Scaling and Sustaining Science Teachers Learning from Lesson Analysis (STeLLA^{R}) Professional Learning Program Evaluation

Final Summative Report

American Institutes for Research and BSCS Science Learning

January 2026

Advancing Evidence. Improving Lives.

Acknowledgments

This work was funded by a 2019 U.S. Department of Education, Office of Elementary and Secondary Education, Education Innovation and Research Mid-Phase grant (U411B190029) to BSCS Science Learning.

Contents

Exhibits

Abstract

The Science Teachers Learning from Lesson Analysis (STeLLA) is a yearlong professional learning (PL) program designed by BSCS Science Learning (BSCS) to equip teachers with science content knowledge, pedagogical content knowledge (PCK), and a framework for continuous professional inquiry intended to support instructional practices aligned with the *Framework for K–12 Science Education* (National Research Council, 2012). BSCS was awarded an Education Innovation and Research grant, extended to 6 years because of the COVID-19 pandemic, to take STeLLA to scale and sustain its use at scale with two scale-up strategies, including program delivery by local leaders and a blended delivery model of in-person, virtual synchronous and asynchronous professional learning experiences. The American Institutes for Research® (AIR®), an independent evaluator of STeLLA, has completed a mixed-methods study to assess four aspects of the program: (a) fidelity of implementation, (b) impact, (c) scale-up, and (d) cost and cost-effectiveness. The evaluation involved a fidelity of implementation study and an impact study using a multicohort, school-level cluster randomized trial (CRT).

This evaluation took place in 58 public elementary schools, 50% of which were rural, across 16 districts in two states, Kentucky and Tennessee. For the impact study, we randomly assigned two cohorts of elementary schools (Cohort 1 in school year 2022–23 and Cohort 2 in school year 2023–24) to receive STeLLA immediately (treatment) or continue with normal practice and receive STeLLA at a later time (i.e., business as usual [BaU] control group). A total of 58 schools (32 treatment and 26 BaU control) with 130 teachers and their 5,582 students in Grades 4 and 5 participated in the cluster randomized trial (CRT). To assist in controlling for between-district, between-locale, and other confounding effects, the random assignment occurred in 21 blocks, collectively defined by cohort, district, locale, and school enrollment size.

The final report describes the STeLLA program; AIR's evaluation methods; findings on the degree to which program components and scale-up strategies were implemented with fidelity; the impact of STeLLA on teacher science content knowledge, PCK, and instructional practice, as well as student science achievement; and cost and cost-effectiveness of the program.

Results of implementation analyses indicated that all core program components were implemented with a high degree of fidelity. Local leaders participated in leadership development prior to the full implementation. During the implementation years, teachers participated in STeLLA PL, which was led by local leaders, and implemented STeLLA in their classrooms.

Results of scale-up analyses indicated that BSCS fully implemented all its STeLLA scaling strategies, including building local leadership capacity, expanding flexible and accessible PL opportunities (especially for rural teachers), and establishing scale-up and sustainability plans and open access to program materials. BSCS met or exceeded each scale-up objective, except that it was slightly under the target threshold for providing STeLLA PL opportunities to 50% or more of the nontreatment teachers.

Results of impact analyses indicated that the STeLLA program had significant positive effects on teachers' instructional practice, science content knowledge, and science PCK, with treatment teachers outperforming their counterparts in BaU schools across all three outcomes. For students, findings were mixed. Treatment students scored significantly higher than their peers in BaU schools on the study-specific science assessment. Regarding the analyses of impact on the state science achievement test, treatment students scored significantly lower than their peers in BaU schools. Potential explanations are offered.

Last, cost analyses estimated that the overall cost of implementing the STeLLA program was approximately $1,523 per student and $6,394 per teacher. In addition, analyses of cost-effectiveness estimated that the STeLLA program produced a one standard deviation (*SD*) increase in student science achievement, as measured by the study-specific science assessment, at the expense of $774 per student and produced a one *SD* increase in teacher content knowledge and instructional practice at the expense of $14,582 and $6,927, respectively.

Background

The development and adoption of research-based teaching practices, as presented by the *K–12 Framework for Science Education* (National Research Council [NRC], 2012) and the *Next Generation Science Standards* (NGSS Lead States, 2013), represent a vision of teaching and learning whereby teachers serve as facilitators of learning as students develop sophisticated knowledge and skills through meaningful opportunities for sensemaking. Enactment of classroom instruction aligned with the *Framework* and the NGSS often requires significant shifts in teachers' practice, and, without sustained professional support, many teachers face difficulty in making those shifts (Darling-Hammond et al., 2017; Osborne, 2014; Wilson, 2013). The challenges are particularly acute for elementary grade teachers who were required to complete few science content courses as part of preservice preparation (6–12 credit hours on average; Banilower et al., 2018) and those in rural areas where on-site professional learning (PL) is limited and less accessible, and schools often have fewer teachers to share expertise. As a result, PL programs like STeLLA can support the development of science content knowledge, pedagogical content knowledge (PCK), and skills for continuous professional inquiry that supports improvement to science teaching.

STeLLA is a science teacher PL program that builds on 15 years of research and development; it uses a 1-year cycle of activities aligned with the vision for teaching and learning embodied in the NGSS. STeLLA engages teachers in sustained yearlong analysis of science teaching and learning in professional learning communities (PLCs). The work of PLCs is grounded in a two-lens conceptual framework (Exhibit 1) focused on student thinking and instructional coherence. STeLLA is intended to stimulate deep PL and support teachers in designing and delivering more effective science instruction, thus leading to improved student achievement. STeLLA has demonstrated positive impacts on both teacher and student outcomes, above and beyond those observed in traditional science PL programs (Taylor et al., 2017). STeLLA has been studied in preservice and in-service settings, across the K–12 continuum, and in in-person and virtual formats.

This project was funded by a 5-year Education Innovation and Research (EIR) grant that was extended to 6 years because of the COVID-19 pandemic. The project aimed to take STeLLA to scale and sustain its use at scale with two scale-up strategies: program delivery by local leaders rather than BSCS experts and a hybrid delivery model that used in-person, virtual synchronous and asynchronous components. At the same time, the study sought to facilitate better comparison of results with those of Taylor et al. (2017), which presented the impact of STeLLA when delivered by BSCS experts solely in person. As a result, the current study was purposefully aligned with Taylor et al. by maintaining the same intervention duration and the same study-specific teacher and student outcome measures. The logic model guiding this project can be found in Exhibit 2.

The American Institutes for Research® (AIR®), an independent evaluator of STeLLA, completed this scale-up study of STeLLA. The study focused on four areas: (a) fidelity of implementation, (b) impact, (c) scale-up, and (d) cost and cost-effectiveness. This final summative report begins by briefly describing the STeLLA program, followed by a summary of implementation evaluation findings. Next, the report describes methods and findings for the implementation and impact studies. The presentation of impact evaluation findings is designed to provide all the information necessary for a What Works Clearinghouse (WWC)evidence review. Then, the report summarizes scale-up and cost study findings. The report concludes with our interpretation of the key findings alongside implications and study limitations.

Exhibit 1. The STeLLA Two-Lens Framework

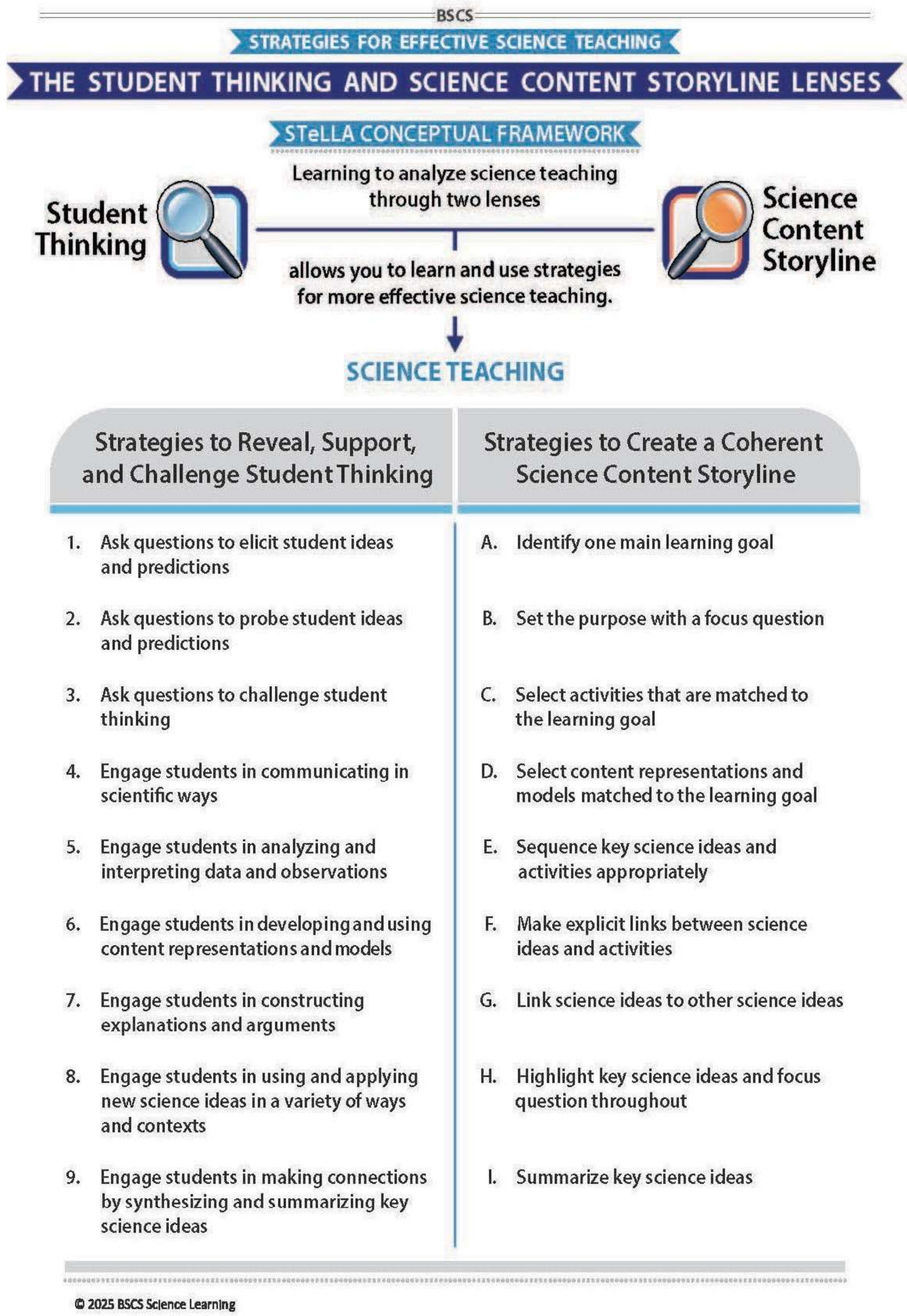
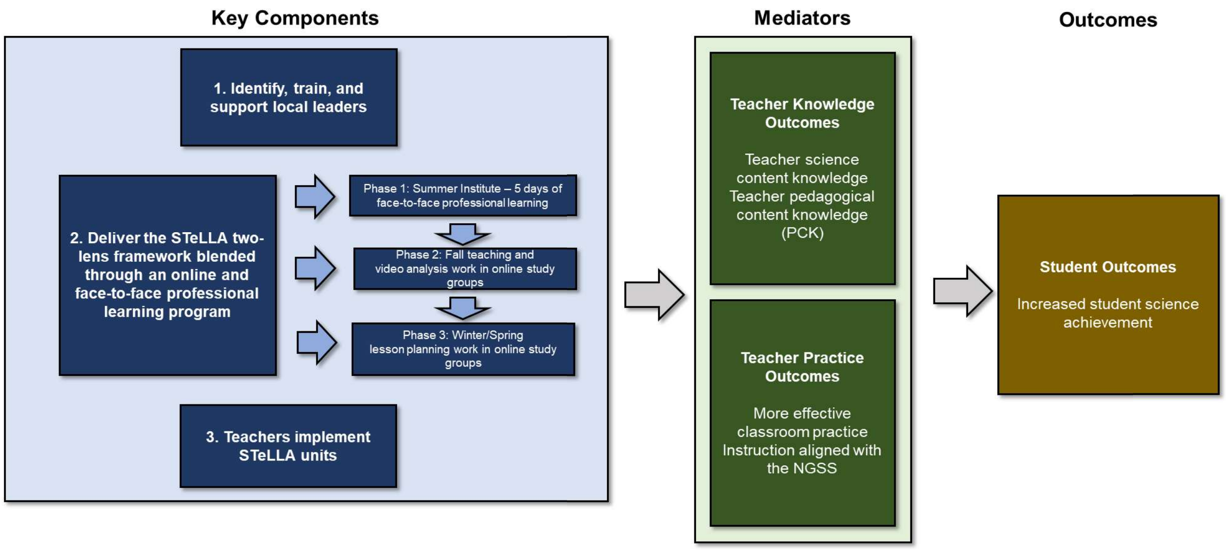


