area. Teachers' content-area knowledge and their own mastery of content-specific concepts and skills impacts student achievement in the subject area (Heller et al., 2012; Lange et al., , 2012).

Taking the above information into account, it is likely that additional teacher professional development and support mechanisms are needed to help teachers meet the learning needs of their EL and low-income students (Buxton & Allexsaht-Snider, 2016; Tong et al., 2017b). Considering the challenge of recruitment and retention in rural school districts, teachers in rural school districts may particularly benefit from virtual professional development and coaching programs related to content area instruction.

Professional development can increase teacher effectiveness and positively impact student achievement when it is (a) sustained over time; (b) linked with curricula; and (c) focused on both pedagogy and academic content (Darling-Hammond & Richardson, 2009; Yoon et al., 2007). Based on prior research on teacher practices and student achievement of EL students, professional development that targets cognitive-academic language proficiency within an academic content area may be particularly appropriate (Irby et al., 2010; Lara-Alecio et al., 2009; Tong, Irby, Lara-Alecio, & Mathes, 2008; Tong, Lara-Alecio, et al., 2008; Tong et al., 2017b). Tarr et al. (2008) assert that consistency between curriculum and instruction is also important in improving outcomes for all students.

In addition to targeted professional development and instructional fidelity, coaching and mentoring also positively impact academic outcomes, teacher-student interactions, and the overall educational climate for EL students (Casteel & Ballantyne, 2010; Delaney, 2012; Pruitt

& Wallace, 2012). Coaching and mentoring may positively impact student achievement, particularly for low-income students, and especially for long-term outcomes (Hagler, 2018; Hurd et al., 2012; Miranda-Chan et al., 2016). Effective teacher mentoring and coaching provide teachers with content and pedagogical expertise, modeling of instructional strategies, and feedback on teacher practice (Pruitt & Wallace, 2012).

The LISTO project builds on evidence-based best strategies for effective professional development and coaching to help teachers improve their content area instruction. LISTO is a validation study of a previous project—Project Middle School Science (MSSELL)—developed by researchers at Texas A&M University (Tong et al., 2014; Lara-Alecio et al., 2012). Project MSSELL was a literacy-infused science instructional and curricular innovation for fifth- and sixth-grade students that was funded by the National Science Foundation. Researchers at Texas A&M evaluated effects of the MSSELL program and found promising evidence of program efficacy in increasing students' likelihood of passing formative benchmark science tests and low- and high-stakes reading assessments (Lara-Alecio et al., 2012).

An overarching goal of LISTO was to validate an expansion of Project MSSELL and analyze the impact of the program on student and teacher outcomes in rural school districts and in schools that serve a relatively large proportion of students from low-income households. The intervention is designed to improve teacher effectiveness and student outcomes through ongoing virtual professional development (VPD), virtual mentoring and coaching (VMC), and literacy-infused science curricula that incorporates best practices in teaching ELs. Therefore, the LISTO project contains the same programmatic elements as the earlier MSSELL program but is implemented in contexts that allow researchers to validate previous findings in new school contexts, including in rural and low-income schools.

Project Description

The purpose of Project LISTO was to improve the instructional capacity of science educators and to validate innovative practices and strategies that integrated literacy-infused science instruction, technology, and standards-based curriculum. LISTO provided educators with standards-aligned, literacy-infused science curricula, ongoing virtual professional development, and ongoing virtual mentoring and coaching to fifth- grade science teachers. As mentioned in the Executive Summary, the first year of the project suffered delays and incomplete implementation, primarily due to Hurricane Harvey's impact on participating school districts and teachers.

Literacy-Infused Science Using Technology Innovation Opportunities (LISTO) Curricula. Participating treatment teachers received LISTO curricular materials, which included 25 weeks of standards-aligned lesson plans, lesson scripts, related resources, and hands-on science activity supplies. Lessons were designed to be implemented within an 80-minute science block. Detailed, scripted lessons were organized using the 5E instructional model (in which at least three of the five E's – engage, explore, explain, elaborate, evaluate – were implemented in each lesson) and included embedded literacy skills to facilitate listening, speaking, reading, and writing. Some of the strategies included working in student groups, direct teaching of science academic vocabulary using visuals and student-friendly definitions, supporting reading through pre-teaching pronunciation of vocabulary and words that are challenging to decode, strategically partnering students for reading, leveled questioning, highlighting expository text features, sentence stems, graphic organizers, and integrating student use of technology via tablets.

LISTO included two sub-components: Family Involvement in Science (FIS) and Scientists as Role Models and Mentors (SRM^2). Although the intent was to implement both of these components starting in Year 1, they were not implemented until Year 2. Therefore there was no influence or impact from these subcomponents on this confirmatory analysis. Family Involvement in Science (FIS) consisted of take-home booklets that included activities to engage family members in science, including vocabulary development, reading selection related to the science concept, family science activities, and science literature resources. During the spring semester of Year 2, FIS kits inclusive of FIS booklets and GoVision goggles were sent to treatment teachers to send home with consented students. The intent of the SRM^2 component was to have university scientists meet via live, synchronous, online sessions with students; however, during the second year of the project, the interaction was limited to pre-recorded video clips embedded into lesson presentations and opportunities for students to pose questions and scientists to respond. During Year 2, the SRM^2 virtual **mentoring component utilized university science featured contributions from eight** university science mentors who were strategically recruited so that their area of science field, interest, and science study, and whose experiences, directly aligned with LIS curriculum units. Videos of the scientists were embedded in the introductory scenarios (setting a real-life context for learning the science content), and also when students encountered the science challenge (a closing unit activity that brings together the skills and content learned in the unit). During Year 2, 19 teachers participated in SRM^2, yielding 951 student questions for scientists. The questions were synthesized and the scientists generated responses.

Virtual Professional Development (VPD). During Year 1, initial onboarding VPD sessions were scheduled weekly during September, 2017. However, Hurricane Harvey adversely impacted 17 of the treatment teachers (29.8%) in six school districts. From October through the beginning of April, treatment teachers attended 90 minutes of virtual training every two weeks focused on implementation of LISTO curriculum and embedded instructional strategies. VPD sessions conducted mid-April through May were related to teacher feedback, surveys, and focus group interviews. During Year 2, treatment teachers received 60 minutes of virtual training every two weeks from September to April, on average totaling two hours per month. The VPD sessions were conducted using GoToTraining, an interactive virtual platform that allows screen sharing, webcam sharing, voice chat, type chat, and breakout sessions. The VPD sessions included professional growth opportunities to develop teachers' knowledge of science content and literacy-integration, including strategies that support listening, speaking, reading, and writing in science – such as vocabulary instruction, reading comprehension, oral language development, and writing in science. VPD sessions also included a preview of upcoming curriculum units, demonstrations and modeling videos, project updates, teacher feedback, and teacher spotlights.

Virtual Mentoring and Coaching (VMC). As part of the technology innovations, participating fifth-grade teachers received the Applied Pedagogical Education Xtra Imaging System (APEXIS) hardware and access to the Hoot Education platform, through which VMC was conducted. Teachers participated in virtual coaching sessions in which coaches provided real-time feedback to teachers as they implemented the LIS curriculum. Due to delays caused by

Hurricane Harvey, it took additional time to get observation equipment in place and to provide training and ongoing supports for teachers to utilize the online platform and classroom technology, VMC was delayed until spring 2018. Therefore, monitoring fidelity of teacher implementation of the LISTO lessons did not occur during the first semester of the project, and teachers were not given feedback during the first semester on their LISTO lesson implementation. During the second semester, coaches conducted two live, real-time coaching sessions and provided written feedback to identify what went well during the lesson and areas of improvement related to lesson plan and instructional strategy implementation. Teachers were asked to reflect on the feedback. Coaches met to discuss trends observed during VMC sessions, and strategically incorporated supports within the ongoing VPD sessions. During Year 2, teachers participated in five VMC sessions including an initial goal setting session and four real-time coaching sessions. In addition to written feedback, teachers also participated in a virtual reflection session each semester in which the teacher and coach met synchronously online to review selected time stamps of a recorded classroom observation and reflect on teacher LISTO lesson implementation and teacher-selected instructional goals.

Evaluation Design

The evaluation of LISTO involved a multisite cluster randomized trial (CRT) designed to meet the Every Student Succeeds Act (ESSA) Tier 2 standards for “moderate” evidence, as well as the What Works Clearinghouse (WWC) standards "with reservations." The study estimated program impacts on both student and teacher outcomes and documented the fidelity of implementation and educators’ perceptions of program quality.

Schools with participating fifth-grade science teachers were randomly assigned to either the treatment or control condition. Schools were randomly assigned within district blocks, when more than one school in a district chose to participate in the study. Fifth-grade science teachers may have participated in the intervention for either one or two years over the 2017–18 and 2018–19 school years, and some teachers were allowed to join the study after the random assignment of schools. Students were exposed to the intervention only in their fifth-grade year, either in the 2017–18 or 2018–19 school year.

LISTO is expected to produce positive outcomes for student and teacher outcomes after two years of professional development supports. The confirmatory contrasts for student outcomes estimated the impact of LISTO on student achievement in science and reading (as measured by the state-mandated STAAR assessments) in the second year of the study (2018–19) and after one year of treatment for students. The confirmatory contrasts for teacher outcomes estimated the impact of LISTO in the second year of the study (2018–19) and after either one or two years of treatment for teachers, depending on when they joined the study. The teacher outcomes were the amount of instructional time teachers spent presenting new science information (in English) while (a) students performed an academic task and/or teachers evaluated the accuracy of student responses, and (b) the class was engaged in listening and/or speaking (as opposed to reading and writing).

Research Questions

1. What is the impact of LISTO on fifth-grade students' science and reading achievement after one year of treatment compared with the business-as-usual condition?
2. What is the impact of LISTO on fifth-grade science teachers' instructional delivery after one or two years of treatment compared with the business-as-usual condition?
3. Was each key component of LISTO implemented with fidelity?
4. How do teachers perceive the effectiveness of the VPD, and do they perceive their practice to improve with reflections included in training?
5. How do teachers and coaches perceive the ease of use and quality of VMC using Hoot Education and APEXIS software and hardware?

Methods

Sample

Prior to the 2017–18 school year, the grantee recruited 71 Texas schools in 37 school districts in which low-income students comprised more than 50% of the student population. Schools were randomized to either the treatment or control condition within district, whenever possible. For seven districts, schools were randomized to either treatment or control within district. For the remaining 30 districts, there was only one participating school per district, and schools were randomized to either the treatment or control condition. Table 1 shows the results of the random assignment of schools.

Table 1
Results of the school random assignment

	Total	Rural	Non-Rural
Treatment school N	35	23	12
Control school N	36	24	12
District N	37	33	4

NOTE—Two districts and three schools left the study prior to implementation due to changes in district administration.

Fifth-grade science teachers in participating schools were then recruited to participate in the study. Initially, a maximum of two teachers per school were recruited to participate. Because a number of rural schools had only one fifth-grade science teacher and there were fewer numbers of teachers than expected, ultimately, all fifth-grade science teachers in rural schools were offered participation in the study. In non-rural schools, up to two fifth-grade science teachers were invited to participate in the study. Given teacher turnover, new teachers were also allowed to join the study after the start of the 2017–18 school year and through the beginning of the 2018–19 school year. One hundred twenty-one teachers participated in the study for 2017–18, 31 teachers joined the study in 2018–19, and 69 participated for two consecutive years. Students were exposed to the program in their fifth-grade year only.[6] This count reflects teachers who had non-missing student outcomes in either of the 2017-18 or 2018-19 school years, or had at least one observation submitted in the 2018-19 school year.

[6] Some teachers were not included in the student and teacher impact analyses, however, due to missing data.

For each school, up to four classes or rotations were selected to participate in the study. The grant could not support providing the intervention to all fifth-grade science classes in study schools. For schools with two fifth-grade science teachers participating in the study, two classes or rotations per teacher were selected to participate in the study. For schools with more than two fifth-grade science teachers participating in the study, one class or rotation per teacher was selected to participate. For schools where study teachers had only one class (e.g., not departmentalized), all of the teacher's students were included in the study.

Students were included in the study if they were in the sampled classes, and if their parents provided consent for them to participate in the study. The student sample was also narrowed to the students who had non-missing test scores on both the pretest and posttest. Similarly, teachers were included in the impact analyses on teacher outcomes when teachers had non-missing observational scores at both the pre- and post-intervention time points. Given potential bias due to non-random selection of participating teachers and students from study schools, baseline equivalence on the pretest measures for each analytic sample was assessed (WWC, 2020).

Table 2 outlines the characteristics of the teacher sample. Note that there were two teacher samples, one for the analyses on student outcomes, and a second for the analyses on teacher outcomes. LISTO and control teachers were relatively similar in terms of background characteristics, although background characteristics were unavailable for roughly one-third to one-half of participating teachers. There were no statistically significant differences in teacher characteristics between the LISTO and control groups for either teachers or their students. The statistical models controlled for alternative certification, as it appeared to be an explanatory covariate.