Exhibit 2. STeLLA Logic Model



STeLLA Program Overview

Scaling and sustaining the STeLLA PL program included three essential program components:

1. Engaging teachers in STeLLA PL
2. Implementation of model science curriculum units by teachers
3. Development of local STeLLA PL leaders

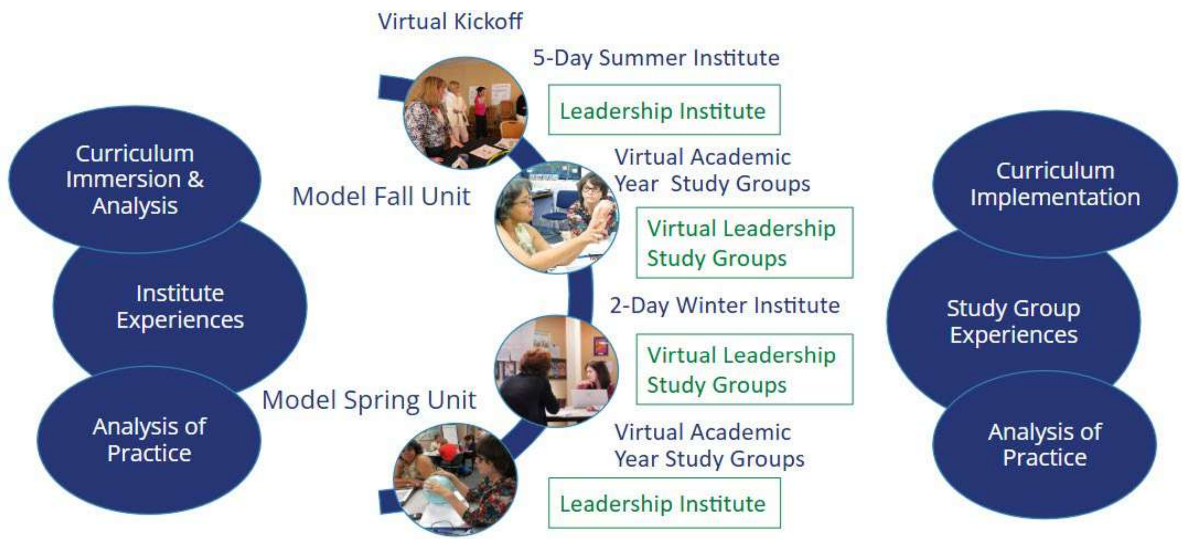
Development and use of the three program components are described next.

Engaging Teachers in STeLLA Professional Learning

The yearlong STeLLA teacher PL program was intended to improve fourth- and fifth-grade teachers' science instruction and students' science achievement in ways consistent with the NGSS and state science standards. Teachers engaged in PL in grade- or course-specific groups (5–20 teachers) during summer and winter institutes and in small, grade-level study groups (5–7 teachers) during the academic year.

The STeLLA teacher PL program (Exhibit 3) began with a 3-hour virtual kickoff in early summer followed by a 5-day on-site summer institute (Phase 1, 40 hours). Phase 2 began during the fall term; teachers participated in four 2-hour virtual study group sessions (8 hours). Phase 3 began with a winter institute (16 hours) followed by four additional 2-hour virtual study group sessions (8 hours).

Exhibit 3. Overview of the STeLLA Professional Learning Program for Teachers and Local Professional Learning Leaders



The three phases of the STeLLA teacher PL program are described in more detail next.

Phase 1. Summer Institute

The STeLLA program began with 5 days of on-site PL to build educator trust, establish norms for collaboration within a learning community, and immerse educators in the model science unit they would teach in the fall term. During immersion in the model science unit, teachers learned science content, the STeLLA two-lens framework and strategies, and the lesson analysis process. Teachers analyzed the model STeLLA units and examined videocases of experienced teachers teaching the model units. Participating teachers practiced identifying STeLLA strategies in use in the videocase classrooms and analyzing the impacts of STeLLA strategies on making student thinking visible.

Videocases are central to the STeLLA teacher PL program. Videocases consist of full-length videos of model STeLLA units (described in the following sections) implemented in classrooms, with 3- to 7-minute clips taken from the full-length video that offer examples of the STeLLA approach in action. In addition, videocases include transcripts of each video clip, a lesson analysis protocol for each clip, and sample student work from unit activities and the postassessment.

Phase 2. First Model Science Unit and Study Group Sessions

During the fall term, teachers implemented the first model science unit and participated in four 2-hour online study group sessions. Between study group sessions, teachers engaged in individual asynchronous professional learning. Teachers collected their own classroom videos and student work and then analyzed classroom video clips from their own classrooms and from those of their peers. Teachers came to the study group sessions prepared to share their analyses. In the virtual study group sessions, they engaged in collaborative analyses of the video clips and student work. Chief goals of study group sessions were to enhance teachers' understanding of three-dimensional content knowledge (i.e., disciplinary core ideas, crosscutting concepts, and science and engineering practices), unpack student ideas during learning, and examine how STeLLA strategies were useful in uncovering and responding to student ideas to move student thinking forward.

Phase 3. Winter Institute, Second Model Science Unit, and Study Group Sessions

The third phase of the STeLLA program began in December or January when teachers gathered for a winter institute. Teachers reflected on lessons learned from implementing the first model unit, then immersed in the second model unit that focused on a new set of disciplinary core ideas with continued development of crosscutting concepts and science and engineering practices. The focus of the monthly study group sessions shifted from planning to implement fully developed model lesson plans (first model unit) to adding the STeLLA strategies to model lesson plans in the second unit, anticipated teacher questions/prompts, and anticipated student talk to the model units. Adding anticipated student talk requires attention to both common science ideas and desired student responses. As in Phase 2, teachers implemented the model unit, collected student work and classroom video, analyzed classroom video asynchronously, and collaborated during four 2-hour study group sessions.

Implementation of Model Science Curriculum Units by Teachers

BSCS curriculum developers with expertise in the STeLLA two-lens framework, the *Next Generation Science Standards*, and both Kentucky and Tennessee Academic Standards developed two model STeLLA units each for Grade 4 and Grade 5 (Exhibit 4). The curriculum development process was iterative, with cycles of development, testing, feedback, and revision. Aspects of each unit were tested in classrooms and by BSCS curriculum developers during the initial development process. Outcomes of testing were used to inform revisions to the final units. Each unit was further revised based on feedback during the pilot phase of the project. Final units are available at https://bscs.org/stella-elementary/instructional-materials/.

Exhibit 4. Model STeLLA Science Curriculum Units

Unit	Grade 4	Grade 5
Fall unit	Energy Everyday, Everywhere	Sun's Effect on Climate and Seasons
Spring unit	Earth's Changing Surface	Matter

Implementation of model STeLLA science units was central to teachers' PL because the model science units provided practical experience with enacting the STeLLA approach to science teaching and learning. Because the science units were designed to meet both Kentucky and Tennessee standards, the pilot institutes and Cohort 1 institute could involve mixed groups of teachers. Each science unit included a unit overview, a teacher-facing content background document, and lesson plans, student-facing handouts, a slide deck, a kit of unit-specific supplies and materials, and student pre- and postassessments along with a teacher-facing feature analysis chart[1] to support teachers in monitoring student learning based on the pre- and postassessments.

Development of Local STeLLA Professional Learning Leaders

In the past, a significant barrier to scaling STeLLA was the program's reliance on BSCS experts to deliver the STeLLA teacher PL program. This approach was difficult to scale because of the limited number of BSCS experts available and geographic constraints. To address this barrier, we implemented a local STeLLA PL leader development program whereby BSCS experts prepared and supported local PL leaders who then led STeLLA PL for teachers who participated in this study. The timing and structure of the local STeLLA PL leader development shifted because of the COVID-19 pandemic. During the first year of the leadership program, local leaders completed a summer leadership institute (see Exhibit 3) and virtual leader study groups during the academic year. To support local STeLLA PL leaders, BSCS created a set of facilitation guides and slides to use during the summer and winter teacher PL institutes and virtual teacher study group sessions. These resources were available to local STeLLA PL leaders who participated in the leadership development program through a BSCS project portal.

Overall, the local STeLLA PL leader development program had dual foci. Local STeLLA PL leaders first experienced the STeLLA teacher PL program as learners and then shifted focus to developing the capacity needed to lead the STeLLA teacher PL program. Three key strategies are used to develop the capacity of local leaders: (a) nested analysis of practice, (b) practice facilitation, and (c) mentorship by experienced BSCS staff as local STeLLA PL leaders implemented the STeLLA teacher PL program with their own groups of teachers.

Nested Analysis of Practice

Analysis of practice is central to the STeLLA PL program. As with the STeLLA teacher PL program, videocases played a central role in the PL of local STeLLA PL leaders. Local leaders first

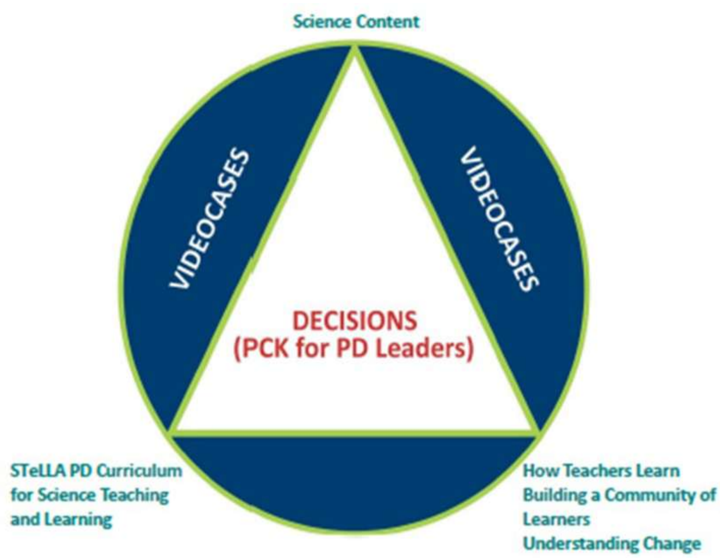
[1] A feature analysis chart is a tool used by teachers to help them track the progress of student ideas across time. The chart includes assessment questions with ideal teacher and student responses and a breakdown of science ideas and skills that serve as evidence of students' progress toward science standards. The feature analysis chart is designed such that teachers can track individual student progress.

engaged in the same videocases used in the STeLLA teacher PL, watching video clips of classroom practice and student work to analyze that classroom practice through the embedded analysis tools in the teacher videocases. Then local leaders analyzed videocases of the STeLLA teacher PL being facilitated by BSCS experts who were leading groups of teachers through the teacher videocases. The goal of this nested analysis of practice—first engaging in videocases as teachers and then engaging in videocases of teacher PL—was to support local STeLLA PL leaders in developing the PCK needed to make decisions about what, when, and how to facilitate STeLLA teacher PL based on their particular (or local) needs (Exhibit 5).

Practice Facilitation

Local STeLLA PL leaders were afforded time during the leadership institutes to study the STeLLA teacher PL facilitator guides. Local leaders were paired to plan for and cofacilitate a 40- to 60-minute session from the STeLLA summer institute for teachers, with others in their leadership cohort serving as participants. As they led the session, local leaders had opportunities to seek support from a BSCS expert. Local leaders were given a "challenge card" that described behavior or ideas often encountered from teachers during a PL session. Each practice facilitation was followed by a coaching session with the option of receiving feedback from a BSCS expert.

Exhibit 5. Framework for the Development of Local Leaders' Pedagogical Content Knowledge for STeLLA Teacher Professional Learning



Mentoring by BSCS Experts

The PL for local STeLLA PL leaders was intentionally scaffolded and faded across time as these local leaders developed the knowledge and skills to facilitate STeLLA teacher PL. In Cohort 1, local leaders cofacilitated the STeLLA teacher PL summer institute with STeLLA experts. Local leaders also observed daily debriefing and reflection by the BSCS team after each day of the summer and winter institutes. Local leaders and BSCS experts also cofacilitated virtual teacher study groups. Each team of local STeLLA PL leaders was supported by a BSCS expert through planning, leading, and reflecting on their practice and teacher learning. In Cohort 2, local STeLLA PL leaders engaged in additional teacher PL practice facilitation but were expected to coplan and cofacilitate with other local leaders the summer and winter institutes and teacher study sessions. BSCS experts supported local leaders in coplanning but did not serve as leads in this process. Local STeLLA PL leaders cofacilitated the summer and winter institutes as BSCS experts observed and offered support as needed. After each day of the institutes, local leaders debriefed and reflected with BSCS experts, using the outcomes to plan for the next day of the institute.

Setting

This school-level, cluster randomized trial (CRT) took place in 58 public elementary schools from 16 districts in two states, Kentucky and Tennessee (Exhibits 6 and 7), across two cohorts (Cohort 1 in school year 2022–23 and Cohort 2 in school year 2023–24). Among all schools, 50% were rural. The two cohorts were similar in terms of student demographic makeup but differed in locale and the percentage of high-needs schools (Exhibit 8). The STeLLA PL was delivered in a standardized format (duration and intensity) across sites.

Exhibit 6. Kentucky Participating Districts Mapped by County

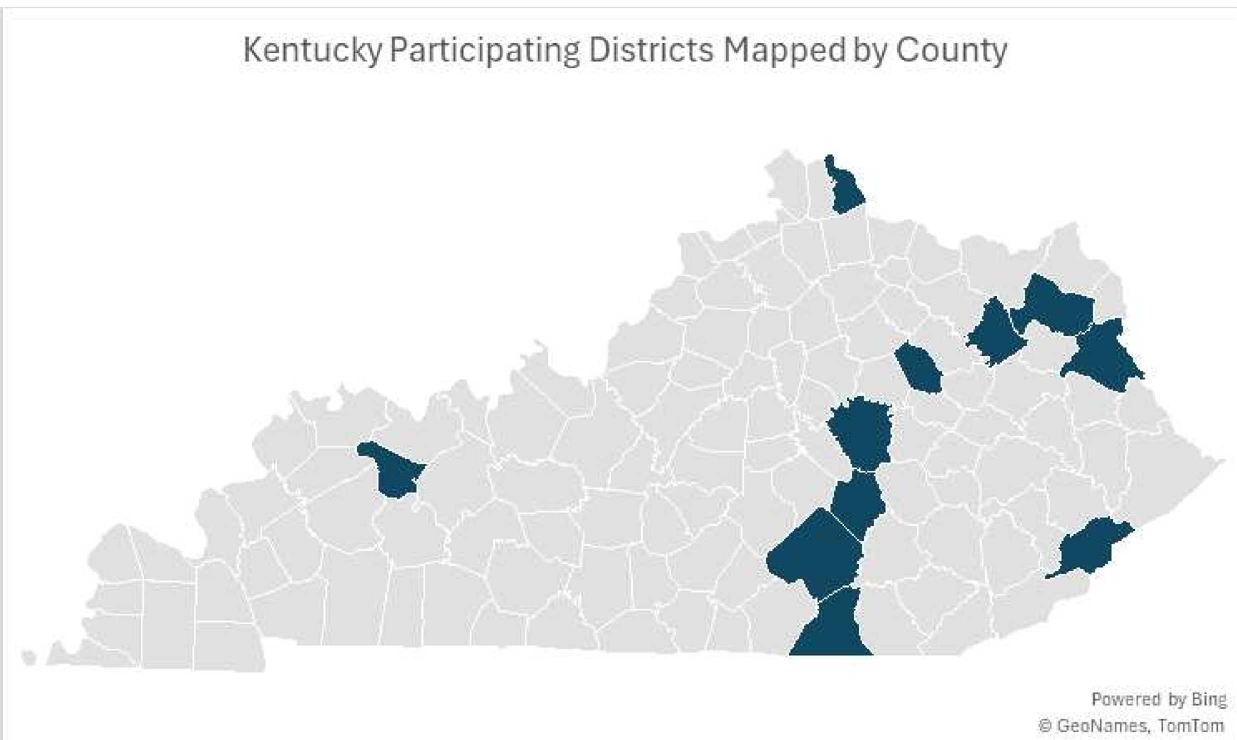


Exhibit 7. Tennessee Participating Districts Mapped by County

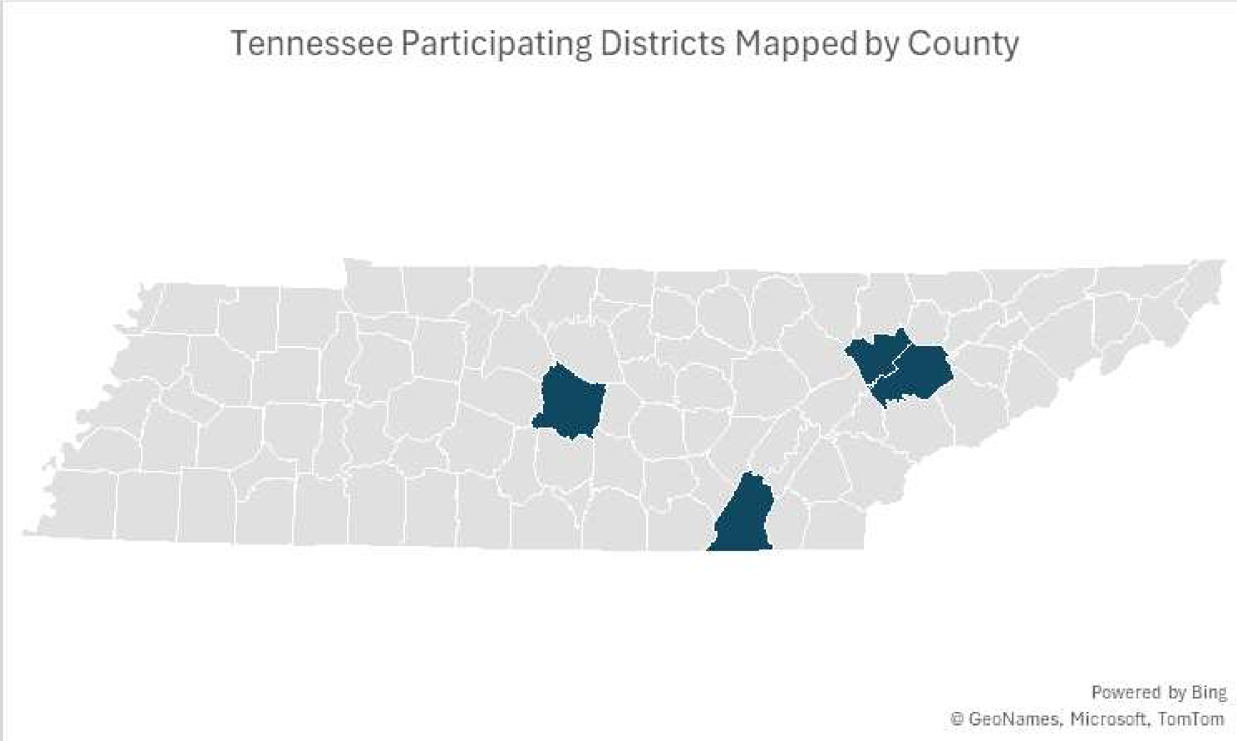


Exhibit 8. School Characteristics by Study Cohort

Characteristics	Cohort 1 (percent)	Cohort 2 (percent)
American Indian or Alaska Native	0	0
Asian	1	1
Black or African American	8	4
Hispanic/Latino	10	6
Native Hawaiian or Other Pacific Islander	0	0
Two or more races	5	4
White	75	85
Male	51	52
Female student	49	48
FRPL student	67	54
High-needs school [a]	85	48
Rural school	39	64

[a] High-needs schools are defined as schools in which more than 50% of students are eligible for free or reduced-price lunch (FRPL).

For each cohort, the study recruited schools during the prior school year (2021–22 for Cohort 1 and 2022–23 for Cohort 2).[2] All regular schools in Kentucky and Tennessee that consented to all programmatic and research requirements and had no prior exposure to STeLLA, including the PL provided in the pilot year of the study (2021–22), were eligible for study participation. Priority selection was used for rural sites to ensure that the study sample met EIR specific requirements for rural representation.