Table 2
Characteristics of the teacher sample

	Analyses on Student Outcomes			Analyses on Teacher Outcomes		
Characteristics	Total	LISTO	Control	Total	LISTO	Control
Female	77.85%	80.00%	75.68%	73.58%	78.26%	70.77%
Science teacher	86.97%	89.86%	83.74%	95.35%	94.44%	96.00%
Certification Alternative	42.41%	43.24%	41.55%	46.15%	47.83%	44.83%
Science	9.85%	10.45%	9.23%	17.65%	17.39%	17.86%
ESL	29.55%	29.85%	29.23%	27.45%	30.43%	25.00%
Bilingual	28.79%	29.85%	27.69%	33.33%	39.13%	28.57%
Average years teaching	Total	LISTO	Control	Total	LISTO	Control
All	10.05	10.81	9.27	10.51	11.57	9.70
Science	6.26	6.24	6.27	7.85	8.96	6.97
5th grade	4.47	4.12	4.83	5.35	5.43	5.28
N	219	100	119	71	33	38

NOTES—1. Descriptive statistics for teachers were based on the analytic samples. Teacher characteristics for the student outcomes analyses were based on the combined analytic samples across the 2017–18 and 2018–19 school years. Teacher characteristics for the teacher outcomes analyses were based on the 2018–19 year only. 2. Teacher characteristics were missing for approximately one-third to one-half of teachers, depending on the characteristic and sample.

Next, we outline characteristics of the student sample. As shown in Table 3, the majority (75.36%) of students were low-income, and about one-third (32.58%) were English learners (ELs). Additionally, the majority (73.67%) of students were Latino, with smaller percentages of White (15.75%) and Black (7.42%) students. Therefore, the student sample reflected the grant priorities to serve low-income students, many of whom were ELs.

Table 3
Characteristics of the student sample

Characteristics	Total (%)	LISTO (%)	Control (%)
Low-income	75.36	78.36	71.91
English learner (EL)	32.58	34.95	29.84
Reclassified EL	2.94	3.36	2.46
Migrant	2.44	2.53	2.33
Special education	7.84	7.69	8.02
504 plan	8.95	8.86	9.04
Female	49.96	49.61	50.38
Latino	73.67	72.92	74.54
White	15.75	15.54	16.00
Black	7.42	7.85	6.91
More than one race	2.46	3.48	1.28
Other race	0.70	0.20	1.28
N	5,180	2,790	2,390

NOTE—Descriptive statistics were calculated for the combined analytic sample across the 2017–18 and 2018–19 school years.

While LISTO and control students were similar in terms of demographic characteristics, there were a few small differences between the two groups of students. A larger percentage of LISTO students were low-income (78.36%) relative to control students (71.91%). In addition, a larger percentage of LISTO students were English learners (34.95%) compared with control students (29.84%). The statistical analysis controlled for all of these student characteristics, as well as baseline achievement.

Measures and Instruments

Student outcomes. The evaluation estimated the impact of LISTO on student performance in science and reading using the following assessments and instruments:

- *State of Texas Assessments of Academic Readiness (STAAR) science* (Texas Education Agency, 2017a): The science test measures student knowledge of science concepts and scientific processes and is administered each spring to all students in Texas in the fifth and eighth grades. This test is primarily administered in English but was administered in Spanish to 0.40% of students in the study.

- *STAAR reading* (Texas Education Agency, 2017b): The reading test measures grade-level reading expectations, including students' critical thinking, inferencing, making connections, understanding, and application in different genres of reading. STAAR reading is administered each spring to all students in Texas in grades 3–8. The test was administered in Spanish to about 2% of students in the study.

- *Iowa Test of Basic Skills (ITBS) science* (Dunbar & Welch, 2015): The science subtest measures student knowledge of science concepts. This test was administered to fifth-grade students by trained testers,[7] in the fall and spring of each study year (e.g., both prior to program implementation and after one year of treatment).

- *Big Ideas in Science Assessment (BISA)* (Lara-Alecio et al., 2018): This instrument measures disciplinary core ideas in both the Next Generation Science Standards and Texas science standards. The instrument was developed by researchers at Texas A&M University, and has internal reliability of .70 (Lara-Alecio et al., 2018). The instrument was administered to students in both the fall and spring of each study year.

- *Science interest survey*: This 5-point Likert scale instrument gauges student motivation and self-efficacy to learn science. It also contains science-related items about family encouragement, teacher efficacy, and English comprehension. The instrument was developed by researchers at Texas A&M University, and was found to have an internal reliability of .86 (Tong et al., 2020). The survey was administered to students in both the fall and spring of each study year.

Student scores on the STAAR science and reading tests in spring 2019 served as the confirmatory contrasts. The remaining student assessments and assessments administered in spring 2018 were analyzed for exploratory purposes. For nearly all student outcomes, the same instrument was used as both the pretest and posttest measure. The one exception is that the pretest for the STAAR science was the ITBS science test administered in the fall of fifth grade, since STAAR science is not administered to students in the fourth grade.

LISTO project personnel at Texas A&M University were responsible for data collection, processing, and scoring. Data were then transferred to the CRRE evaluation team, and the evaluation team checked, merged, and analyzed the data.

Teacher outcomes. Teacher outcomes for this impact study were improved instructional delivery per pedagogical transitional bilingual theory. Teacher outcomes were assessed using the following instruments:

- *Science Teacher Observation Record (STOR)* (Lara-Alecio et al., 2012): The STOR was developed by researchers at Texas A&M University and documents the extent to which teachers implement best practices while teaching science content, particularly to ELs. The STOR asked raters to rate teachers on approximately 10 items that capture teacher preparation for and delivery of science instruction.[8] Topics included: teacher and material preparation; lesson pacing; technology utilization; questioning strategies; opportunities for student writing and reading in science; connections to prior knowledge; reading comprehension supports; use of scientific inquiry; and student reflection. The STOR used

[7] Testers were hired by CRRE and trained by LISTO project personnel.
[8] The inter-rater reliability of STOR was 0.86 (Lara-Alecio et al., 2012).

a 4-point scale in 2017–18 and a 5-point scale in 2018–19, and scores were created by CRRE.[9]

- *Transitional Bilingual Observation Protocol (TBOP)* (Lara-Alecio et al., 2009): The TBOP was previously developed and validated by researchers at Texas A&M University from the four-dimensional bilingual pedagogical classroom theory (Lara-Alecio & Parker, 1994). TBOP captures certain pedagogical behaviors (e.g., language of instruction, language content, activity structure, communication mode, English as a second language (ESL) strategies, etc.) during classroom instruction (Lara-Alecio et al., 2009; Tong et al., 2017b). The TBOP asks raters to record the frequency of such behaviors; therefore, the TBOP score denoted the proportion of instructional time the teacher demonstrated the particular behavior.[10] Frequency data were provided to the CRRE by Texas A&M University, and the CRRE calculated teachers' TBOP scores. TBOP scores were used to document changes in teacher practices over time. The two domains of interest for this study were the proportion of time the teacher spent presenting new science content while (a) teachers were overseeing students perform an academic task or evaluating the accuracy of student responses, and (b) teachers explicitly focused on academic oral language.

All teachers, treatment and control, were observed by trained observers three times annually and rated on both the TBOP and STOR instruments. LISTO project personnel were extensively trained on the instruments by Texas A&M University researchers and then observed and scored teachers virtually using videos of classroom practice. Observations occurred at the beginning, middle, and end of the school years. The first round of observations occurred approximately 1-2 months after program implementation began, typically 1-2 weeks after completion of student consent and baseline assessments.

Teachers' TBOP scores and STOR ratings were not analyzed for the 2017–18 school year. Due to Hurricane Harvey, many teachers did not submit their instructional videos, and therefore, these data were missing for most teachers. Note, however, that the scores from fall 2017 were used as the pretest when not missing; otherwise, scores from fall 2018 were used as the pretest. Scores from the final observation in spring 2019 were used as the confirmatory contrast.

Fidelity of implementation. Fidelity of implementation was measured using teacher attendance for virtual professional development and coaching sessions, as well as evidence that curricula materials were mailed to teachers. Perceived quality of the program was also captured by teacher perceptions about the professional development, curriculum materials, and coaching. Two qualitative data sources were used to capture teacher perceptions about program quality:

---

[9] Scores were created by calculating the mean rating across all items. There was no item-level missing values for teachers who had non-missing STOR scores.

[10] Prior studies have found inter-rater agreement using the TBOP ranging from 0.65 to 0.98 in Kappa values (Bruce, Lara-Alecio, Parker, Hasbrouck, Weaver, & Irby, 1997; Breunig, 1998; Irby et al., 2007; Irby et al., 2010). However, given the multi-dimension-multi-rater nature of the instrument, a more rigorous process was developed to establish inter-rater reliability (IRR) using Gwet's (2012) AC_1 coefficient; the IRR using this approach ranged from .724 to .945 (Tong et al., 2017a).

- *Teacher surveys.* At the end of each school year, researchers at Texas A&M University administered online surveys to treatment teachers. Using a combination of Likert-type and open-ended questions, the survey asked teachers to rate their experiences with the Virtual Professional Development (VPD) sessions. A total of 49 teachers completed the survey in year one; 37 teachers participated in year two.

- *Teacher focus groups.* Texas A&M University researchers conducted virtual focus groups for treatment teachers in May of each school year. Facilitators used video conferencing software to conduct interviews that lasted approximately 45 minutes. The protocols asked teachers to provide their perceptions of LISTO on student engagement and academic development, as well as the quality of program curricula, professional development, and coaching. In year one, a total of 20 teachers participated in seven different focus groups; there was a total of 30 teacher participants in eight different focus groups in year two.

Analytic Approach

Impact study. The impact of LISTO on student and teacher outcomes was estimated using hierarchical linear modeling. Propensity score weighting was also used to estimate program impacts on teacher outcomes due to large differences on the pretest measure.

Hierarchical linear modeling. The impacts of LISTO on student and teacher outcomes were estimated separately by school year. Due to Hurricane Harvey in the summer of 2017, the first year of LISTO implementation became more of a pilot year, and confirmatory contrasts were conducted on outcomes collected in spring 2019. Impacts of LISTO were estimated using a hierarchical linear model with students or teachers nested within schools (Raudenbush & Bryk, 2002). The model to estimate impacts of LISTO on student outcomes was as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{01} treatment_j + \gamma_{10} pretest_{ij} + \gamma_{20} X_{ij} + \gamma_{02} Y_j + u_{0j} + r_{ij}$$

where:
Y_{ij}: Test score for student i in school j
γ_{00}: Grand mean for students in control condition
γ_{01}: Average treatment effect
$Treatment_j$: Treatment indicator for school j
γ_{10}: Regression coefficient for the pretest
$pretest_{ij}$: Pretest score for student i in school j
γ_{02}: Vector of regression coefficients for student covariates
X_{ij}: Vector of student covariates (outlined in the appendix)
γ_{02}: Vector of regression coefficients for the district dummy indicators
Y_j: Vector of district dummy indicators for school j
u_{0j}: Random school effect for school j
r_{ij}: Residual for student i in school j

The model to estimate the impacts of LISTO on teacher outcomes was identical to the one above, except that teachers (instead of students) were the unit of analysis. This model controlled for alternative certification of teachers and the pretest.[11] The independent variables, except for the treatment indicator, were grand-mean centered to facilitate interpretation of the intercept (Enders & Tofighi, 2007).

For all models, students or teachers were included in the analysis if they had non-missing pretest and outcome scores. Students or teachers with missing background variables were included in the analysis, using a simple imputation method for missing values and dummy indicators (WWC, 2020).

Similar hierarchical linear models—without the covariates or district dummy indicators—were used to estimate baseline equivalence on each pretest measure for each analytic sample. Baseline equivalence was satisfied (≤ 0.25 standard deviations) for all student and teacher outcomes, after applying propensity score weighting for teacher outcomes (WWC, 2020).

Propensity score weighting. Baseline equivalence was not satisfied for the teacher analytic samples (> 0.25 standard deviations) because the pretests were collected after treatment had already begun. To account for these baseline differences, propensity score weighting was incorporated into the hierarchical linear model outlined above for teacher outcomes—both in models estimating program impacts and in models estimating baseline differences between treatment and control groups. Propensity score weighting was designed to make the weighted samples equivalent on the pretest measure (WWC, 2020).

To obtain the propensity score weights and calculate the average treatment effect for the treated (ATT), we first regressed the logit of treatment group assignment on the pretest. Then, propensity score weights were calculated using $weight = 1$ for the treatment group and $weight = \frac{probability}{1-probability}$ where probability is the likelihood of being in the treatment group. Propensity scores and weights were determined separately for each outcome measure and analytic sample to achieve baseline equivalence.[12]

Implementation study. To determine whether LISTO was implemented with fidelity, we analyzed the percentage of teachers and schools who participated at high levels of fidelity in each of the key program components—virtual teacher professional development (VPD), virtual mentoring and coaching (VMC), and distribution of curricula materials (LIS). High fidelity was determined based on the criteria in Table 4.

Table 4
Criteria for high fidelity of implementation

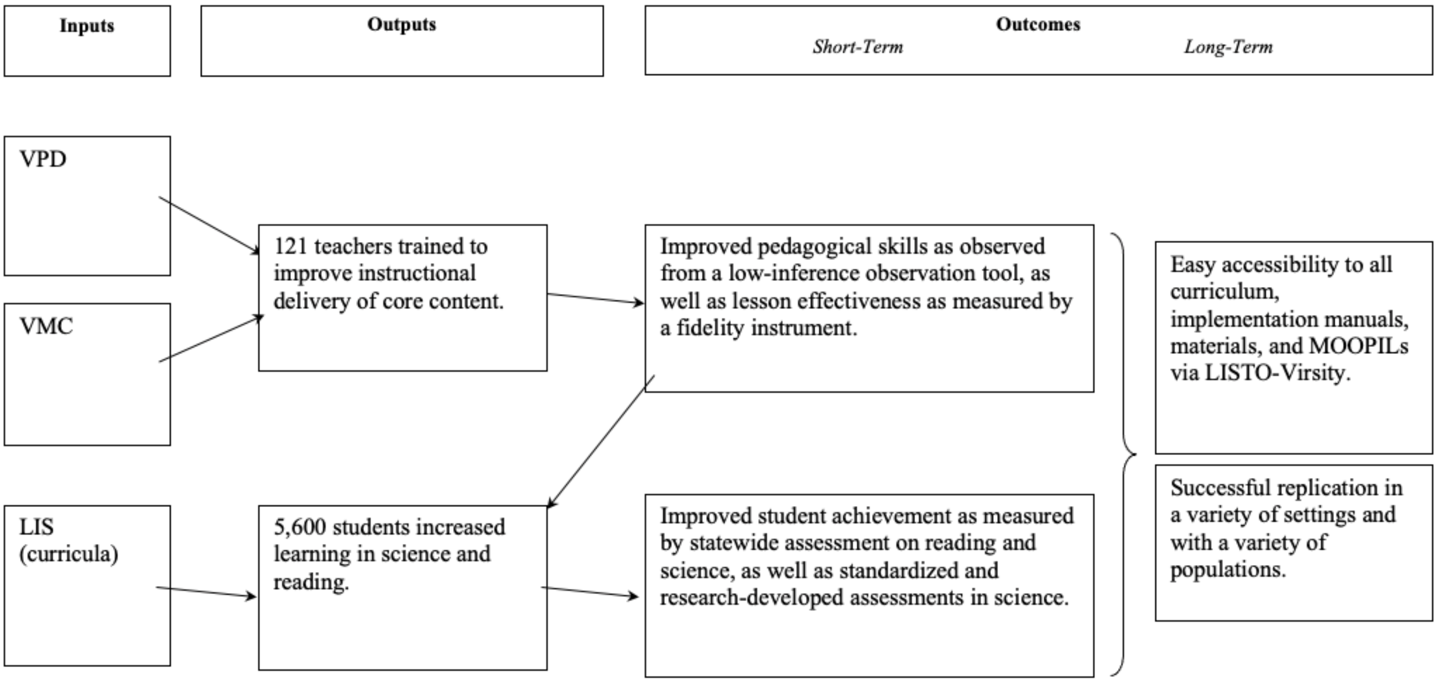
[11] For each teacher outcome, the pretest used the same instrument as the outcome but was administered at an earlier time point. The pretest was the score from fall 2017, and for Year 1 teachers with missing pretest data and all teachers who joined in Year 2, the pretest was the score from fall 2018. The only exception was for STOR; due to large baseline differences in LISTO and comparison teachers in fall 2018, only the pretest from fall 2017 was used.
[12] To incorporate propensity score weights into the hierarchical linear model, we used Stata with the [pweight=weight] option in the level-1 model. We also used Stata's svy command to calculate the means and standard deviations of the pretest and posttest scores.

Key Program Component	Data Source	Definition of High Fidelity (Teacher Level)	Definition of High Fidelity (School Level)	Definition of High Fidelity (Sample Level)
Virtual Professional Development (VPD)	Teacher training attendance record	Teacher participates in at least 90% of PD sessions	100% of participating teachers have high fidelity	At least 90% of schools have high fidelity
Virtual Mentoring and Coaching (VMC)	Coach observation feedback rubric	Teacher participates in at least 90% of coaching sessions	100% of participating teachers have high fidelity	At least 90% of schools have high fidelity
Curricular Materials (LIS)	Delivery receipts	Teacher receives curriculum	100% of participating teachers receive curriculum	At least 90% of schools have high fidelity

Fidelity of VPD, VMC, and curricular materials were measured at the teacher, school, and sample levels. VPD was considered to have been implemented with fidelity in a school if all treatment teachers in the school participated in 90% of the professional development sessions, which equated to attendance in at least 15 of the 17 sessions. VMC was considered to have been implemented with fidelity in a school if all treatment teachers in the school participated in 90% or more of the coaching sessions, which equaled attendance in all four sessions offered. The distribution of curricular materials was considered to be implemented with fidelity if the school received the curriculum materials. At the program component level, 90% of schools had to have achieved high fidelity for the program component to be implemented with fidelity at the sample level.

The fidelity of implementation for each program component was analyzed separately for the 2017–18 and 2018–19 school years. Teachers were excluded from the fidelity sample if (a) they did not attend any of the VPD training sessions; (b) they (or their schools) withdrew from the study; or (c) they left their schools. The key components of LISTO and how they theoretically relate to outcomes are detailed in the logic model, as shown in Figure 1.

Figure 1
LISTO logic model



There are three inputs: Virtual Professional Development (VPD), Virtual Mentor Coaching (VMC), and Curricular Materials. The output for VPD and VMC is to train 121 teachers to improve instructional delivery. The output for the curricular materials is to increase learning in science and reading for 5,600 students. The short-term outcomes are to improve pedagogical skills as observed from a low-inference observation tool, as well as lesson effectiveness as measured by a fidelity instrument. Improved student achievement as measured by statewide assessment on reading and science, as well as standardized and research-developed assessments in science is also a short-term outcome. Long-term outcomes include easy accessibility to all curriculum, implementation manuals, materials, and MOOPILs via LISTO-Virsity and successful replication in a variety of settings and with a variety of populations.

Qualitative data sources—treatment and control teacher surveys and treatment teacher focus groups—were analyzed thematically. The analyst initially reviewed the data, searching for recurring themes in participants' responses; these themes were cross-referenced with data from teacher surveys, and the findings were categorized and reported by theme.

Findings

Program Impacts

The following program impacts should be interpreted cautiously due to the aforementioned limitations of delayed and incomplete implementation during the first year, as the baseline year of the project. LISTO resulted in increased teacher capacity to implement research-based strategies while teaching science content, yet this improvement did not necessarily translate into improved student achievement in science or reading. The LISTO professional development and coaching supplied teachers with pedagogical strategies for teaching science, including those that have been shown to improve literacy and be particularly effective for ELs. Findings showed that LISTO teachers implemented these research-based pedagogical strategies to a greater extent than did control teachers. Despite a number of barriers to implementation, LISTO was directly responsible for benefitting teachers' instructional practices, especially those who implemented LISTO with more fidelity.

There was a statistically significant difference in science achievement on the STAAR science assessments for students in LISTO versus control classrooms in 2017–18. Students in LISTO classrooms also expressed slightly lower average interest in science than students in control classrooms. In 2018–19, students in LISTO classrooms had lower average science achievement on the state test than did students in control classrooms, as well as average lower BISA scores. However, qualitative data collected from treatment teachers suggested that students had improved science vocabulary as a result of LISTO participation, which led to improvements in student self-efficacy and engagement. There were no differences in reading achievement for LISTO and control students in either study year.

Science achievement. Fifth-grade students in LISTO classrooms did not outperform similar, control peers on the state accountability science test (e.g., STAAR science), or on formative science assessments (e.g., ITBS science, BISA) in either the 2017–18 or 2018–19 school years. There was a statistically significant difference in science achievement between LISTO and control students in 2017–18 ($p<.05$) on the STAAR science assessment, with LISTO students underperforming control students by about 48 points. LISTO students underperformed control students on the STAAR science test in 2018–19 by roughly 73 points or -0.13 standard deviations ($p<.05$), but there were no statistically significant differences in student performance on formative science assessments in that year.

Table 5 shows the impacts of LISTO on student outcomes in science relative to control students. Specifically, the table outlines the unadjusted mean for the control students, impact estimate, standard error of the estimate (SE), p-value of the impact estimate, and standardized effect size. The standardized effect size provides the effect of LISTO in terms of standard deviations.

Table 5
Estimated impacts of LISTO on science outcomes

Outcome	Unadjusted control mean	Impact estimate	Standard error	P-value	Std. effect size
2017–18					
STAAR science	3841.79	-48.15*	24.50	0.049	-0.10
ITBS science	213.64	-0.90	1.56	0.566	-0.03
BISA	19.92	-0.17	0.29	0.548	-0.03
Science interest	3.19	-0.07*	0.03	0.012	-0.14
2018–19					
STAAR science	3904.85	-72.67*	35.58	0.041	-0.13
ITBS science	213.28	-2.15	1.78	0.226	-0.07
BISA	17.17	-0.34	0.41	0.413	-0.06
Science interest	3.08	-0.02	0.02	0.285	-0.06

NOTE—*$p<.05$, **$p<.01$.

LISTO students had slightly lower average interest in science (determined by a student survey) than control students in 2017–18 by 0.07 points on a 5-point survey scale, or -0.14 standard deviations ($p<.05$). There was no statistically significant difference in science interest between LISTO and control students in 2018–19. Across both years, there was a statistically significant difference in science achievement for LISTO and control students. Directionally, results generally showed negative program effects in science achievement and interest.

Outcomes collected in the 2017–18 school year were considered to be exploratory, given the timing of Hurricane Harvey, which hit Texas in August of 2017. Outcomes in the 2018–19 school year served as the confirmatory contrasts. In both school years, students were exposed to the program through their teachers in only their fifth-grade year. One year of exposure for students may have been insufficient to increase student achievement in science.

Science vocabulary. While quantitative data did not yield positive impacts of LISTO on students' science achievement, qualitative data collected from LISTO teachers via focus groups and interviews indicated that teachers believed students had improved in their knowledge of science vocabulary as a result of participating in LISTO. The most frequently cited response from teachers was that LISTO directly impacted the way that students talk about science both academically and conversationally. Specifically, the literacy component of LISTO provided students with a common language that included grade-level, aligned, academic scientific vocabulary terms. Prior to this, one teacher explained, "They know what's happening outside, but they don't realize that it is actually related to science. They see the rain, but they don't realize it's a process." In turn, the literacy component "…is a big deal because it helps make the connection from what they're seeing to text." By experiencing science through a narrative lens – that is, learning about scientific concepts and vocabulary through reading activities – students were able to grasp concepts in more authentic ways that were meaningful.

LISTO teacher respondents also noted that the literacy-infused strategies improved students' scientific writing. One teacher noted that students gradually integrated scientific vocabulary into their writing, "…almost two times more often than my other two [non-LISTO] classes," and that this progression in writing "…just started to become natural." The literacy-infused instruction helped students to elaborate in their writing, as observed by one teacher: "I saw my students adding a whole lot more detail and more explanation than they used to know, and they would use the correct academic vocabulary." Some teachers found problems with the LISTO vocabulary, saying that it was "too advanced," and that students had difficulty connecting the reading passages with the vocabulary terms. Still, teachers generally acknowledged the benefits of literacy-based science instruction with a focus on science vocabulary, particularly for struggling readers.

Reading achievement. Improving student literacy was another focus of LISTO, in addition to increasing students' science achievement. There were no statistically significant differences between LISTO and control students on the state reading assessment (e.g., STAAR) in either school year. Directionally, LISTO students had higher average scores on STAAR reading than control students, controlling for student characteristics, but these differences were small and not statistically significant.

As shown in Table 6, LISTO students had higher STAAR reading achievement by an average of 2.65 points in 2017–18 and 4.09 points in 2018–19. The standard errors of these estimates were large, and therefore, these differences were not statistically significant. The differences translated into effect sizes of +0.02 in 2017–18 and +0.03 in 2018–19.

Table 6
Estimated impacts of LISTO on reading outcomes

Outcome	Unadjusted control mean	Impact estimate	Standard error	P-value	Std. effect size
2017–18					
STAAR reading	1558.46	2.65	5.01	0.597	0.02
2018–19					
STAAR reading	1564.06	4.09	5.73	0.476	0.03

NOTE—There were no statistically significant differences.

As noted above, outcomes from the 2017–18 school year were exploratory, given the timing of Hurricane Harvey, and outcomes in the 2018–19 school year served as the confirmatory contrasts.