All fourth- or fifth-grade teachers in participating schools were invited to participate. Ten schools in Cohort 1 and 14 in Cohort 2 had only one eligible teacher who taught science. The remaining eligible schools had at least two teachers, in Grades 4 or 5, who were self-contained or science teachers. No teachers participated in STeLLA PL prior to 2022–23.

[2] The study recruited schools during the prior school year to allow enough time for the completion of district- and school-level approvals and necessary onboarding activities.

Fidelity of Implementation Study

Study Description

The implementation study addressed the following research question:

To what extent were the STeLLA program's key components implemented with fidelity?

To address the research question, BSCS and AIR collected and analyzed implementation data across 3 years (local leader PL, Cohort 1, and Cohort 2). The first year involved leadership development for 16 local leaders, 13 of whom served as PL leaders for both Cohorts 1 and 2, two of whom served as a PL leader for Cohort 1 only, and one of whom served as a PL leader for Cohort 2 only. Cohort 1 consisted of 34 treatment teachers from six districts, and Cohort 2 consisted of 20 treatment teachers from 11 districts. The implementation study sample represented approximately 81% of the overall CRT sample assigned to the treatment condition.

Key Components and Data Sources

To guide implementation data collection and analysis, BSCS and AIR collaboratively developed a fidelity of implementation (FOI) matrix, which assessed three key components of STeLLA:

Component 1: BSCS identifies, develops, and supports local leaders.

Component 2: Local leaders deliver the STeLLA two-lens framework blended through a virtual and in-person PL program.

Component 3: Treatment teachers implement STeLLA in their classrooms.

For each STeLLA component, a set of fidelity indicators was identified and used to calculate the component-specific FOI score. Implementation data were obtained from a variety of sources, including program data from BSCS (e.g., attendance records), video observation data of local leaders' facilitation of PL activities, video observation data of teachers' classroom instruction, and survey data collected from teachers.

To calculate a component-specific FOI score, indicators assigned the same weight, or priority category (i.e., high, medium, or low), were averaged to create a single score for each category. The average score for each category was then weighted based on defined weight level (.6 for high, .3 for medium, and .1 for low).[3] Finally, weighted indicator scores were summed to create a component score, with possible scores ranging from 0% to 100%. The threshold for fidelity for

[3] Weight levels were established a priori, in consultation with BSCS and based on features that are central in the design principles.

each component was ≥ 75%. In the section that follows, we provide more detail about our analytic approach.

Data Analysis

To assess FOI related to Component 1 (i.e., identifying, developing, and supporting local leaders), BSCS and AIR identified 12 indicators, as listed in Exhibit 9. Component 1 indicators included participation of local leaders in various PL opportunities led by BSCS expert PL leaders. These sessions were designed to equip local leaders with the knowledge, skills, and tools needed to implement STeLLA in their local contexts. During the leader development year of the study, STeLLA program data (i.e., attendance records) were collected and analyzed to determine whether Component 1 was implemented as intended.

Exhibit 9. Component 1 Fidelity of Implementation Indicators

Development activity	Intended dosage	Priority category	Data source
Kickoff leadership development series	40 hours	Medium	Attendance records
Summer orientation meeting	3 hours	Low	Attendance records
Summer institute	40 hours	High	Attendance records
Summer institute debrief	5 hours	High	Attendance records
Summer leadership institute	36 hours	High	Attendance records
Fall lesson analysis study groups	8 hours	High	Attendance records
Fall leader reflection sessions	4 hours	High	Attendance records
Winter institute	16 hours	Medium	Attendance records
Winter institute debrief	2 hours	High	Attendance records
Spring lesson analysis study groups	4 hours	Medium	Attendance records
Spring leader reflection sessions	2 hours	High	Attendance records
Spring leader planning session(s)	24 hours	Medium	Attendance records
Total hours of development	**184 hours**		

To assess FOI related to Component 2 (i.e., local leaders deliver the STeLLA two-lens framework blended through a hybrid PL program), BSCS and AIR identified seven indicators, as listed in Exhibit 10. Component 2 indicators included teachers' participation in summer and winter institutes facilitated by local leaders, teachers' completion of fall and spring homework tasks, teachers' attendance at virtual lesson analysis sessions (i.e., study groups), and the quality of local leaders' facilitation of the study groups as determined by video observation of a fall and spring study group session. Component 2 implementation data were collected during Cohort 1 and Cohort 2 of the study.

Exhibit 10. Component 2 Fidelity of Implementation Indicators

Activity	Intended dosage	Priority category	Data source
Summer institute	40 hours	High	Attendance records
Fall homework tasks	4 tasks (≥ 1 hour per task)	Medium	Teacher survey
Fall lesson analysis sessions	8 hours	High	Attendance records
Winter institute	16 hours	High	Attendance records
Spring lesson analysis sessions	8 hours	High	Attendance records
Spring homework tasks	4 tasks (≥ 1 hour per task)	Medium	Teacher survey
Quality of implementation [a]	--	High	Video observations
Total hours of facilitation/ engagement with teachers	**80 hours**		

[a] Quality of implementation scores were based on video observations of one fall and one spring lesson analysis session facilitated by each local leader group for each cohort of teachers they supported. Two BSCS expert PL leaders used an observation protocol developed by BSCS and AIR to independently code each video. During the coding period, AIR's implementation lead met with the coders monthly to compare scores and reach consensus. Fall and spring scores for each local leader group video were then averaged (i.e., an average of the percentage of points possible). Finally, the local leader group received a "quality of implementation" score of 1 (meets) or 0 (does not meet) depending on whether their average score was equal to or greater than the predetermined implementation threshold (≥ 80% of points possible).

For each of the seven Component 2 indicators, we first calculated the average score for teachers within the same study group. For example, for Indicator 1, we averaged the summer institute attendance hours for each teacher in the same study group (i.e., total hours; percentage of possible hours). Then we calculated a component average of the study group–level averages. To account for the impact of missing data, we calculated raw averages (i.e., average score for all teachers in the same study group, including those who did not have data reported for that indicator), as well as adjusted averages (i.e., average score for all teachers in the same study group, including only those with data reported for that indicator).

Finally, to assess FOI related to Component 3 (i.e., treatment teachers implement STeLLA in their classrooms), BSCS and AIR identified three indicators, as listed in Exhibit 11. Component 3 indicators included teachers' self-reports on the extent to which STeLLA-based instruction was implemented in their classrooms, as well as teachers' implementation quality score on their spring lesson sequence (as determined by an independent video coding team that watched and scored videos of BaU control and treatment teachers' classroom instruction[4]). Component 3 implementation data were collected during Cohort 1 and Cohort 2 of the study.

Exhibit 11. Component 3 Fidelity of Implementation Indicators

Activity	Priority category	Data source
Teach fall lesson sequence	Medium	Teacher survey
Teach spring lesson sequence	Medium	Teacher survey
Quality of implementation	High	Video observations

As with Component 2, for each of the three Component 3 indicators, we calculated the average score for teachers within the same study group first, then calculated a component average of the study group–level averages.[5]

Findings

In the following sections, we present implementation findings for each component of the STeLLA program. Detailed indicator-level scores for each of these components are presented in Appendix A.

Component 1. Identify, Develop, and Support Local Leaders

Analyses revealed that the first program component—identifying, developing, and supporting local leaders—was implemented with fidelity. Exhibit 12 presents the average indicator score and weighted score for the 12 indicators that comprise Component 1, with seven high-priority indicators, four medium-priority indicators, and one low-priority indicator. A component score of 87% indicates that FOI was achieved and that STeLLA local leaders were identified, developed, and supported as intended.

Exhibit 12. Component 1 Fidelity of Implementation Score

Priority category	Average indicator score (%)	Weighted score (%)
High priority	88	53
Medium priority	84	25
Low priority	94	9
Component 1 score		**87** **Program fidelity = Yes [a]**

[a] Minimum threshold = 75%.

[4] The independent coding group did not know the classroom condition (i.e., treatment or BaU control) of the videos they observed and scored.

[5] As a reminder, to calculate the component score, indicators assigned the same weight (priority) category (i.e., low, medium, or high) were averaged to create a single score for each weight category. Weighted indicator scores were then summed to create a component score, with a possible range of 0%–100%.

Component 2. Local Leaders Deliver the STeLLA Two-Lens Framework Blended Through a Virtual and In-Person PL Program

Analyses revealed that the second program component—local leaders' delivery of STeLLA PL—was implemented with fidelity. Exhibit 13 presents the adjusted average indicator score and weighted score for the seven indicators that comprise Component 2, accounting for missing indicator data, with five high-priority indicators and two medium-priority indicators. A component score of 86% indicates that FOI was achieved and that local leaders delivered the STeLLA two-lens framework blended through a virtual and face-to-face PL program as intended.

Exhibit 13. Component 2 Fidelity of Implementation Score (Adjusted)

Priority category	Average indicator score (%)	Weighted score (%)
High	93	56
Medium	99	30
Component 2 score (adjusted)		**86** **Program fidelity = Yes [a]**

[a] Minimum threshold = 75%.

Component 3. Treatment Teachers Implement STeLLA in Their Classrooms

Analyses revealed that the third program component—teachers' implementation of STeLLA in classrooms—was delivered with fidelity. Exhibit 14 presents the average indicator score and weighted score for the three indicators that comprise Component 3, with one high-priority indicator and two medium-priority indicators. A component score of 81% indicates that FOI was achieved and that treatment teachers implemented STeLLA practices in their classrooms as intended.

Exhibit 14. Component 3 Fidelity of Implementation Composite Score

Priority category	Average indicator score (%)	Weighted score (%)
High	93	56
Medium	83	25
Component 3 score		**81** **Program fidelity = Yes [a]**

[a] Minimum threshold = 75%.

Implementation Evaluation Limitations

One limitation in interpreting findings from the implementation study is the presence of incomplete survey data. Although instances were not common or widespread, we did see cases of participants skipping specific items or failing to complete the survey altogether, which introduces gaps that can compromise the reliability and representativeness of the results. These missing responses may not be random; for instance, individuals who experience challenges with the program may be less likely to respond, potentially skewing the data toward more favorable outcomes. This nonresponse bias may obscure important nuances in program fidelity or participant experience.

The study team handled missing data in two ways. For teachers who participated in the survey but skipped some items, AIR analysts used mean imputation to address missing survey data. In addition, anticipating that some incomplete survey data could impact FOI monitoring, AIR and BSCS established priority categories for each component indicator before data collection began. By assigning priority categories, the impact of missing survey data was mitigated.

Impact Study

Study Description

The impact study used a school-level CRT in 58 public elementary schools in 16 districts across two states, Kentucky and Tennessee, and two cohorts (2022–23 and 2023–24) to assess the effects of full implementation of STeLLA on teacher and student outcomes (see Exhibit 15). AIR randomly assigned 32 schools to treatment (STeLLA) and 26 schools to a BaU control group. The random assignment occurred within 21 blocks, defined by cohort and district and, for a small number of blocks, by cohort, district, locale, and school enrollment size.