The qualitative interview and focus group data indicated that numerous LISTO teachers found that the program instilled confidence in reading for their struggling readers. LISTO introduced new approaches to teaching reading, such as placing an emphasis on the features of a text. Some teacher comments included:

> *The biggest change I saw was the reading with confidence.*
>
> *I have very low, struggling readers… They don't like to read in front of anybody, but because they were paired up…they were eager to read and work together… They really enjoyed it.*
>
> *Even my low students, who were embarrassed to read in front of the class before [LISTO] – it helped them out a lot.*
>
> *[LISTO] really helped my low readers.*

Clearly, the literacy-infused strategies had a distinguished effect on struggling readers, but teachers found that advanced readers also favored the science-related readings over a traditional science textbook. In sum, teachers indicated that LISTO improved the confidence of struggling readers, as well as increased engagement in reading for all students.

Teacher outcomes. With teacher outcomes the primary goal of LISTO, the evaluation team analyzed program impact on teachers' instructional delivery and found improvements in teachers' capacity to implement research-based strategies while teaching science content. Specifically, LISTO teachers outperformed control teachers by 0.45 points (out of 5 points) on the STOR instrument ($p<.05$), which translated into an effect size of +1.12. These findings indicate that of the teachers who participated in two continuous years (both treatment and control), the treatment teachers yielded increased quality of science lesson delivery (e.g., teacher and material preparation; lesson pacing; technology utilization; questioning strategies; opportunities for student writing and reading in science; connections to prior knowledge; reading comprehension supports; use of scientific inquiry; and student reflection). Due to impacts of Hurricane Harvey and issues with teachers submitting the first round of classroom observations, there was a low return on the first round of classroom observations during Year 1; therefore, this finding should be interpreted with caution, given the relatively small sample size of eight LISTO teachers and 22 comparison teachers. However, there were no significant differences between

LISTO and control teachers' focus on academic tasks, student feedback, or oral language when presenting new science content. Table 7 outlines these findings.

Table 7
Estimated impacts of LISTO on teacher outcomes

Outcome	Unadjusted control mean	Impact estimate	Standard error	P-value	Std. effect size
2018–19					
TBOP (share of instructional time spent teaching new science content while students performed academic task or received feedback)	0.10	-0.02	0.04	0.549	-0.20
TBOP (share of instructional time spent teaching new science content with an explicit focus on oral language)	0.22	-0.05	0.07	0.469	-0.27
STOR (research-based practices when teaching science)	2.70	0.45	0.18	0.012	+1.12*

NOTES—All models also incorporated propensity score weighting to establish baseline equivalence. Treatment teachers were exposed to the intervention prior to the baseline measure.

*p<.05.

Program impact on teacher outcomes was estimated for the 2018–19 school year only. While teacher outcomes were collected during the 2017–18 school year, the response rate was low due to Hurricane Harvey. Therefore, teacher outcomes for the first year of implementation were not analyzed as part of the study.

Fidelity of Program Implementation

LISTO included three major program components: virtual professional development (VPD), virtual mentoring and coaching (VMC), and literacy-infused science using technology opportunities curricula (LISTO). The VPD and VMC components were made available to all participating treatment teachers. Fidelity of VPD, VMC, and curricular materials were each measured at teacher, school, and component levels (see Table 4). High fidelity for each program component was defined at the sample level and if 90% of participating schools had high fidelity, as outlined in Table 4.

Programmatic fidelity was measured in this study via VPD and VMC fidelity as determined by teacher attendance rates, and programmatic fidelity of implementation was measured by the timely acquisition and delivery of curricular materials as determined by delivery receipts of materials. At the individual teacher level, participation VPD and VMC failed to meet the fidelity threshold in either year of program implementation (2017–18 or 2018–19). Overall, program fidelity could not be achieved because of a lack of observation data and delayed onset of all components of the intervention due to the effects from Hurricane Harvey. Between 77-80% of teachers participated with high fidelity of the VPD, and between 70-74% of teachers participated in the VMC with high fidelity. Similarly, at the school level, VPD and VMC also did not meet the teacher attendance threshold of fidelity in either year of program implementation. Between 62–72% of schools had high fidelity of participation in the VPD, and between 54–73% of schools had high fidelity of participation in the VMC, depending on the school year. These percentages fell short of the high fidelity criterion for these two key program components VPD and VMC. The delivery of curricular materials was met with high fidelity in both of the 2017–18 and 2018–19 school years, however. Therefore, this key program component (LIS) was implemented with fidelity in both study years. Table 8 summarizes the fidelity for each program component by implementation year.

Table 8
Fidelity of implementation for each of the three components

Implementation Year	Key Component	Sample Size	Fidelity Score	Implemented With Fidelity?
2017-18	VPD	44 teachers	80%	N
		32 schools	72%	N
	VMC	42 teachers	74%	N
		33 schools	73%	N
	LIS	32 schools	100%	Y
2018-19	VPD	33 teachers	77%	N
		26 schools	62%	N
	VMC	30 teachers	70%	N
		24 schools	54%	N
	LIS	26 schools	100%	Y

The low levels of teacher participation in VMC and VPD might be explained in Year 1 due to a highly disruptive weather event, Hurricane Harvey, which interrupted the school year and likely impacted program fidelity. However, Year 2 saw an even further decline, particularly in the school level of teacher participation in VMC and VPD; taken together, the low levels of implementation at the teacher and school levels might be an explanatory factor in the results of the first two years of the LISTO program.

Perceived Program Quality

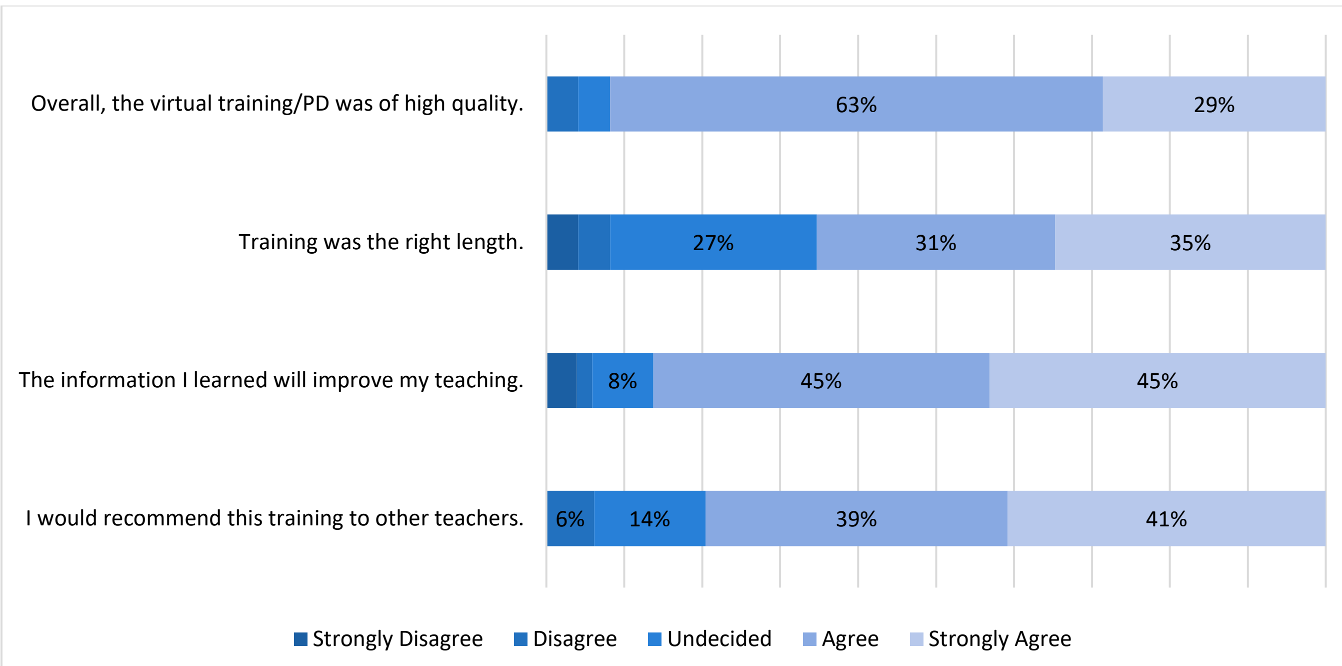
Teacher focus groups and interviews were conducted and teacher surveys were administered in order to understand teacher perceptions of LISTO and the professional development and coaching associated with it. LISTO teachers were also asked to identify various challenges with implementing LISTO and provide recommendations for program improvement. The focus group and interview protocols differed slightly between Year 1 and Year 2 cohorts (see Appendix B), but generally, the participants were asked to comment on their personal experiences with LISTO professional development and coaching; the perceived benefits that LISTO had on their teaching practices and on student learning, specifically with regard to the observable changes in student confidence; and engagement in science. The following sections summarize teacher responses.

Professional development and coaching. Overall, teachers responded positively to the virtual professional development (VPD) opportunities. Ultimately, 98% (n = 49) of Year 1 teachers and 92% (n = 37) of Year 2 teachers reported that the VPD either met or exceeded their expectations, as shown in Figures 2 and 3. The VPD sessions, according to teachers, helped to create a more collaborative environment in which other teachers could watch and learn from their LISTO colleagues. One teacher respondent stated:

> *What I liked about the large group VPD… as teachers, we rarely have the opportunity to actually do and see another colleague teach because we are busy teaching our own classes. We were able to see, 'I'm not the only one in this, and I'm not alone.'*

The professional development sessions fostered a sense of community and camaraderie among teachers. This led to an environment where teachers felt "very supported." In particular, the majority of teachers in Year 2 (62%, n = 23) were in agreement that they felt a relationship with others participating in the VPD.

Figure 2
Teachers' perceptions of the VPD (n =49) in Year 1

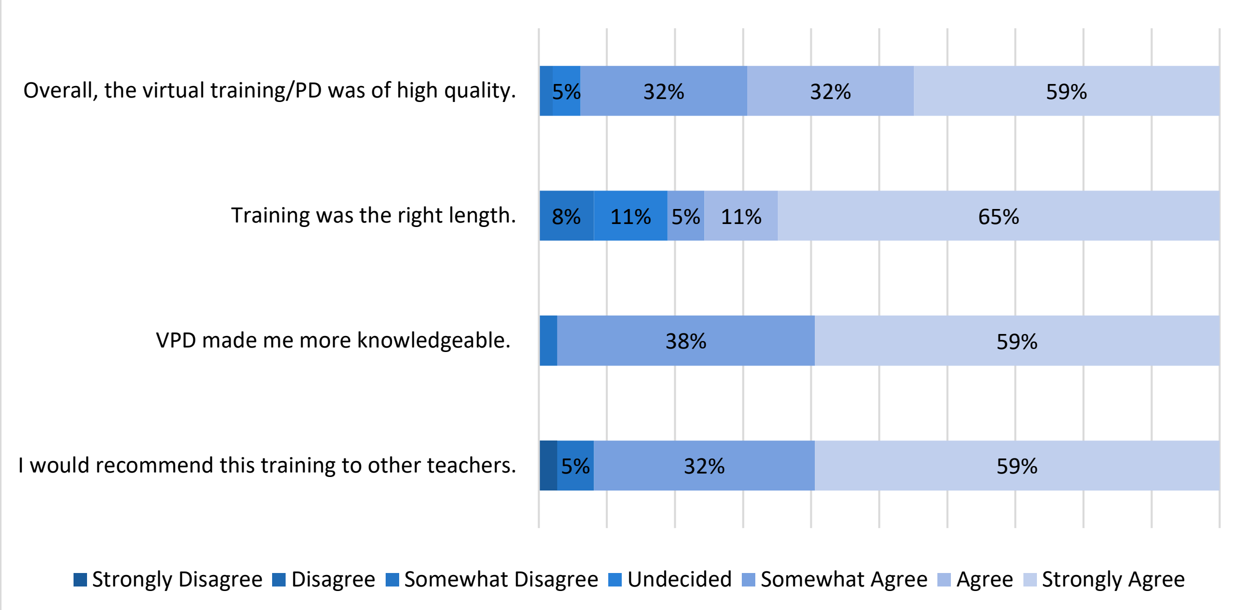


Note: Values < 5.0% are not labeled.

Teachers rated the VPD on a 5-point Likert scale where 1 = strongly disagree; 2 = disagree; 3 = undecided; 4 = agree; and 5 = strongly agree. The results follow:

- Overall, the virtual training/PD was of high quality: 63% agree; 29% strongly agree
- Training was the right length: 27% undecided; 31% agree; 35% strongly agree
- The information I learned will improve my teaching: 8% undecided; 45% agree; 45% strongly agree
- I would recommend this training to other teachers: 6% disagree; 14% undecided; 39% agree; 41% strongly agree

Figure 3
Teachers' perceptions of the VPD (n =37) in Year 2



Notes: 1. Values < 5.0% are not labeled. 2. The Likert-type scale used in Year 2 differed from the Year 1 scale, adding the options of "Somewhat Agree" and "Somewhat Disagree."