Exhibit 15. Number of Schools by Treatment Status and Cohort

Cohort	Treatment (STeLLA)	Business as usual (BaU)	Overall
Cohort 1	18	15	33
Cohort 2	14	11	25
Overall	**32**	**26**	**58**

The impact study examined both teacher and student outcomes and provided intent-to-treat (ITT) impact estimates for all outcomes. The ITT approach estimates impact based only on outcome availability and randomly assigned treatment condition, ignoring variation in treatment uptake or fidelity. Teacher outcomes included science content knowledge and science PCK, both of which were measured by study-developed assessments, and instructional quality as measured by classroom observations. Student outcomes included science achievement, which was measured by a study-developed assessment and state standardized test scores.

Impact Research Questions

The study addressed four *confirmatory* research questions focused on student outcomes and teacher outcomes. The study also addressed four *exploratory* research questions focused on the moderating effects of student, teacher, or school characteristics, as well as one exploratory research question on the mediating effects of teacher outcomes on student achievement.

Impacts of STeLLA on Student Science Achievement

1. Relative to business as usual (BaU) control schools, what impact does the 1-year STeLLA program, as delivered by local leaders, have on the science achievement of fourth- and fifth-grade students? *(confirmatory main effects for study-specific measure and state test)*
2. To what extent does the impact of STeLLA on the science achievement of fourth- and fifth-grade students, compared to BaU, differ by student, teacher/classroom, or school characteristics? *(exploratory moderating/differential effects)*
3. To what extent is the impact of STeLLA on the science achievement of fourth- and fifth-grade students, compared to BaU, mediated by teachers' instructional practice, science content knowledge, and pedagogical content knowledge? *(exploratory mediating effects)*

Impacts of STeLLA on Teacher Instructional Practice

1. Relative to BaU control schools, what impact does the 1-year STeLLA program, as delivered by local leaders, have on the science instructional practice of fourth- and fifth-grade teachers? *(confirmatory main effects)*
2. To what extent does the impact of STeLLA on the science instructional practice of fourth- and fifth-grade teachers, compared to BaU, differ by teacher or school characteristics? *(exploratory moderating/differential effects)*

Impacts of STeLLA on Teacher Science Content Knowledge

1. Relative to BaU control schools, what impact does the 1-year STeLLA program, as delivered by local leaders, have on the science content knowledge of fourth- and fifth-grade teachers? *(confirmatory main effects)*
2. To what extent does the impact of STeLLA on the science content knowledge of fourth- and fifth-grade teachers, compared to BaU, differ by teacher or school characteristics? *(exploratory moderating/differential effects)*

Impacts of STeLLA on Teacher Science Pedagogical Content Knowledge

1. Relative to BaU control schools, what impact does the 1-year STeLLA program, as delivered by local leaders, have on the science pedagogical content knowledge of fourth- and fifth-grade teachers *(confirmatory main effects)*
2. To what extent does the impact of STeLLA on the science pedagogical content knowledge of fourth- and fifth-grade teachers, compared to BaU, differ by teacher or school characteristics? *(exploratory moderating/differential effects)*

Intervention Condition

Two cohorts of teachers participated in the STeLLA PL program, including the implementation of the STeLLA units in their classrooms. For both cohorts, teachers participated in a 5-day summer institute (July 2022 for Cohort 1 and July 2023 for Cohort 2), academic year study group sessions, and a 2-day winter institute (January 2023 for Cohort 1 and December 2024 for Cohort 2). BSCS experts and local leaders cofacilitated the full PL program beginning with the virtual kickoff and summer institute. Cohort 2 was led by local leaders with support from a BSCS expert.

Control Condition

Teachers in the schools assigned to the BaU control condition conducted business as usual instruction without participating in the STeLLA PL or implementing the STeLLA units. Teachers were informed about the learning goals and focal content areas of the treatment condition so as to align their extant curriculum and instruction. Students in the BaU control group were expected to receive standard fourth- and fifth-grade science curriculum and instruction aligned with STeLLA focal content areas.

Teachers in the BaU control group had an opportunity to receive the STeLLA treatment a year later (2023–24 for Cohort 1 and 2024–25 for Cohort 2). Contamination across the experiment conditions was limited because of the school-level randomization design.

Study Participants

After schools agreed to participate in the study, AIR randomly assigned 32 schools to the treatment condition and 26 schools to the BaU control condition. Concurrently with random assignment, we obtained consent from teachers while withholding school randomization results. At the time of random assignment, the teacher sample included 67 treatment teachers and 63 BaU control teachers. The analytic sample of teacher outcomes included 41 (61%) treatment teachers and 44 (70%) BaU control teachers across outcomes and did not include any teacher joiners.

After random assignment, the study team obtained parental consent for fourth- and fifth-grade students enrolled in participating schools at the start of the school year (fall 2022 for Cohort 1 and fall 2023 for Cohort 2). There were no restrictions on the eligibility of individual students in the study. In addition, participation in the study was voluntary, and participants could withdraw at any time without penalty. The analytic sample of student outcomes included all eligible students in study schools.

Sample Alignment With Those Served by the Program

The evaluation sample included 32 (52%) of the 62 schools that were offered the intervention during the evaluation. The project offered STeLLA programming to 30 schools that were not randomly assigned or otherwise included in the evaluation sample.

Design and Measures

Independence of the Impact Evaluation

AIR conducted an independent and external evaluation of BSCS's STeLLA PL program. To ensure the independence of the evaluation, AIR conducted all aspects of the evaluation and had no role in the development or the implementation of the STeLLA intervention except to share implementation analyses as feedback. This structure ensured the independence of key evaluation activities including random assignment, outcome data collection, analysis, and reporting.

AIR was responsible for conducting random assignment of schools, collecting key outcome data, conducting analyses of intervention impacts, and reporting findings about intervention impacts. Data analysis and findings were not subject to BSCS's review and approval.

Study Design Preregistration

The study was preregistered on the Registry of Efficacy and Effectiveness Studies (REES; #14900.3v1; #14900.1v1).

Study Design

The impact evaluation was a multisite CRT that examined impacts of the 1-year STeLLA PL program for fourth- and fifth-grade science teachers compared to the BaU control condition on postintervention teacher and student outcomes. The randomization employed a 50% assignment probability to each condition and 21 blocks, collectively defined by cohort, district, school locale, and school enrollment size.

School-level randomization was appropriate for this intervention because STeLLA PL was designed to facilitate within-building collaboration among grade-alike teachers through its study group structure during the academic year. The sample included two cohorts of schools (33 schools in 2022–23 and 25 schools in 2023–24).

Measures

All data collection activities were conducted identically across both study conditions to enable unbiased estimates of the contrasts between STeLLA and the BaU control group. To provide a valid, reliable measure of student science achievement, AIR worked with participating teachers to administer science assessments at baseline and posttest. Teacher content assessments, a video analysis task focused on PCK, and instructional practice videos were administered at baseline and used as teacher outcomes and in mediation analyses. Exhibit 16 shows a summary of study measures.

Exhibit 16. Outcome Domains, Measures, and Timing of Measurement at Outcome and at Baseline

Domain	Outcome measures		Baseline measures	
Outcome domain	Measure	Timing of measurement (cohort)	Measure	Timing of measurement (cohort)
Student outcomes				
Science achievement	State science test	Spring 2023 (C1) Spring 2024 (C2)	State science test	Spring 2022 (C1) Spring 2023 (C2)
	Study-administered science test		Study-administered science test	Fall 2022 (C1) Fall 2023 (C2)
Teacher outcomes				
Instructional practice	Observer rating of teacher instructional practice (in video)	Spring 2023 (C1) Spring 2024 (C2)	Observer rating of teacher instructional practice (in video)	Spring 2022 (C1) Spring 2023 (C2)
Science content knowledge	Science assessment	Spring 2023 (C1) Spring 2024 (C2)	Science assessment	Pre–summer institute Summer 2022 (C1) Summer 2023 (C2)
Science pedagogical content knowledge	Video-based lesson analysis task	Spring 2023 (C1) Spring 2024 (C2)	Video-based lesson analysis task	Pre–summer institute Summer 2022 (C1) Summer 2023 (C2)

Note. C1 = Cohort 1; C2 = Cohort 2.

Teacher Outcome Measures

Videos of Instructional Practice. AIR researchers collected videos of classroom instruction at two points in time—baseline (prior spring) and spring of the intervention year—for all participating teachers in the treatment and BaU control schools in both cohorts. Data from baseline and spring of the intervention year were used in the impact analysis to add precision to impact estimates and control for residual baseline imbalance across treatment conditions.

Researchers (raters) blind to treatment condition and timepoint coded two videos per teacher, one at baseline and one postintervention, with a coding scheme focused on NGSS three-dimensional science instruction to measure treatment–control contrast in the quality of science instruction. Across all raters and instructional practice dimensions, interrater reliability was assessed for a 20% sample of duplicate coded videos. For these videos, rater scores were highly consistent (Cohen's kappa [weighted] = .88).

In addition, we used Rasch modeling to examine the internal consistency of the measure, based on the analytic sample and postintervention data. The Rasch person reliability estimate was .96.

***Teacher Content Knowledge Assessment*.** Teacher content knowledge was assessed pre- and postintervention (pre–summer institute, post–spring semester) with two assessments adapted from a prior impact study of STeLLA (Taylor et al., 2017). These assessments were pilot tested prior to the experiment and exhibited strong psychometric properties. Consequently, the assessments were used in the experiment without modification.

Each topic-specific assessment had 26 multiple choice items and addressed the same topics as the corresponding student assessments. Each teacher responded to one topic-specific assessment, based on the grade level they teach, and this assessment included items from the other grade-level assessment. Common items allowed for teacher responses to be equated and combined across grade levels.

AIR used Rasch analysis, with common items, to establish a posttest ability/item calibration scale. Using the posttest item calibrations, baseline responses were anchored to establish a singular baseline score for each teacher, on the same scale as the posttest. We also calculated the reliability coefficient of the teacher content knowledge assessments based on the study sample and posttest data. For the current sample, the Rasch person reliability of the assessment was .57.[6]

[6] The low person reliability estimate may be due to a relatively small sample size. For the full teacher sample including joiners, the Rasch person reliability was .61.

***Teacher Pedagogical Content Knowledge Assessment*.** A video-based lesson analysis task was used to generate a single outcome measure of teachers' PCK. Teachers watched video clips of science teaching and responded to two open-ended prompts asking them about the science content, the teaching, and the students. Responses were coded by independent raters outside the evaluation and implementation teams for attention to specific STeLLA strategies and teachers' understanding of more general aspects of PCK. The evaluation team used Rasch modeling to examine the internal consistency of the measure, based on the analytic sample and postintervention data. The Rasch person reliability estimate was .73.

Student Outcome Measures

Study-Administered Science Assessments. We administered a science assessment at posttest—after 1 year of the intervention, as well as at baseline—at the start of the school year (fall 2022 for Cohort 1 and fall 2023 for Cohort 2). The study-administered science assessments were used in a prior impact study of STeLLA (Taylor et al., 2017). These two assessments, one for each grade level, included multiple choice items and covered two content areas—Earth's Changing Surface for fourth grade and Matter for fifth grade. Each assessment was pilot tested early in the study and exhibited strong psychometric properties. As a result, the assessments were used in the experiment without modification.