Teachers rated the VPD in Year 2 on a 7-point Likert scale where 1 = strongly disagree; 2 = disagree; 3 = somewhat disagree; 4 = undecided; 5 = somewhat agree; 6 = agree; and 7 = strongly agree. The results follow:

- Overall, the virtual training/PD was of high quality: 5% undecided; 32% somewhat agree; 32% agree; 59% strongly agree
- Training was the right length: 8% somewhat disagree; 11% undecided; 5% somewhat agree; 11% agree; 65% strongly agree
- VPD made me more knowledgeable: 38% somewhat agree; 59% strongly agree
- I would recommend this training to other teachers: 5% somewhat disagree; 32% somewhat agree; 59% strongly agree

Although the virtual togetherness was beneficial for teachers, many of them found issues with the VPD component, namely, the time demand and the relevance of the sessions. Several teachers commented that the VPD sessions felt too lengthy at times and often took place at the end of an already exhaustive school day. According to one respondent:

> *As a classroom teacher it is difficult to extend my day even further for a PD. I at times felt tired and sometimes disconnected depending on what happened that day.*

Another qualm with the VPD sessions centered on relevance. As teachers represented districts across the state in this large-scale study, and followed their district-specific academic calendar and pacing, the science topic of the VPD may not have aligned exactly to what all participating teachers were implementing at a specific time. However, teachers did have access to recorded VPD and other support materials, so relevancy was mostly subjective and not a typical complaint from teachers. In sum, teachers saw more value in the VPD when it covered a topic that they were presently teaching.

Others appreciated the flexibility and convenience of the virtual trainings, with one teacher stating that "Virtual training is time effective." Compared to face-to-face trainings, most participants in Year 2 found the VPD better in terms of convenience of timing (78%, n = 29) and location (97%, n = 36), interaction with colleagues and mentors (68%, n = 25), and ongoing connections to their own classroom practices (89%, n = 33). Based on teacher feedback, Year 2 VPD sessions were reduced from 90 minute to 60 minute sessions. Further, the VPD sessions were recorded so that teachers could go back and review if needed.

Teachers also responded very favorably to the virtual mentor coaching because of its individualized approach and the useful feedback that they received from coaches. The overwhelming majority of LISTO teachers found VMC beneficial. As stated by one teacher, "Coaching feedback was excellent; I would have loved to have had them in my ear more." Still, as with the VPD, the single most common dissatisfaction was the demand that VMC placed on teachers' time, particularly at the end of the school day: "[Virtual coaching] was quite a bit long when we have long days."

Participants in Year 2 responded to questions specifically aimed at the improvements made to the curriculum and support that occurred between Years 1 and 2. The vast majority of teachers reported that their experience was either better, somewhat better, or much better compared to the previous year in all areas, including: vocabulary supports, reading passages and guides, using Nearpod as a delivery mechanism and as formative assessment, student engagement, support videos, monthly progress reports, participation checklists, and flex days.

Despite the general positivity towards VPD and VMC experiences and content, some teachers noted having technological issues. Many reported problems—including connectivity, hardware malfunctions, and an initial unfamiliarity with the software—that impacted the virtual experiences in negative ways. Although some LISTO teachers indicated that the VPD and VMC were time-consuming because of their duration and frequency, the sessions were not always relevant, and technological issues persisted, the overwhelming majority of teachers agreed with the sentiment that "the benefits outweighed the challenges."

Curricula materials. Teachers overwhelmingly agreed that LISTO strongly influenced the pedagogical landscape in their classrooms through literacy- and technology-infused strategies. One of the greatest benefits for LISTO teachers was the curricular materials, and first-year teachers benefited the most. According to one,

> *The thing I liked… is having all the supplies; I don't really have to plan a lot. It's handed to me, and as a first-year teacher… I don't have to spend all weekend long planning what I'm going to do.*

LISTO essentially supplied an instructional playbook for teachers; this helped assuage the uncertainty of first-year teaching and provided a structural framework for the class, where it previously might not have existed. A common refrain among teachers was that "I've learned to pace myself [with] more structure than what I had before."

As mentioned previously, the subcomponent of SRM^2 was not implemented during the first year of the project. The intent of this component was to have university scientists meet via live, synchronous, online session with students; however, during the second year of the project, the interaction was limited to pre-recorded video clips embedded into lesson presentations and opportunities for students to pose questions and scientists to respond. Teachers pointed out that the students found the SRM^2 component ineffective, saying it was difficult to make a connection with mentors through video, and this undercut the value of mentorship. Some specific comments included:

> *I don't think the kids really saw [the videos] too much as mentors because I guess it was just like a video that they watched, you know, like any other thing they would watch on YouTube or things like that… I mean the videos were interesting, you know, but I don't think the kids saw them as mentors just, you know, scientists that were there somewhere far away.*

> *They kind of didn't connect… That's really not that far from where I am. It was just kind of 'oh, it's another adult on the screen, you know telling me something.' They didn't connect it to a mentor.*

This view was consistently reinforced by other teachers, who acknowledged that while the videos were interesting, they did not achieve the intended effect of mentoring students. This likely decreased their effectiveness, or at the very least, reshaped their usefulness in the classroom.

Similarly, teachers gave mixed reviews on the FIS take-home science kits. The expectation is that all treatment teachers send home FIS booklets, and send home GoVision glasses with consented students only to record family interactions while working through the activities. During the second year of the project, 18 treatment teachers returned 251 microSD cards from the GoVision glasses. An end of the year family survey (n = 82), reported that 85% of families considered the FIS family activities fun, 91% considered the activities a valuable learning experience, 84% reported that FIS activities helped the family engage in science-related conversations, and 87% reported the learner's (student's) attitude toward science improved. While some teachers neglected to send the kits home with students entirely, some teachers cited low levels of participation due to limited family involvement, a lack of time, and because the activities were optional. A small number of teachers described the familial involvement with the science kits as "disappointing" and "disengaged." Others described more barriers to home implementation of the science kits because households lacked the necessary materials (such as ice trays) or because parents objected to the idea of introducing recording devices into the home. A teacher respondent said simply of her students, "There's no support at home." Perhaps a more common rationale was that the science kits took a backseat to preparing students for the STAAR test. Still, most teachers gave positive feedback on the science kits, finding that the activities were educational but "a different kind of homework." The family involvement science kits were successful for those students and families who embraced them. A teacher respondent summarized the general sentiment towards LISTO saying, "There may be some difficulties, but it is an overall excellent program."

Perceived program benefits for teachers. Teachers identified numerous benefits that LISTO had on their instructional practices. Most commonly, respondents valued the LISTO program as being a "roadmap" for learning. The curriculum and materials that were provided helped teachers (and students) clearly understand where they were and where they were headed by articulating clear goals and objectives. This helped teachers to "see the bigger picture." Other benefits of LISTO included the ability to identify struggling learners earlier on and the provision of materials for teachers.

Some respondents commented further:

[LISTO] gave you a map so that you could work your way through the lessons really easily and it lets you know exactly where you were going with each lesson. You could follow it to know what the kids should know at the end of the lesson.

[LISTO] definitely helps you on track sort of like where you get help on exactly where you are at. We can look at what we should be able to do and when exactly you will be done.

I love that they had a big picture idea, but they had little pieces for the kids to connect to get the education out of it.

It has helped me a lot with being able to apply those higher order thinking questions towards my students.

I definitely appreciated having all of the material. And all the supplies. Because that's always been a huge issue with us.

Perceived program benefits for students. Teachers indicated that LISTO provided benefits to students in terms of engagement and confidence with regard to science-based content. Teachers attributed the increased interest in science directly to LISTO strategies and to the associated technology. One teacher reported that "There [was] definitely a change in the enthusiasm for learning science when they got to use the technology." In short, instructional technology promoted student engagement with the content. Technology also democratized student participation; one teacher recalled, "I saw a big change in my quiet kids. There was no hiding in class anymore." Technology provided more reserved students with greater opportunities to participate than the traditional call-and-respond lecture style allows for, and therefore improved overall engagement.

In addition, numerous teacher respondents said that their students were excited to go home and talk to their parents about what they learned in science that day. Collectively, teachers agreed that because of LISTO, "[students] were more excited, they were more interested, they were more positive." These changes were most noticeable in lower-performing students. According to two teacher respondents:

It was a great experience to see [students] grow and really become passionate about science.
It's been wonderful to see in our lower students how much more confident they are.

Multiple teacher respondents recounted that the newfound interest in science and the resulting increase in content knowledge translated into learner confidence. A LISTO teacher summed up the change in their students' mindset towards science: "[T]he students felt more confident, they had more knowledge, and they were more interested in the subject." LISTO also empowered students, with one teacher stating, "They're not afraid of taking risks anymore."

Barriers to implementation. The implementation of LISTO was not without its challenges, however. An emergent theme from teacher responses included issues with the pacing of LISTO. Despite many respondents (predominantly first-year teachers) who appreciated the structure of the curriculum provided to them, many found the pacing to be the "hardest part" of LISTO, specifically noting that there was "not enough time for review." Other teachers elaborated:

> *We never really had time to finish everything.*
>
> *At the end, you really had to pick and choose because you were running out of time.*
>
> *I really appreciated when [LISTO] backed off of the expectations to put so many activities in; I felt stressed when I couldn't get to them.*

Although time constraints affected the pacing of LISTO, teacher participants in their second year remarked that it had improved substantially from the first year.

In addition to the time management issue, teachers experienced a variety of technological setbacks, which also may have impacted the quality of implementation. LISTO teachers reported issues with their personal technologies, which impacted their VPD and VMC sessions. These issues included, but were not limited to, audio and Bluetooth connectivity used specifically during VMC, and lagging internet connections at home and at school that challenged use of online software and student use of tablets in the classroom. Consequently, teachers expressed frustration in these areas, and this was reflected in the focus group interviews and teacher surveys.

The LISTO-issued technology devices presented some issues. Aside from the physical challenges and degradation of the tablets (e.g., broken screens, missing chargers, etc.), many teachers noted that the school's internet connectivity caused serious lag time issues, which prevented some content from loading. In response, LISTO replaced devices and chargers once teachers notified project personnel and also worked with district/campus IT support to offer improved Wi-Fi service (e.g., via router or MiFi device). Consequently, several teachers reported that they instead used the school-provided devices (e.g., iPads that campuses already had integrated in their classrooms before LISTO) instead of LISTO-issued tablets. Each teacher had a unique set of difficulties with regard to technology – some more than others – and this likely informed perceptions of the LISTO program, overall. Still, support for LISTO remained overwhelmingly strong.

Recommendations for improvement. LISTO teacher participants in the focus groups were asked to provide recommendations for program improvement. The following recommendations were the most frequently cited. For the most part, these recommendations were not unpacked further, in terms of their justification or rationale. Regardless, these recurring themes for improvement provide valuable insights for program improvement:

- Begin the LIS lessons at the start of the school year rather than introducing them later in the semester. Whereas, due to research required student consent and baseline student testing processes, LIS lessons typically started 4-6 weeks into the fall semester. Teachers felt that students should be introduced to LIS lessons from the beginning in order to establish and uphold expectations for the remainder of the school year.
- Include more dynamic types of assessments besides quizzes. Teachers expressed a desire for more creative and diverse assessment options, even if informal.
- Offer more synchronous options for connecting students with scientists in order to improve the authenticity.
- Dilute the number and complexity of vocabulary words and provide the ability for teachers to add new vocabulary. Teachers requested that blank cards be added to the

vocabulary sets so that they can add relevant terms as they see fit. Additionally, some teachers felt that some of the vocabulary words were too advanced and that they might not align with appropriate reading levels.
- Consider laptops in lieu of tablets, as they provide more functionality and are less fragile.

As challenges are expected with any large-scale implementation of a program, these issues are opportunities for program improvement.

Conclusion

LISTO (Valid 45), and the corresponding VPD, VMC, and curricula resources benefitted teachers and their instructional practices, despite unforeseen barriers to implementation and a subsequently shortened intervention period. While student achievement is important, it is a necessary but tangential aspect of this study, the main focus was to provide teachers with innovative, research-based strategies for instruction and to improve teacher sustainability with diminishing support. In this view, LISTO successfully facilitated the teaching experience. Specifically, LISTO had positive effects on teacher practices for a subsample of teachers, specifically on increased delivery of research-based instruction to teach science content as rated on a rubric by external reviewers (ES = +1.12). There were no differences in two other teacher outcomes, however, focused on the share of instructional time spent teaching new science content while performing various activities.

The LISTO teachers who participated in the program reported a high level of satisfaction with the VPD and VMC opportunities. At times, teachers found the VPD and VMC sessions lengthy and covered upcoming lesson units, not necessarily the unit some teachers were implementing at that time, yet the VPD allowed for greater teacher collaboration, and overall, teachers found the VPD and VMC to be very helpful and useful. The curricula were also appreciated by the teachers, with first-year teachers in particular benefitting from the pacing guides. Teachers also reported some barriers to implementation, including technological issues with the hardware and software and inadequate instructional time to fully engage with and implement the program. But for those who implemented LISTO with high fidelity, the teaching experience and the quality of instruction showed marked improvement.

However, LISTO did not necessarily lead to improved student achievement in science or reading for students in the state of Texas. There was a negative program impact on students' science achievement in both 2017–18 (ES = -0.10), and in 2018–19 (ES = -0.13). These quantitative findings were in conflict with qualitative data collected from LISTO teachers who indicated that the program led to improvements in both science vocabulary and engagement and self-efficacy in science for students. LISTO teachers also indicated that the program had benefited their struggling readers, but there was no observed program impact on student reading achievement in either 2017–18 or 2018–19. While LISTO may have yielded some benefits for students, these benefits were not well captured on standardized tests or survey instruments after only one year of exposure to LIS lessons.