Items from the two study-administered assessments were combined to construct a single measure of science achievement across grade levels. Student responses across the two assessments—one for each grade level—were equated using common item equating whereby a small set of common items was embedded in each grade-level assessment. These common items were used to establish a single ability scale and subsequent ability estimate (person measure) for each student that will become the ultimate outcome measure score. Common item equating was used on the two pretests to establish a singular baseline score for each student. The baseline and postintervention versions of each assessment were identical. We also calculated the reliability coefficient of the student assessments based on the study sample and posttest data. For the current sample, the Rasch person reliability of the assessment was .76.

State-Administered Science Assessments*.* AIR collected student-level data from state-administered science assessments; WWC (2022) standards indicate that state assessments are assumed valid and reliable. Student data from spring of the prior academic year on state science, math, and English language arts assessments provided baseline data. Students' test scores were converted to *z*-scores based on the grade-specific statewide means and standard deviations to allow pooling of scores across states and grades.

Analysis Approach

Main Effects

Main effects of the STeLLA intervention were estimated from multilevel (two-level) models that nested either students within schools or teachers within schools, depending on analysis of student or teacher outcomes, respectively. Because schools were the unit of random assignment, treatment effects were estimated at Level 2 (school level) of the model and controlled for both individual- and school-level covariates using grand-mean centering. Impact estimates also were calculated as Hedges' *g* effect sizes in accordance with WWC guidance for effect size reporting from cluster-assignment studies. Specifically, the effect sizes were computed using Equation E.20 in the *WWC Standards and Procedures Handbook 5.0* (WWC, 2022). The standard errors of the effect sizes, needed for the 95% confidence intervals, were computed using Equation E.24 from that same handbook. As an additional way to express the magnitude of impacts, effect sizes were used to calculate WWC Improvement Indices (WWC, 2022, p. 186). Based on the comparison group distribution, these indices represent the difference in percentile rank for an average intervention group student versus an average comparison group student.

The final analytic sample for impacts included those individuals who had complete observations at both outcome and baseline. For missing covariate data, we used dummy code imputation to calculate a constant for missing values and adding a binary missing indicator to the model to account for the imputation. No imputed values for outcomes or baseline values are included in the impact analysis data frame. Full model specifications for main effects of treatment are included in Appendix B.

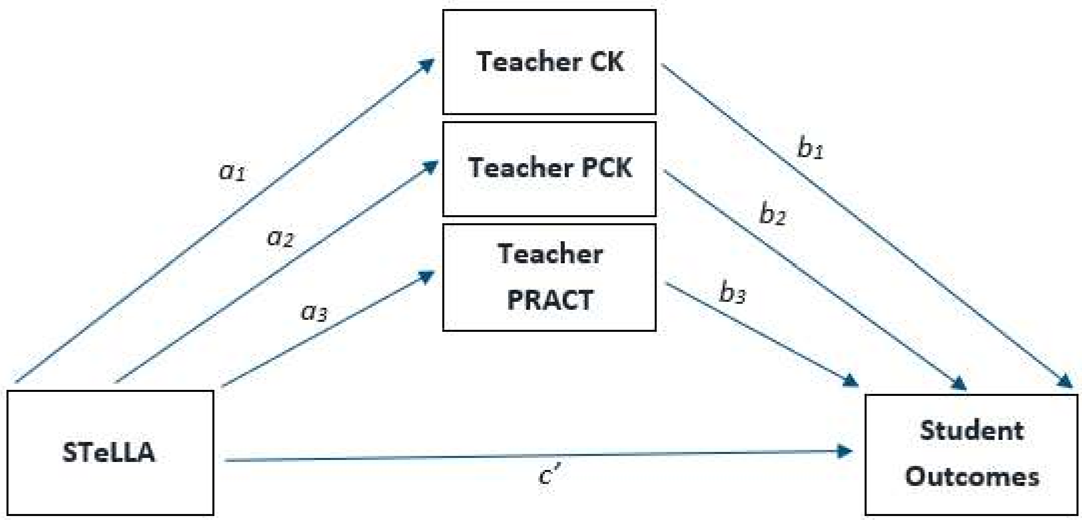
Moderation and Mediation

Exploratory moderation and mediation effects were estimated from three-level models nesting students within teachers and teachers within schools. For moderation, interactions were tested between student demographics and both teacher outcomes and treatment condition (i.e., cross-level interactions). Statistically significant interactions (at the 5% level) were flagged for discussion.

Mediation analysis entailed a suite of four models (Exhibit 17). The first three are the main effects of STeLLA treatment on each of the three teacher outcomes listed in the logic model. These outcomes are content knowledge (CK), pedagogical content knowledge (PCK), and instructional practice (PRACT), which also are hypothesized as mediators of the effect of STeLLA on student outcomes. These main effects supply the *a* paths of the exhibit.

The fourth model was a three-level model nesting students within teachers and teachers within schools. This model provides estimates of the relationship between teacher and student outcomes (the *b* paths), with teacher outcomes measured prior to student outcomes, as well as the mediated effect of STeLLA on student outcomes (c'). For each indirect effect of the teacher outcome/mediators, the significance of the mediation effect through that path (e.g., a_1b_1) was tested using the Sobel Test (Sobel, 1982) of the *ab* product. All model specifications are provided in Appendix B.

Exhibit 17. Hypothesized Mediation Model



Note. CK = content knowledge; PCK = pedagogical content knowledge; PRACT = instructional practice.

Sample Sizes and Attrition

Student Study-Specific Assessment. To measure students' science achievement using the study-specific assessment, we constructed a student analytical sample using listwise deletion to handle missing baseline and outcome data, retaining only students who took both the pretest and posttest. Missing values for nonbaseline covariates were imputed using the variable-specific mean. For each included variable that contained imputed data, a parallel missingness indicator variable also was included in the model. The student analytical sample includes 3,098 students (1,458 treatment, 1,640 comparison) who were in fourth or fifth grade during 2022–23 (Cohort 1) and 2023–24 (Cohort 2) across 45 schools in Tennessee and Kentucky.

At the cluster (school) level, the overall attrition is 22.4%, and the differential attrition is 5.8%. The cluster-level attrition is low under the WWC optimistic boundary but high under the cautious boundary. At the student level, the overall attrition is 34.1%, and the differential attrition is 3.8%. The student-level attrition rates were estimated using student counts in the retained clusters (schools). The student-level attrition is low under the optimistic boundary but high under the cautious boundary set by the WWC (2022).

Student State Achievement. To measure students' science achievement using state achievement, we constructed a student analytical sample using listwise deletion to handle missing baseline and outcome data, retaining only students with both prior math or science state achievement scores and science scores from the intervention year. Missing values for nonbaseline covariates were imputed using the variable-specific mean. For each included variable that contained imputed data, a parallel missingness indicator variable also was included in the model. The student analytical sample includes 4,603 students (2,500 treatment, 2,103 comparison) who were in fourth or fifth grade during 2022–23 (Cohort 1) and 2023–24 (Cohort 2) across 52 schools in Tennessee and Kentucky.

At the cluster (school) level, the overall attrition is 10.3%,[7] and the differential attrition is 2.2%. The cluster-level attrition is low under both the optimistic boundary and the cautious boundary. At the student level, the overall attrition is 13.5%, and the differential attrition is 5.5%. The student-level attrition is low under both the optimistic boundary and the cautious boundary.

Teacher Content Knowledge Assessment. To measure teachers' science content knowledge, we constructed a teacher analytical sample using listwise deletion to handle missing baseline and outcome data, retaining only teachers who took both the pretest and posttest. Missing values for nonbaseline covariates were imputed using the variable-specific mean. The teacher analytical sample includes 84 teachers (40 treatment, 44 comparison) who taught fourth or fifth grade during 2022–23 (Cohort 1) and 2023–24 (Cohort 2) across 37 schools in Tennessee and Kentucky.

At the cluster (school) level, the overall attrition is 36.2%, and the differential attrition is 2.9%. The cluster-level attrition is low under both the optimistic boundary and the cautious boundary. At the teacher level, the overall attrition is 16.8%, and the differential attrition is 3.0%. The teacher-level attrition is low under both the optimistic boundary and the cautious boundary.

Teacher PCK Assessment. To measure teachers' science PCK, we constructed a teacher analytical sample using listwise deletion to handle missing baseline and outcome data, retaining only teachers who took both the pretest and posttest. Missing values for nonbaseline covariates were imputed using the variable-specific mean. The teacher analytical sample

[7] We did not obtain state standardized test data from two districts, which included three participating schools.

includes 85 teachers (41 treatment, 44 comparison) who taught fourth or fifth grade during 2022–23 (Cohort 1) and 2023–24 (Cohort 2) across 39 schools in Tennessee and Kentucky.

At the cluster (school) level, the overall attrition is 32.8%, and the differential attrition is 3.6%. The cluster-level attrition is low under both the optimistic boundary and the cautious boundary. At the teacher level, the overall attrition is 18.3%, and the differential attrition is 0.5%. The teacher-level attrition is low under both the optimistic boundary and the cautious boundary.

Teacher Instructional Practice. To measure teachers' science instructional practice, we constructed a teacher analytical sample using listwise deletion to handle missing baseline and outcome data, retaining only teachers who took both the pretest and posttest. Missing values for nonbaseline covariates were imputed using the variable-specific mean. For each included variable that contained imputed data, a parallel missingness indicator variable also was included in the model. The teacher analytical sample includes 81 teachers (40 treatment, 41 comparison) who taught fourth or fifth grade during 2022–23 (Cohort 1) and 2023–24 (Cohort 2) across 39 schools in Tennessee and Kentucky.

At the cluster (school) level, the overall attrition is 32.8%, and the differential attrition is 3.6%. The cluster-level attrition is low under both the optimistic boundary and the cautious boundary. At the teacher level, the overall attrition is 22.1%, and the differential attrition is 4.1%. The teacher-level attrition is low under both the optimistic boundary and the cautious boundary.

Information on sample sizes and attrition for all student and teacher outcomes is presented in Exhibit 18.

Exhibit 18. Sample Sizes at Randomization and in Analytic Sample Needed to Assess Attrition for an RCT With Cluster-Level Assignment

Outcome measure	Comparison group				Treatment group			
	Clusters [a]		Individuals [b]		Clusters [a]		Individuals [b]	
	No. Randomized	No. Analytic sample	No. Randomized	No. Analytic sample	No. Randomized	No. Analytic sample	No. Randomized	No. Analytic sample
Student study-specific assessment	26	21	2,421	1,640	32	24	2,279	1,458
Student state achievement	26	23	2,517	2,103	32	29	2,807	2,500

Outcome measure	Comparison group				Treatment group			
	Clusters [a]		Individuals [b]		Clusters [a]		Individuals [b]	
	No. Random-ized	No. Analytic sample	No. Random-ized	No. Analytic sample	No. Random-ized	No. Analytic sample	No. Random-ized	No. Analytic sample
Teacher content knowledge assessment	26	17	52	44	32	20	49	40
Teacher PCK assessment	26	18	54	44	32	21	50	41
Teacher Instructional practice	26	18	54	41	32	21	50	40

Note. PCK = pedagogical content knowledge.