One potential reason for the lack of observed positive effects on student outcomes was that the teacher participation in VPD and VMC components of the program were not programmatically implemented with fidelity. More specifically, teachers attended 90% of the VPD sessions in 62–72% of schools and 90% of the VMC sessions in 54–73% of schools, depending on the implementation year. Therefore, LISTO teachers may not have participated in the program to the extent needed to observe program impacts on student and teacher outcomes.

In sum, LISTO appeared to improve instructional practices for a small sample of teachers who implemented the program for two years, but did not positively impact student outcomes more broadly. One likely reason for the lackluster effects were the issues that impacted the first year of the project, such as incomplete implementation of all proposed project components, which were exacerbated by the disruptions from the hurricane, causing a late start in many districts during the first year and delayed component implementation. Arguably, having limited years (and here, less total program time than originally planned) to learn and implement a new curriculum reduces the capacity of teachers to perfect instructional strategies and consequently impact student achievement relative to control-group colleagues who may employ less innovative but more familiar curricula. Likewise, only one year's exposure by students to novel ways of learning science could limit the development of positive attitudes or translate increases in learning quality from LISTO to higher achievement on standardized science and reading assessments. Encouragingly, teachers' overall positive reactions to the program suggest its potential to improve student affect and learning, but more extensive implementation experience by teachers and multi-year exposure by students starting from early grades may be needed to yield measurable benefits. Clearly, such focuses emerge as a highly recommended topic for future research.

References

Beesley, A. (2011). Keeping Rural Schools Up to Full Speed. *T H E Journal, 38*(9), 26–27.

Ballantyne, K.G., Sanderman, A.R., & Levy, J. (2008). *Educating English language learners: Building teacher capacity.* Washington, DC: National Clearinghouse for English Language Acquisition. Available at http://www.ncela.gwu.edu/practice/mainstream_teachers.htm

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 11*(1), 1–29. https://doi.org/10.2167/beb392.0

Breunig, N. A. (1998). Measuring the instructional use of Spanish and English in elementary transitional bilingual classrooms. *Dissertation Abstracts International, 59*(04), 1046A.

Bruce, K. L., Lara-Alecio, R., Parker, R., Hasbrouck, J. E., Weaver, L., & Irby, B. (1997). Accurately describing the language. *Bilingual Research Journal, 21*(2&3), 123-145.

Bryan, L. A., & Atwater, M. M. (2002). Teacher beliefs and cultural models: A challenge for science teacher preparation programs. *Science Education, 86*(6), 821-839. https://doi.org/10.1002/sce.10043

Buxton, C., & Allexsaht-Snider, M. (2016). *Supporting K-12 English language learners in science: Putting research into teaching practice*. New York, NY: Routledge.

Casteel, C.J., & Ballantyne, K.G. (Eds.). (2010). *Professional development in action: Improving teaching for English learners.* Washington, DC: National Clearinghouse for English Language Acquisition. Available at http://www.ncela.gwu.edu/files/uploads/3/PD_in_Action.pdf

Correll, P. K. (2016). *Teachers' preparation to teach English language learners (ELLs): An investigation of perceptions, preparation, and current practices* (Doctoral dissertation). Retrieved from https://uknowledge.uky.edu/edc_etds/19/

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Trowbridge, England: Cromwell Press Ltd.

Darling-Hammond, L., & Richardson, N. (2009). Teacher learning: What matters? *Educational Leadership, 66*(5), 46-53. Retrieved from http://eds.b.ebscohost.com.proxy1.library.jhu.edu/ehost/pdfviewer/pdfviewer?vid=4&sid=3497d70b-1467-4196-a72c-08a6c258cedc%40sessionmgr103

Davis, C. R., Cannon, S. R., & Fuller, S. C. (2021). The storm after the storm: the long-term lingering impacts of hurricanes on schools. *Disaster Prevention and Management: An International Journal*. https://doi.org/10.1108/DPM-03-2020-0055

Davis, M. H., & McPartland, J. M. (2012). High school reform and student engagement. In S. L. Christenson, A. L. Reschly, & C. Wiley (Eds.), *Handbook of research on student engagement* (pp. 515–539). Boston, MA: Springer.

Delaney, Y. A. (2012). Research on mentoring language teachers: Its role in language education. *Foreign Language Annals, 45*(1), 184-202. doi:10.1111/j.1944-9720.2011.01185.x

Dunbar, S., & Welch, C. (2015). *Iowa assessments: Research and development guide*. Orlando, FL: Houghton Mifflin Harcourt. Retrieved from http://itp.education.uiowa.edu/ia/documents/Research-Guide-Form-E-F.pdf

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138. doi:10.1037/1082-989X.12.2.121

Formby, B. (2018). Nearly 10 percent of Texans displaced by Harvey still haven't gone home, survey says. *The Texas Tribune*. https://www.texastribune.org/2018/08/23/fema-and-texas-dont-know-how-many-people-hurricane-harvey-displaced/

Friedrichsen, P., Chval, K.B. and Teuscher, D. (2007). Strategies and Sources of Support for Beginning Teachers of Science and Mathematics. *School Science and Mathematics, 107*: 169-181. doi:10.1111/j.1949-8594.2007.tb17781.x

Fullan, M. (2004). *Leading in a culture of change*. San Francisco, CA: Wiley.

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-Learning, 2*(1), 5.

Gwet, K. L. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Glover, T. A., Nugent, G. C., Chumney, F. L., Ihlo, T., Shapiro, E. S., Guard, K., Koziol, N., Bovaird, J. (2016). Investigating rural teachers' professional development, instructional knowledge, and classroom practice. *Journal of Research in Rural Education, 31*(3), 1-16.

Hagler, M. (2018). Processes of natural mentoring that promote underrepresented students' educational attainment: A theoretical model. *American Journal of Community Psychology*, *62*(1–2), 150–162. https://doi.org/10.1002/ajcp.12251

Heller, J.I., Daehler, K.R., Wong, N., Shinohara, M. and Miratrix, L.W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching, 49,* 333-362. doi:10.1002/tea.21004

Hurd, N. M., Sanchez, B., Zimmerman, M. A., & Caldwell, C. H. (2012). Natural mentors, racial identity, and educational attainment among African American adolescents: Exploring pathways to success. *Child Development, 18*, 1199–1216

Irby, B. J., Tong, F., Lara-Alecio, R., Mathes, G. P., Acosta, S., & Guerrero, C. (2010). Quality of instruction, language of instruction, and Spanish-speaking English learners' performance on a state high-stakes reading assessment. *TABE Journal, 12*(1), 1-42.

Irby, B., Tong, F., Lara-Alecio, R., Meyer, D., & Rodríguez, L. (2007). The critical nature of language of instruction compared to observed practices and high stakes tests in transitional bilingual classroom. *Research in the Schools, 14*(2), 27–36.

Irby, B. J., Tong, F., Nichter, M, Lara-Alecio, R., Hassey, F., & Guerrero, C. (2011). Hispanic English learners' self-esteem related to instructional program type, language of instruction, and gender. *TABE Journal, 13*(1), 26-48.

Lange, K., Kleickmann, T., & Moller, K. (2012). Elementary teachers' pedagogical content and student achievement in science education. In C. Bruguiere, A. Tiberghien & P. Clement (Eds.), *Science Learning and Citizenship*. Proceedings of the Ninth ESERA-Conference 2011. Lyon.

Lara-Alecio, R., & Parker, R. I. (1994). A pedagogical model for transitional English bilingual classrooms. *Bilingual Research Journal*, *18*(3-4), 119-133.

Lara-Alecio, R., Irby, B., Tong, F., Guerrero, C., Koch, J., & Sutton-Jones, K. (2018). Assessing conceptual understanding via literacy-infused, inquiry-based science among middle school English learners and economically-challenged students. *Education Sciences*, *8*(1), 27. doi:10.3390/educsci801002

Lara-Alecio, R., Tong, F., Irby, B. J., Guerrero, C., Huerta, M., & Fan, Y. (2012). The effect of an instructional intervention on middle school English learners' science and English reading achievement. *Journal of Research in Science Teaching*, *49*(8), 987-1011.

Lara-Alecio, R., Tong, F., Irby, B. J., & Mathes, G. P. (2009). Teachers' pedagogical differences among bilingual and structured English immersion kindergarten classrooms in a randomized trial study. *Bilingual Research Journal, 32*(1), 77-100. doi:10.1080/15235880902965938

Lee, O., & Buxton, C. A. (2013). Teacher professional development to improve science and literacy achievement of English language learners. *Theory Into Practice, 52*, 110-117. doi:10.1080/00405841.2013.770328

Lee, O., Hart, J. E., Cuevas, P., & Enders, C. (2004). Professional development in inquiry-based science for elementary teachers of diverse student groups. *Journal of Research in Science Teaching, 41*(10), 1021-1043. doi:10.1002/tea.20037

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). London, England: SAGE Publications, Inc.

Miranda-Chan, T., Fruiht, V., Dubon, V., & Wray-Lake, L. (2016). The functions and longitudinal outcomes of adolescents' naturally occurring mentorships. *American Journal of Community Psychology, 57*, 47–59.

Monk, D. H. (2007) Recruiting and Retaining High Quality Teachers in Rural Areas. Retrieved from : https://files.eric.ed.gov/fulltext/EJ795884.pdf

Morath, M. (2017). Superintendents and commissioner of education testify before house public education committee on needs resulting from Hurricane Harvey. In: Texas House of Representative: Hearings before the Committee on Public Education, 85th Legislative Session (Testimony of Mike Morath). https://house.texas.gov/video-audio/committee-broadcasts/85

Nguyen, T. D., Cannata, M., & Miller, J. (2018). Understanding student behavioral engagement: Importance of student interaction with peers and teachers. *The Journal of Educational Research, 111*(2), 163-174.

Peske, H. G., & Haycock, K. (2006). Teaching Inequality: How Poor and Minority Students Are Shortchanged on Teacher Quality: A Report and Recommendations by the Education Trust. Education Trust. Retrieved from: https://edtrust.org/resource/teaching-inequality-how-poor-and-minority-students-are-shortchanged-on-teacher-quality/

Pickrom, G. M. (2015). Comparing levels of principal support for science, technology, engineering and mathematics education in urban, suburban and rural schools (Order No. 3664248). Available from ProQuest Central; ProQuest Dissertations & Theses Global. (1727735129). Retrieved from https://search.proquest.com/docview/1727735129?accountid=11752

Pruitt, S. L., & Wallace, C. S. (2012). The effect of a state department of education teacher mentor initiative on science achievement. *Journal of Science Teacher Education, 23*, 367-385. doi:10.1007/s10972-012-9280-5

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed., Vol. 1). London, England: SAGE Publications, Inc.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances.* Russell Sage Foundation. Retrieved from: https://inequality.stanford.edu/sites/default/files/The%20Widening%20Income%20Achievement%20Gap%20Between%20the%20Rich%20and%20The%20Poor.pdf

Reardon, S. F. (2013). The widening income achievement gap. *Faces of Poverty, 70*(8), 10-16.

Samson, J. F., & Collins, B. A. (2012). *Preparing all teachers to meet the needs of English language learners: Applying research to policy and practice for teacher effectiveness.* Washington, DC: Center for American Progress. Available from https://files.eric.ed.gov/fulltext/ED535608.pdf

Tarr, J. E., Reys, R. E., Reys, B. T., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middle-grades mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education, 39*(3), 247-280. Retrieved from http://www.jstor.org.proxy1.library.jhu.edu/stable/pdf/30034970.pdf?refreqid=excelsior%3A38419830eaefc0fd7e535f384bff4591

Texas Education Agency. (2017a). *Enrollment in Texas public schools, 2016-17*. (Document No. GE17 601 12). Austin TX: Author. Available at https://tea.texas.gov/acctres/enroll_index.html

Texas Education Agency. (2017b). *Texas Academic Performance Report: 2016-17 State Performance.* Austin, TX: Author. Available at https://rptsvr1.tea.texas.gov/perfreport/tapr/2017/state.pdf

Tong, F., Irby, B. J., Lara-Alecio, & Koch, J. (2014). Integrating literacy and science for English language learners: From learning-to-read to reading-to-learn. *The Journal of Educational Research, 107*(5), 410-426. https://doi.org/10.1080/00220671.2013.833072

Tong, F., Irby, B. J., Lara-Alecio, R., & Mathes, G. P. (2008). English and Spanish acquisition by Hispanic second grades in developmental bilingual programs: A 3-year longitudinal randomized study. *Hispanic Journal of Behavioral Sciences, 30*(4), 500-529.

Tong, F., Irby, B. J., Lara-Alecio, R., Yoon, M., & Mathes, G. P. (2010). Hispanic English learners' response to a longitudinal English instructional intervention and the effect of gender: A multilevel analysis. *The Elementary School Journal, 110*(4), 542-566.