[a] Reported only for cluster-assignment evaluations. Not applicable for individual-assignment evaluations.

[b] Report the number of individuals in *nonattrited* clusters only, for cluster-assignment evaluations.

Data Analysis and Findings

Baseline Equivalence

According to the *WWC Procedures and Standards Handbook, Version 5.0* (WWC, 2022), satisfying the baseline equivalence standard requires that baseline group differences, or effect sizes, not exceed 0.25 standard deviations on key covariates. Exhibit 19 presents effect sizes for individual-level baseline measures (student- or teacher-level, depending on the outcome), estimated on the analytic sample. These effect sizes were generated using impact estimates from a multilevel model that regressed baseline measures on the treatment indicator, as well as block indicators. No other covariates were included.

Student Study-Specific Assessment: Effect size for student pretest is −0.091. The absolute value falls between 0.05 and 0.25, indicating that baseline equivalence is satisfied with statistical adjustment.

Student State Achievement: Effect size for student pretest is −0.120. The absolute value falls between 0.05 and 0.25, indicating that baseline equivalence is satisfied with statistical adjustment.

Teacher Content Knowledge Assessment: Effect size for teacher pretest is −0.285. The absolute value exceeds 0.25, and baseline equivalence is not satisfied.

Teacher PCK Assessment: Effect size for teacher pretest is 0.183. The absolute value falls between 0.05 and 0.25, indicating that baseline equivalence is satisfied with statistical adjustment.

Teacher Instructional Practice: Effect size for teacher pretest is 0.437. The absolute value exceeds 0.25, and baseline equivalence is not satisfied.

Exhibit 19. Results From Baseline Equivalence Assessment

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Student study-specific assessment baseline	1,640	0.19	6.50	1,458	−0.21	6.84	−0.61	0.61	0.12	−0.091
Student state achievement baseline	2,103	0.05	0.96	2,500	−0.04	1.03	−0.12	0.08	0.09	−0.120
Teacher content knowledge assessment baseline	44	0.36	5.03	40	−0.40	6.38	−1.63	1.09	0.12	−0.285
Teacher PCK assessment baseline	44	−0.68	8.04	41	0.74	8.05	1.48	1.68	0.26	0.183
Teacher instructional practice baseline	41	−3.99	18.11	40	4.09	15.97	7.50	3.32	0.25	0.437

Note. AMD = adjusted mean difference; *SE* = standard error; ICC = intraclass correlation coefficient; PCK = pedagogical content knowledge.

Representativeness of Individuals in Clusters

We assessed representativeness at both baseline (pretest) and follow-up (posttest) by comparing the number of individuals contributing to posttest and pretest mean with the total number of individuals enrolled in clusters. The numbers in the baseline and follow-up tables (Exhibits 20 and 21) are identical because, for all five outcome measures, the analytic samples include only individuals with complete baseline and outcome data.

Student Study-Specific Assessment: 67.7% of comparison students and 64.0% of treatment students contributed to posttest and pretest mean.

Student State Achievement: 83.6% of comparison students and 89.1% of treatment students contributed to posttest and pretest mean.

Teacher Content Knowledge Assessment: 84.6% of comparison teachers and 81.6% of treatment teachers contributed to posttest and pretest mean.

Teacher PCK Assessment: 81.5% of comparison teachers and 82.0% of treatment teachers contributed to posttest and pretest mean.

Teacher Instructional Practice: 75.9% of comparison teachers and 80.0% of treatment teachers contributed to posttest and pretest mean.

Exhibit 20. Postintervention Cluster Sample Sizes and Enrollment Needed to Assess Representativeness for Cluster-Level Assignment Study

Outcome measure	No. of clusters in analysis	Comparison group: No. of individuals contributing to posttest mean	Comparison group: No. of individuals enrolled in clusters	Treatment group: No. of individuals contributing to posttest mean	Treatment group: No. of individuals enrolled in clusters
Student study-specific assessment	45	1,640	2,421	1,458	2,279
Student state achievement	52	2,103	2,517	2,500	2,807
Teacher content knowledge assessment	37	44	52	40	49
Teacher PCK assessment	39	44	54	41	50
Teacher instructional practice	39	41	54	40	50

Note. PCK = pedagogical content knowledge.

Exhibit 21. Baseline Cluster Sample Sizes and Enrollment Needed to Assess Representativeness for Cluster-Level Assignment Study

Outcome measure	No. of clusters in analysis	Comparison group: No. of individuals contributing to posttest mean	Comparison group: No. of individuals enrolled in clusters	Treatment group: No. of individuals contributing to posttest mean	Treatment group: No. of individuals enrolled in clusters
Student study-specific assessment	45	1,640	2,421	1,458	2,279
Student state achievement	52	2,103	2,517	2,500	2,807
Teacher content knowledge assessment	37	44	52	40	49
Teacher PCK assessment	39	44	54	41	50
Teacher instructional practice	39	41	54	40	50

Note. PCK = pedagogical content knowledge.

Results

Main Effects of STeLLA Treatment

In Exhibit 22, we present the results for our analysis of main effects. For student outcomes we had two confirmatory impact analyses (study-specific science assessment and state science achievement test). All three teacher impact analyses were confirmatory. The impact estimates in the AMD (adjusted mean difference) column are those extracted from multilevel models that included all covariates listed in the analytic plan preregistered with REES (see **Study Design Preregistration**). For other specifications for the impact models, see Appendix B.

Exhibit 22. Impact Analysis Results (Cluster-Level Assignment Study)

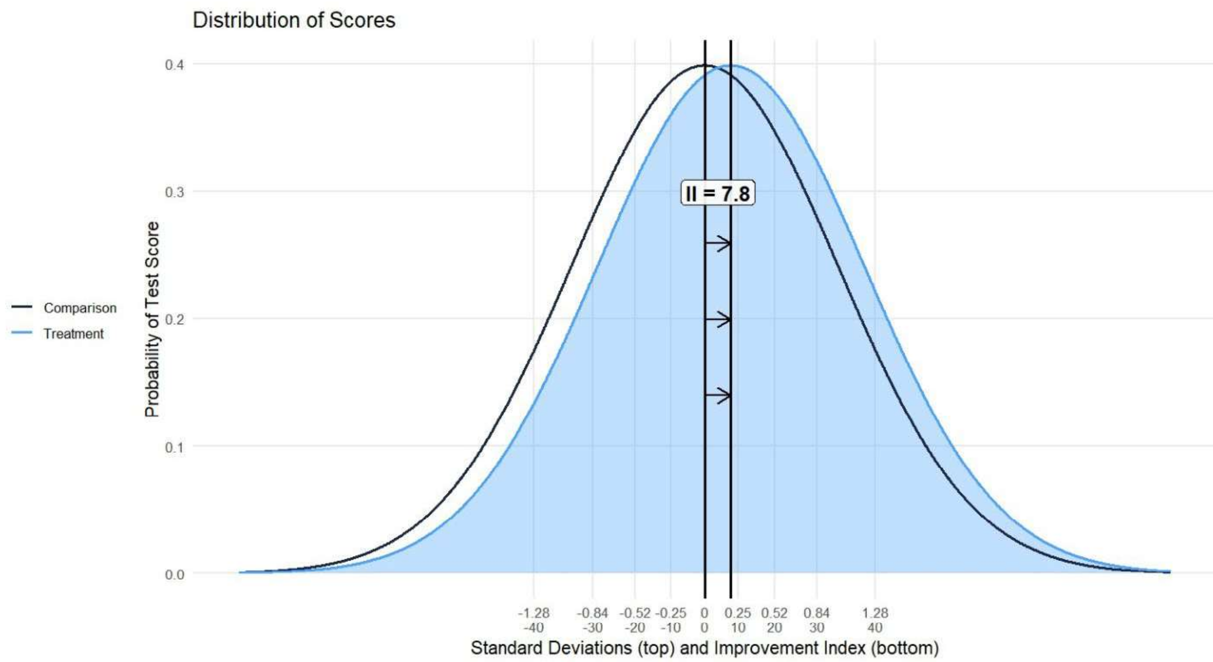
Outcome measure	Comparison group				Treatment group				AMD	*SE* for AMD	*p*-value	ICC	Effect size
	Sample size		*M*	*SD*	Sample size		*M*	*SD*					
	No. of clusters	No. of individuals			No. of clusters	No. of individuals							
Study-specific science assessment	21	1,640	49.45	7.32	24	1,458	51.01	8.71	1.57	0.44	< .001	0.15	0.196
Student state science achievement test	23	2,103	0.10	0.93	29	2,500	-0.08	1.05	-0.11	0.03	.001	0.09	−0.110
Teacher content knowledge assessment	17	44	49.33	5.23	20	40	52.65	4.99	2.49	0.89	.005	0.04	0.486
Teacher PCK assessment	18	44	28.64	9.65	21	41	38.50	7.57	7.76	1.71	< .001	0.20	0.888
Teacher instructional practice	18	41	36.72	17.05	21	40	53.64	11.13	14.51	2.99	< .001	0.46	0.999

Note. AMD = adjusted mean difference; *SE* = standard error; ICC = intraclass correlation coefficient; PCK = pedagogical content knowledge.