Tong, F., Irby, B., Tang, S., Lin, S., Guerrero, C., Lara-Alecio, R., & Lopez, T. (2017a). *Indicators of inter-rater reliability for classroom observation instruments as fidelity of implementation in large-scale RCTs*. Paper presented at the annual meeting of American Educational Research Association, San Antonio, TX.

Tong, F., Lara-Alecio, R., Irby, B. J., Mathes, G. P., & Kwok, O. (2008). Accelerating early academic oral English development in transitional bilingual and structured English immersion programs. *American Educational Research Journal, 45*(4), 1011-1044.

Tong, F., Luo, W., Irby, B. J., Lara-Alecio, R., & Rivera, H. (2017b). Investigating the impact of professional development on teachers' instructional time and English learners' language development: A multilevel cross-classified approach. *International Journal of Bilingual Education and Bilingualism, 20*(3), 292-313. http://dx.doi.org/10.1080/13670050.2015.1051509

Tong, F., Tang, S., Guerrero, C. L., Zhen, F., Irby, B. J. & Lara-Alecio, R. (2020). *Validating a Student Science Interest Survey for Economically Challenged Middle School Students From a Randomized Study* [Paper Session]. AERA Annual Meeting (Conference Cancelled): San Francisco, CA. http://tinyurl.com/stwaw56

U. S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. (2017). *What Works Clearinghouse: Procedures and standards handbook version 4.0.* Washington, DC: Author.

Valdés, G. (2004). Between support and marginalisation: The development of academic language in linguistic minority children. *International Journal of Bilingual Education and Bilingualism, 7*(2), 102-132.

Webb, M. L. (2006-2007). Rural School Districts Struggle to Meet the Demands of No Child Left Behind. *San Joaquin Agricultural Law Review, 16,* 191-218.

Yoon, K. W., Duncan, T., Lee, S. W., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement.* (Issues & Answers Report, REL 2007-No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from https://files.eric.ed.gov/fulltext/ED498548.pdf

Appendices

The technical appendices include required i3 tables and instruments used in the study.

Appendix A: i3 Tables

This appendix contains all tables required of evaluations funded by the Investing in Innovation (i3) Fund. Tables include:

- Master list of contrasts
- Impact tables
- Cluster attrition tables
- Baseline equivalence tables
- Fidelity of implementation tables

Master list of contrasts. Table A1 provides a master list of student contrasts, and Table A2 provides a master list of teacher contrasts. These tables also include the outcome and pretest measures, as well as the timing of the administration of the measures. Finally, these tables include whether the contrast was confirmatory (C) or exploratory (E).

Table A1

Master list of student contrasts

Contrast ID	T-Group	C-Group	Domain	Outcome Measure		Pretest Measure		C/E
T_Students_1_Y1	T in Y1	C in Y1	Science	STAAR Science	Spring 2018	ITBS Science	Fall 2017	E
T_Students_2_Y2	T in Y2	C in Y2	Science	STAAR Science	Spring 2019	ITBS Science	Fall 2017 or 2018	C
T_Students_3_Y1	T in Y1	C in Y1	Science	ITBS Science	Spring 2018	ITBS Science	Fall 2017	E
T_Students_4_Y2	T in Y2	C in Y2	Science	ITBS Science	Spring 2019	ITBS Science	Fall 2018	E
T_Students_5_Y1	T in Y1	C in Y1	Science	BISA	Spring 2018	BISA	Fall 2017	E
T_Students_6_Y2	T in Y2	C in Y2	Science	BISA	Spring 2019	BISA	Fall 2018	E
T_Students_7_Y1	T in Y1	C in Y1	Science	Science survey	Spring 2018	Science survey	Fall 2017	E
T_Students_8_Y2	T in Y2	C in Y2	Science	Science survey	Spring 2019	Science survey	Fall 2018	E
T_Students_9_Y1	T in Y1	C in Y1	Reading	STAAR Reading	Spring 2018	STAAR Reading	Spring 2017	E
T_Students_10_Y2	T in Y2	C in Y2	Reading	STAAR Reading	Spring 2019	STAAR Reading	Spring 2018	C

NOTES—1. The research design for all domains was CRT with school assignment. 2. In all cases, exposure to the treatment was one school year. 3. The unit of observation for all domains was the student. 4. The student sample included all study participants who had non-missing pretest and posttest scores. 5. The scale for all measures was continuous.

Table A2
Master list of teacher contrasts

Contrast ID	T-Group	C-Group	Domain	Outcome Measure		Pretest Measure		C/E
T_Teachers_1_Y2	T in Y2	C in Y2	Science	TBOP 1	Spring 2019	TBOP 1	Fall 2017 or 2018	C
T_Teachers_2_Y2	T in Y2	C in Y2	Science	TBOP 2	Spring 2019	TBOP 2	Fall 2017 or 2018	C
T_Teachers_3_Y2	T in Y2	C in Y2	Science	STOR	Spring 2019	STOR	Fall 2017 or 2018	C

NOTES—1. The research design for all domains was CRT with school assignment. 2. Exposure to the treatment was either one or two school years, depending on when teachers joined the study. 3. The unit of observation for all domains was the teacher. 4. The teacher sample included all study participants who had non-missing pretest and posttest scores. 5. The scale for all measures was continuous; note that TBOP is a proportion. 6. TBOP 1 was the share of instructional time spent teaching new science content while students performed academic tasks or received feedback. TBOP 2 was the share of instructional time spent teaching new science content with an explicit focus on oral language. 7. The pretest was taken from fall 2017 whenever possible, but if data were missing for a teacher, the pretest was taken from fall 2018.

Impact tables. Table A3 provides the impact estimates of LISTO on student outcomes. Table A4 provides the impact estimates for teacher outcomes. Table A5 lists the statistical models that were used to estimate program impacts. All impact estimates were estimated separately by school year.

The hierarchical linear models to estimate program effects on student outcomes included the following covariates:

- District
    - Rural status
    - District dummy indicators
- Student
    - Gender
    - Free and reduced-price meals
    - Race (e.g., Black, White, Latino, other, multi)
    - English learner status
    - Reclassified status
    - Migrant status
    - Special education
    - 504 status
    - Dummy indicator for took the Spanish version of the STAAR reading test in 4th grade
    - Baseline achievement (varies by outcome)
    - STAAR grade 4 reading score (in analyses not already including STAAR reading score as a pretest)
    - Teacher's alternative certification dummy indicator

  - Missing variable flags
- Teacher
  - Alternative teacher certification dummy indicator

For all analyses, no participants were dropped from the analytic sample due to missing values on background characteristics. For each characteristic, missing values were imputed and a dummy indicator was created to flag participants who had missing values.

The hierarchical linear models to estimate program effects on teacher outcomes included the following covariates:

- District
  - Rural status
  - District dummy indicators
- Teacher
  - Baseline performance (varies by outcome)
  - Alternative certification dummy indicator
  - Missing variable flags

Note also that propensity score weighting was used in the teacher outcomes analyses.

Table A3

Impact estimates for student outcomes

Contrast ID	Outcome Measure	T School N	C School N	T Stu. N	C Stu. N	Unadj. T SD	Unadj. C SD	Pooled SD	Impact Est.	Impact SE	Std. Effect Size	P
T_Students_1_Y1	STAAR Science	32	35	1188	1128	459.41	469.41	464.31	-48.15	24.50	-0.10	0.049
T_Students_2_Y2	STAAR Science	23	30	1346	1084	556.17	546.77	552.00	-72.67	35.58	-0.13	0.041
T_Students_3_Y1	ITBS Science	33	35	1112	1053	30.37	30.30	30.33	-0.90	1.56	-0.03	0.566
T_Students_4_Y2	ITBS Science	24	31	1289	1040	29.76	29.74	29.75	-2.15	1.78	-0.07	0.226
T_Students_5_Y1	BISA	33	35	1113	1061	5.57	5.58	5.57	-0.17	0.29	-0.03	0.548
T_Students_6_Y2	BISA	24	31	1300	1043	5.52	5.44	5.49	-0.34	0.41	-0.06	0.414
T_Students_7_Y1	Science survey	33	35	1108	1064	0.56	0.51	0.54	-0.07	0.03	-0.14	0.012
T_Students_8_Y2	Science survey	24	31	1272	1037	0.37	0.34	0.35	-0.02	0.02	-0.06	0.285
T_Students_9_Y1	STAAR Reading	32	35	1181	1160	128.22	130.56	129.39	2.65	5.01	0.02	0.597
T_Students_10_Y2	STAAR Reading	22	29	1293	993	132.56	128.98	131.02	4.09	5.73	0.03	0.476

NOTE—1. The degrees of freedom for all models were infinity.

Table A4

Impact estimates for teacher outcomes

Contrast ID	Outcome Measure	T School N	C School N	T Teach. N	C Teach. N	Unadj. T SD	Unadj. C SD	Pooled SD	Impact Est.	Impact SE	Std. Effect Size	P
T_Teachers_1_Y2	TBOP 1	19	25	33	38	0.10	0.14	0.12	-0.02	0.04	-0.20	0.5488
T_Teachers_2_Y2	TBOP 2	19	25	33	38	0.13	0.22	0.18	-0.05	0.07	-0.27	0.4686
T_Teachers_3_Y2	STOR	6	17	8	22	0.34	0.43	0.41	0.45	0.18	1.12	0.0118

NOTES—1. All measures failed baseline equivalence and were adjusted using propensity score weighting. 2. The degrees of freedom for all models were infinity.

Table A5

Statistical models used to estimate program impacts on student and teacher outcomes

Contrast ID	Outcome Measure	Model
T_Students_1_Y1	STAAR Science	mixed staar_science_post treat grand_* if year1==1 & !missing(staar_science_post) & !missing(itbs_pre) || schid: ,
T_Students_2_Y2	STAAR Science	mixed staar_science_post treat grand_* if year2==1 & !missing(staar_science_post) & !missing(itbs_pre) || schid: ,
T_Students_3_Y1	ITBS Science	mixed itbs_post treat grand_* if year1==1 & !missing(itbs_post) & !missing(itbs_pre) || schid: ,
T_Students_4_Y2	ITBS Science	mixed itbs_post treat grand_* if year2==1 & !missing(itbs_post) & !missing(itbs_pre) || schid: ,
T_Students_5_Y1	BISA	mixed bisa_post treat grand_* if year1==1 & !missing(bisa_post) & !missing(bisa_pre) || schid: ,
T_Students_6_Y2	BISA	mixed bisa_post treat grand_* if year2==1 & !missing(bisa_post) & !missing(bisa_pre) || schid: ,
T_Students_7_Y1	Science survey	mixed sciencesurvey_post treat grand_* if year1==1 & !missing(sciencesurvey_post) & !missing(sciencesurvey_pre) || schid: ,
T_Students_8_Y2	Science survey	mixed sciencesurvey_post treat grand_* if year2==1 & !missing(sciencesurvey_post) & !missing(sciencesurvey_pre) || schid: ,
T_Students_9_Y1	STAAR Reading	mixed staar_read_post treat grand_* if year1==1 & !missing(staar_read_post) & !missing(staar_read_pre) || schid: ,
T_Students_10_Y2	STAAR Reading	mixed staar_read_post treat grand_* if year2==1 & !missing(staar_read_post) & !missing(staar_read_pre) || schid: ,
T_Teachers_1_Y2	TBOP 1	mixed round3_actstruct10_1819 treat grand_* if !missing(round3_actstruct10_1819) & !missing(round1_actstruct10_pre) [pweight=ps_actstruct10_y2] || schid: ,
T_Teachers_2_Y2	TBOP 2	mixed round3_mode15_1819 treat grand_* if !missing(round3_ mode15_1819) & !missing(round1_ mode15_pre) [pweight=ps_ mode15_y2] || schid: ,
T_Teachers_3_Y2	STOR	mixed round3_stor_1819 treat grand_* if !missing(round3_ stor_1819) & !missing(round1_ stor_pre) [pweight=ps_ stor_y2] || schid: ,

NOTES—1. Stata version 16.1 was used to estimate all models. 2. Grand_* indicates that all covariates (e.g., the pretest, student covariates, teacher alternative certification, and district dummy variables) were included in the model, and all were grand-mean centered. 3. Note that propensity score weighting was used to estimate the models on the teacher outcomes.

Cluster attrition tables. The following tables provide the cluster (school) attrition rates. Table A6 provides the cluster attrition for the student analyses. The cluster attrition rates (overall and differential) for all outcomes were acceptable for Year 1 student outcomes according to the WWC (2020) standards, but cluster attrition standards were not met for Year 2 student outcomes. Two districts and three schools attrited from the study prior to program implementation due to changes in district administration. District data was not collected for another two schools that participated in 2017–18. Another nine schools attrited before the end of the 2018–19 school year.

Table A7 provides the cluster attrition for the teacher outcomes analyses. Cluster attrition standards were not met for any of the teacher outcomes (WWC, 2020). Because collecting the teacher outcomes required teachers to self-video a lesson and submit the video to the project team, cluster attrition was higher for the teacher outcomes than for the student outcomes.

Table A6

Cluster attrition for student outcomes

Contrast ID	Outcome Measure	T School N	C School N	N School Randomized to T	N School Randomized to C	Attrited T School	Attrited C School	Overall Sch. Attrition Rate (%)	Diff. Sch. Attrition Rate (%)
T_Students_1_Y1	STAAR Science	32	35	35	36	3	1	5.63	5.79
T_Students_2_Y2	STAAR Science	23	30	35	36	12	6	25.35	17.62
T_Students_3_Y1	ITBS Science	33	35	35	36	2	1	4.23	2.94
T_Students_4_Y2	ITBS Science	24	31	35	36	11	5	22.54	17.54
T_Students_5_Y1	BISA	33	35	35	36	2	1	4.23	2.94
T_Students_6_Y2	BISA	24	31	35	36	11	5	22.54	17.54
T_Students_7_Y1	Science survey	33	35	35	36	2	1	4.23	2.94
T_Students_8_Y2	Science survey	24	31	35	36	11	5	22.54	17.54
T_Students_9_Y1	STAAR Reading	32	35	35	36	3	1	5.63	5.79
T_Students_10_Y2	STAAR Reading	22	29	35	36	13	7	28.17	17.70

Table A7

Cluster attrition for teacher outcomes

Contrast ID	Outcome Measure	T School N	C School N	N School Randomized to T	N School Randomized to C	Attrited T School	Attrited C School	Overall Sch. Attrition Rate (%)	Diff. Sch. Attrition Rate (%)
T_Teachers_1_Y2	TBOP 1	19	25	35	36	16	11	38.03	15.16
T_Teachers_2_Y2	TBOP 2	19	25	35	36	16	11	38.03	15.16
T_Teachers_3_Y2	STOR	6	17	35	36	29	19	67.61	30.08

Baseline equivalence tables. For all analytic samples, baseline equivalence on pretests was assessed using the same analytic model to estimate program impacts, except without the covariates. In other words, the baseline mean difference was estimated using an HLM model with the pretest as the dependent variable and the treatment indicator as the independent variable. Table A8 shows the baseline equivalence for the student outcomes, and Table A9 shows the baseline equivalence for the teacher outcomes.

Baseline equivalence was initially not established for teacher outcomes. Therefore, for teacher outcomes, propensity score weighting was applied to the models used to estimate the baseline mean difference (as well as the models used to estimate impacts); consequently, all baseline differences between treatment and comparison groups were <0.25 standard deviations. Note that all statistical models estimating program effects included the pretest as a covariate.

Table A8
Baseline equivalence for student outcomes

Contrast ID	Pretest Measure	T Student N	C Student N	Unadj T SD at Pretest	Unadj C SD at Pretest	Pooled SD for T and C	C Mean at Pretest	T/C Diff. at Pretest	Std. T/C Diff. at Pretest
T_Students_1_Y1	ITBS Science	1188	1128	24.85	25.05	24.95	197.94	-4.16	-0.17
T_Students_2_Y2	ITBS Science	1346	1084	23.90	26.04	24.88	195.82	-4.57	-0.18
T_Students_3_Y1	ITBS Science	1112	1053	24.74	25.02	24.87	198.66	-4.69	-0.19
T_Students_4_Y2	ITBS Science	1289	1040	24.12	25.45	24.72	194.95	-3.75	-0.15
T_Students_5_Y1	BISA	1113	1061	5.24	5.15	5.20	14.57	-0.68	-0.13
T_Students_6_Y2	BISA	1300	1043	5.20	5.32	5.25	14.62	-0.62	-0.12
T_Students_7_Y1	Science survey	1108	1064	0.52	0.48	0.50	3.29	0.01	0.01
T_Students_8_Y2	Science survey	1272	1037	0.37	0.35	0.36	3.15	0.03	0.09
T_Students_9_Y1	STAAR Reading	1181	1160	131.74	140.69	136.25	1492.08	-14.19	-0.10
T_Students_10_Y2	STAAR Reading	1293	993	130.33	131.73	130.94	1502.75	-13.48	-0.10

NOTES—1. The source for the standard deviations was the sample. 2. The outcome measure was the same as pretest measure for all domains except when the outcome was STAAR science. The pretest for STAAR Science was ITBS science.

Table A9

Baseline equivalence for teacher outcomes

Contrast ID	Pretest Measure	T Teacher N	C Teacher N	Unadj T SD at Pretest	Unadj C SD at Pretest	Pooled SD for T and C	C Mean at Pretest	T/C Diff. at Pretest	Std. T/C Diff. at Pretest
T_Teachers_1_Y2	TBOP 1	33	38	0.15	0.15	0.15	0.12	0.00	0.00
T_Teachers_2_Y2	TBOP 2	33	38	0.18	0.25	0.22	0.27	0.00	0.02
T_Teachers_3_Y2	STOR	8	22	0.19	0.26	0.25	2.19	0.00	0.01

NOTES—1. The source for the standard deviations was the sample. 2. The outcome measure was the same as pretest measure. 3. All measures initially failed baseline equivalence and were adjusted using propensity score weighting.

Fidelity of implementation. Table 8 in the report shows that key components of LISTO were not implemented with fidelity. Table A10 shows the fidelity of implementation for each of the key program components by school year. LISTO included three major program components: virtual professional development (VPD), virtual mentoring and coaching (VMC), and literacy-infused science curricula (LIS). Fidelity of VPD, VMC, and curricular materials were each measured at teacher, school, and component levels (see Table 4). High fidelity for each program component was defined at the sample level and if 90% of participating schools had high fidelity, as outlined in Table 4.

Table A10
Fidelity of implementation of each key program component by school year

Key Component 1 (of 3) – Virtual Professional Development (VPD). Fidelity Matrix and Fidelity Results Reporting Table

Indicators	**Definition**	**Unit of implementation**	**Data Source(s)**	**Data Collection (who, when)**	**Score for levels of implementation at unit level**	**Threshold for adequate implementation at unit level**	**Roll-up to next higher level if needed (score and threshold): Indicate level**		**Roll-up to sample level (score and threshold for adequate implementation at sample level)**	**Expected sample for fidelity measure (n = # units in which the intervention is being implemented)**	**Expected years of fidelity measurement**
VPD for teachers	# of PD sessions attended by teacher	Teacher	VPD-Use Survey GoToTraining reports	Collected by TAMU Annually in spring of Years 1 & 2	% of PD sessions attended by teacher	Adequate =1 1 = Teacher participates in at least 90% of PD sessions 0= Teacher participates in < 90% of PD sessions	Adequate = 1 1 = 100% of teachers in a school have score of 1. 0 = <100% of teachers in a school have score of 1		Adequate = 1 1= at least 90% of schools have a score of 1. 0 < 90% of schools have a score of 1	All treated 5th grade science teachers in all treatment schools (Y1, n = 44 teachers; Y2, n = 33	2017-18 (Year 1) 2018-19 (Year 2)

Indicators	Definition	Unit of implementation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level		Roll-up to sample level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure (n = # units in which the intervention is being implemented)	Expected years of fidelity measurement
										teachers)	
VPD for teachers	# of teachers that completed VPD	School	VPD-Use Survey GoToTraining reports	Collected by TAMU Annually in spring of Years 1 & 2	% of teachers that completed VPD	Adequate =1 1 = 100% of participating teachers have high fidelity 0 = Less than 100% of participating teachers have high fidelity	Adequate = 1 1 = 100% of teachers in a school have score of 1. 0 = <100% of teachers in a school have score of 1	.	Adequate = 1 1= at least 90% of schools have a score of 1. 0 < 90% of schools have a score of 1	All treated schools with teacher participants (Year 1, n = 32 schools; Y2, n = 26 schools)	2017-18 (Year 1) 2018-19 (Year 2)
Key component score							Adequate if: 90% of schools have a sum score of 2				
Fidelity Results							**Threshold**		90% of schools have a sum score of 2	**# of Units Measured (of n=35 schools)**	**Year of Measurement**

Indicators	Definition	Unit of implementation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold) : Indicate level		Roll-up to sample level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure (n = # units in which the intervention is being implemented)	Expected years of fidelity measurement
							Achieved Score at the Sample Level		**80% of teachers** **72% of schools**	**44 teachers** **32 schools**	**2017-18**
							Met Threshold Implemented with Fidelity (Yes, No, N/A)		**No**		
							Achieved Score at the Sample Level		**77% of teachers** **62% of schools**	**33 teachers** **26 schools**	**2018-19**
							Met Threshold Implemented with Fidelity (Yes, No, N/A)		**No**		

Key Component 2 (of 3) – Virtual Mentoring and Coaching (VMC). Fidelity Matrix and Fidelity Results Reporting Table

Indicators	Definition	Unit of imple-mentation	Data Source(s)	Data Collection (who, when)	Score for levels of implementa-tion at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level		Roll-up to sample level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure (n = # units in which the intervention is being implemented)	Expected years of fidelity measurement
VMC for teachers	# of coaching sessions attended by teacher	Teacher	VMC-Use Survey GoToTraining reports	Collected by TAMU Annually in spring of Years 1 & 2	% of coaching sessions attended by teacher	Adequate =1 1 = Teacher participates in at least 90% of coaching sessions 0= Teacher participates in < 90% of coaching sessions	Adequate = 1 1 = 100% of teachers in a school have score of 1. 0 = <100% of teachers in a school have score of 1	.	Adequate = 1 1= at least 90% of schools have a score of 1. 0 < 90% of schools have a score of 1	All treated 5^{th} grade science teachers in all treatment schools (n=35)	2017-18 (Year 1) 2018-19 (Year 2)
VMC for teachers	# of teachers that completed VMC	School	VMC-Use Survey GoToTraining reports	Collected by TAMU Annually in spring of Years 1 & 2	% of teachers that completed VMC	Adequate =1 1 = 100% of participating teachers have high fidelity 0 = Less than 100% of participating teachers have high fidelity	Adequate = 1 1 = 100% of teachers in a school have score of 1. 0 = <100% of teachers in a school have score of 1		Adequate = 1 1= at least 90% of schools have a score of 1. 0 < 90% of schools have a score of 1	All treated 5^{th} grade science teachers in all treatment schools (n=35)	2017-18 (Year 1) 2018-19 (Year 2)
Key component score							Adequate if: 90% of schools have a sum score of 2				
Fidelity Results					**Threshold**				90% of schools have a	**# of Units Measured**	**Year of Measurement**

Indicators	Definition	Unit of imple-mentation	Data Source(s)	Data Collection (who, when)	Score for levels of implementa-tion at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level		Roll-up to sample level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure (n = # units in which the intervention is being implemented)	Expected years of fidelity measurement
									sum score of 2	**(of n= 35 schools)**	
							Achieved Score at the Sample Level		**74% of teachers** **73% of schools**	**42 teachers** **33 schools**	**2017-18**
							Met Threshold Implemented with Fidelity (Yes, No, N/A)		**No**		
							Achieved Score at the Sample Level		**70% of teachers** **54% of schools**	**30 teachers** **24 schools**	**2018-19**
							Met Threshold Implemented with Fidelity (Yes, No, N/A)		**No**		

Key Component 3 (of 3) – LISTO curriculum. Fidelity Matrix and Fidelity Results Reporting Table

Indicators	Definition	Unit of imple-mentation	Data Source(s)	Data Collection (who, when)	Score for levels of implementa-tion at unit level	Threshold for adequate implementa-tion at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level		Roll-up to sample level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure (n = # units in which the intervention is being implemented)	Expected years of fidelity measurement
LISTO curriculum	Teacher receives curriculum	Teacher	VPD-Use Survey GoToTraining reports	Collected by TAMU Annually in spring of Years 1 & 2	1 = teacher receives curriculum 0=teacher does not receive curriculum	Adequate =1 1 = Teacher receives curriculum 0= teacher does not receive curriculum	Adequate = 1 1 = 100% of teachers in a school have score of 1. 0 = <100% of teachers in a school have score of 1	.	Adequate = 1 1= at least 90% of schools have a score of 1. 0 < 90% of schools have a score of 1	All treated 5^{th} grade science teachers (approx.. 70) in all treatment schools (35)	2017-18 (Year 1) 2018-19 (Year 2)
Key component score							Adequate if at least 90% of schools have adequate implementation				

Fidelity Results				
	Threshold	At least 90% of schools have adequate implementation	**# of Units Measured (of n=35 schools)**	**Year of Measurement**
	Achieved Score at the Sample Level	**100% of schools**	**YY schools**	**2017-18**
	Met Threshold Implemented with Fidelity (Yes, No, N/A)	**Yes**		
	Achieved Score at the Sample Level	**100% of schools**		**2018-19**

Indicators	Definition	Unit of imple-mentation	Data Source(s)	Data Collection (who, when)	Score for levels of implementa-tion at unit level	Threshold for adequate implementa-tion at unit level	Roll-up to next higher level if needed (score and threshold) : Indicate level	Roll-up to sample level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure (n = # units in which the intervention is being implemented)	Expected years of fidelity measurement
					Met Threshold Implemented with Fidelity (Yes, No, N/A)			**Yes**	**YY schools**	