Study-Specific Science Assessment

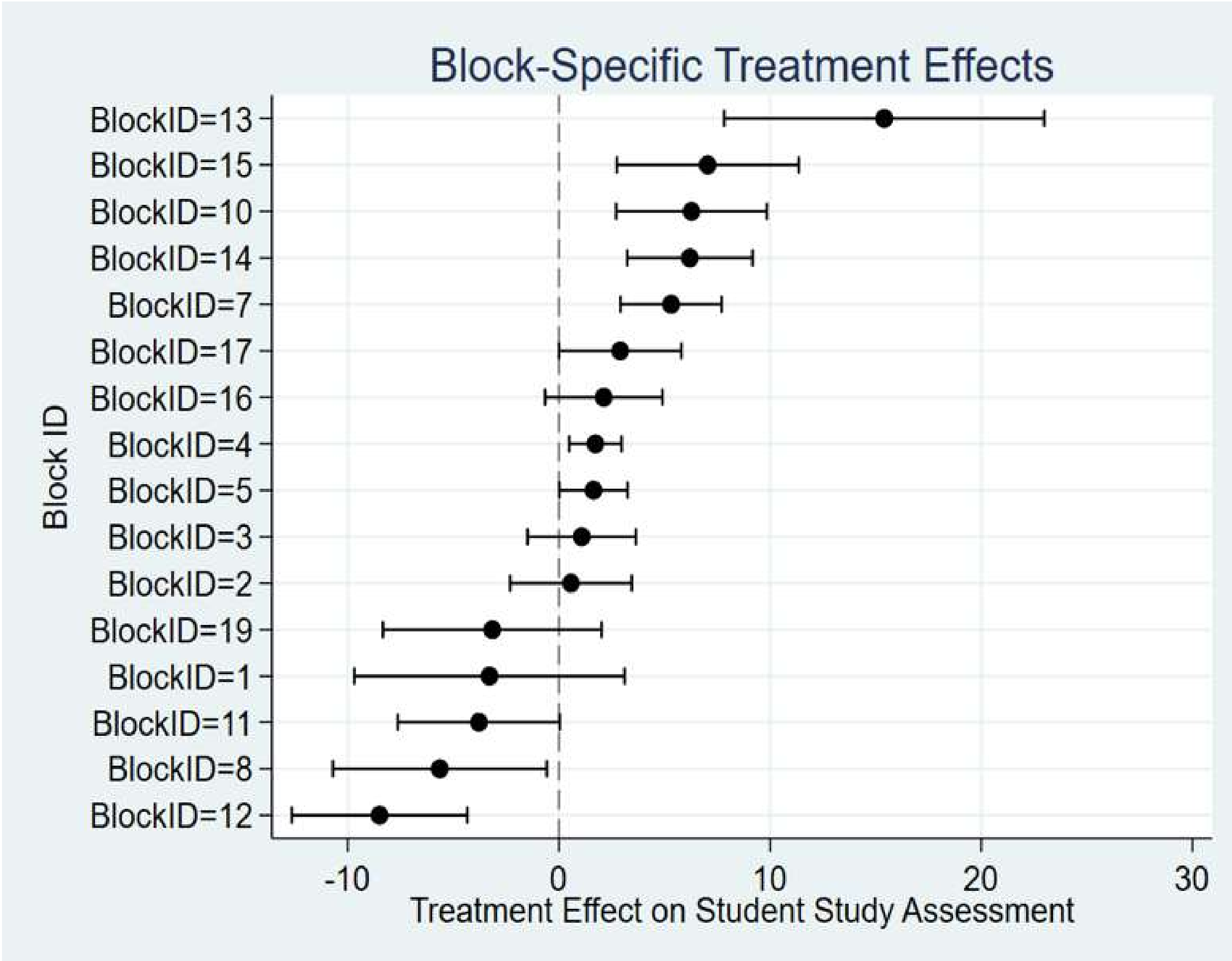
Exhibit 23 shows the main effects results of the multilevel model of STeLLA impact on the equated study-specific achievement measure, scaled from 0 to 100 within the Rasch model. The overall impact estimate suggests that students in the STeLLA treatment condition outperformed students in the BaU control group by 1.57 scale score units, on average ($B = 1.57$, $p < .001$). This effect is statistically significant at the .05 level, and the corresponding Hedges' g effect size is 0.196 *SD*. The corresponding WWC improvement index is 7.77 percentile points, indicating that the average STeLLA student would fall between the 57th and 58th percentiles of the BaU control group distribution, scoring higher than approximately 57% of the BaU control group (see Exhibit 23).

Exhibit 23. Improvement Index for STeLLA Impact on the Study-Specific Science Assessment



The overall impact estimate is a weighted mean of the block-specific impact estimates. An ancillary model was run that generated interaction effects between the treatment and block indicators, producing block-specific treatment effects. As Exhibit 24 illustrates, most block-specific effects are positive, but there is notable heterogeneity in the treatment effects, suggesting further exploration of the extent to which STeLLA was implemented as designed in each of the respective random assignment blocks.

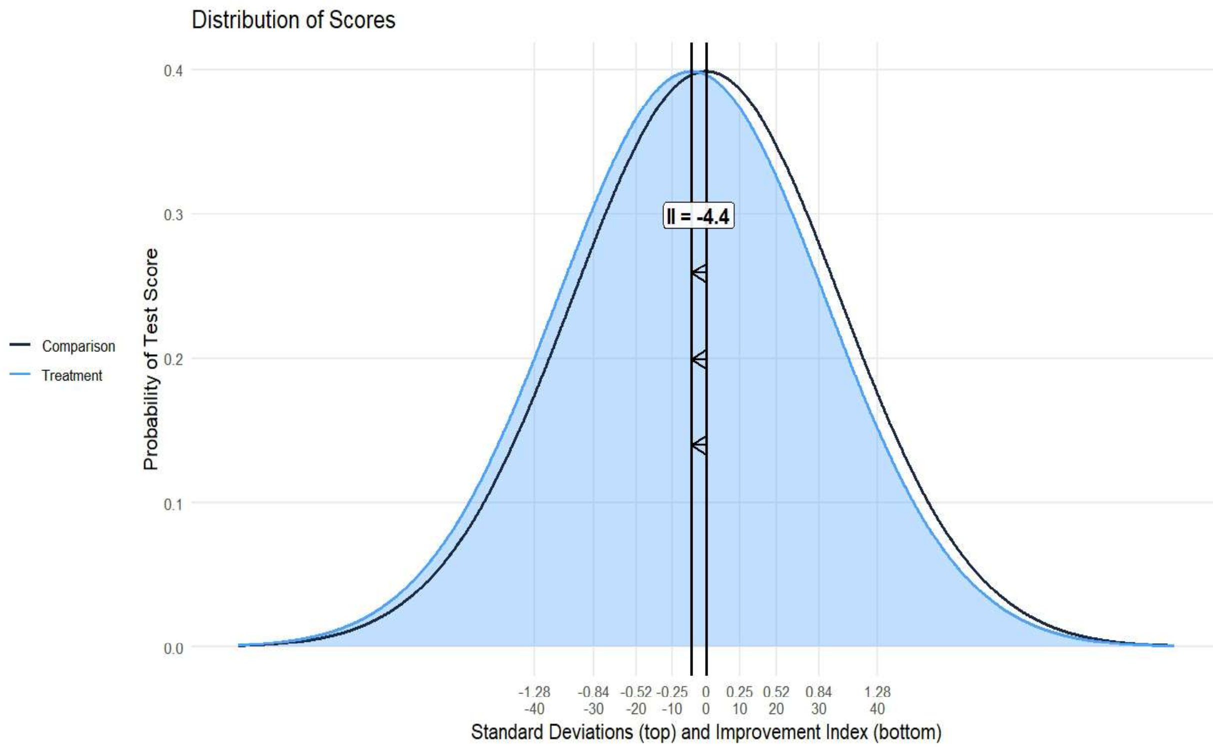
Exhibit 24. Block-Specific Treatment Effects for Study-Specific Science Outcome



Student State Science Achievement Test

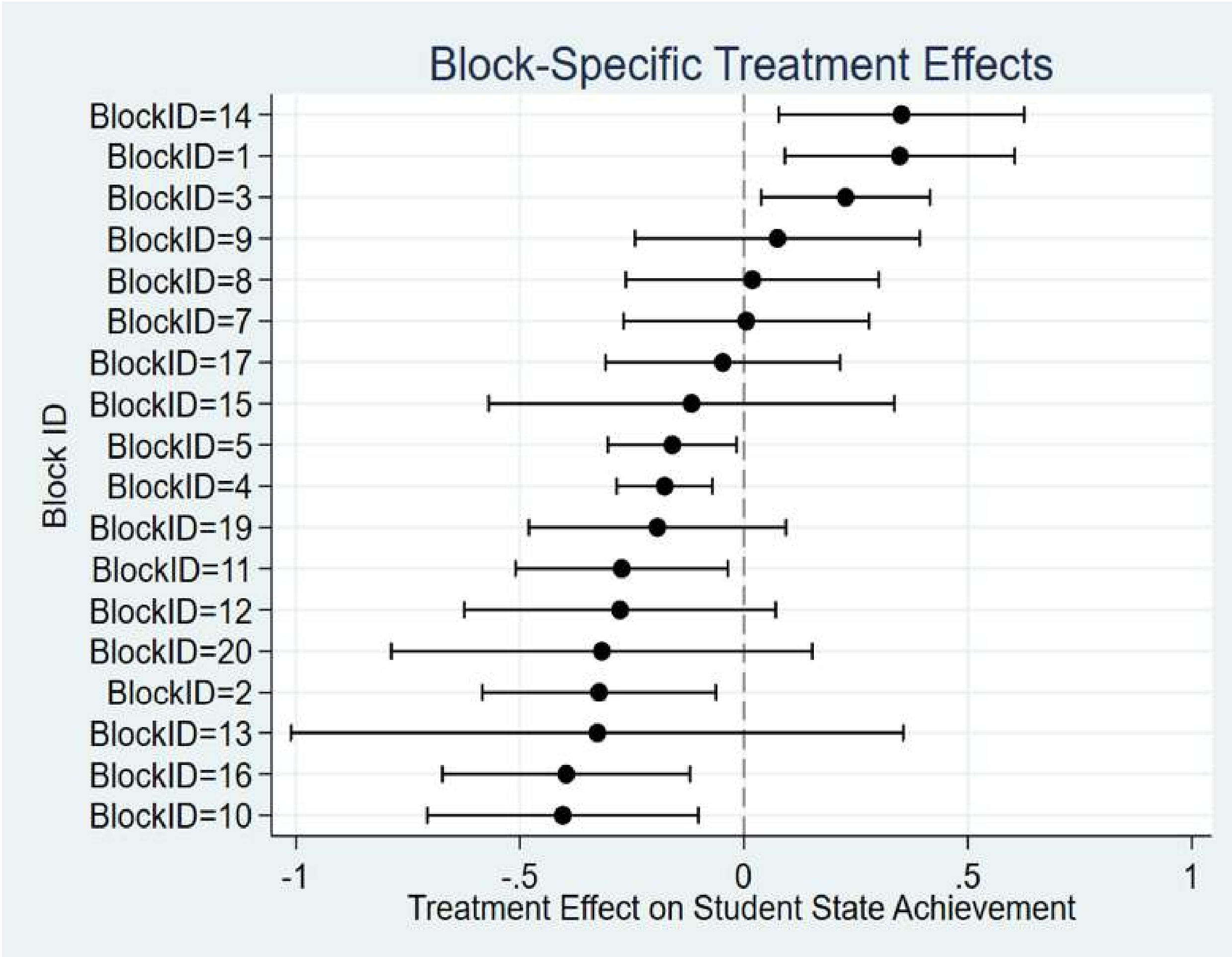
Exhibit 25 shows the main effect of STeLLA on standardized state test scores (*z*-scores), where students' state scale scores were standardized within state and grade level. The ITT impact estimate suggests that students assigned to the BaU control schools outperformed students in the STeLLA condition by 0.11 standard deviation units, on average (B = -0.11, p = .001). This effect is statistically significant at the .05 level. For this impact estimate, the corresponding Hedges' g effect size is also −0.11 *SD* and the improvement index is −4.38, indicating that the average STeLLA student would fall between the 45th and 46th percentiles of the BaU control group distribution, scoring lower than approximately 54% of the BaU control group (see Exhibit 25).

Exhibit 25. Improvement Index for STeLLA Impact on the State Science Achievement Test



The overall impact estimate is a weighted mean of the block-specific impact estimates. An ancillary model was run that generated interaction effects between the treatment and block indicators, producing block-specific treatment effects. As Exhibit 26 illustrates, most block-specific effects are negative, but some heterogeneity exists in the block-specific treatment effects, also suggesting further exploration of STeLLA implementation and of contextual factors present in each of the respective random assignment blocks.

Exhibit 26. Block-Specific Treatment Effects for the State Science Achievement Test

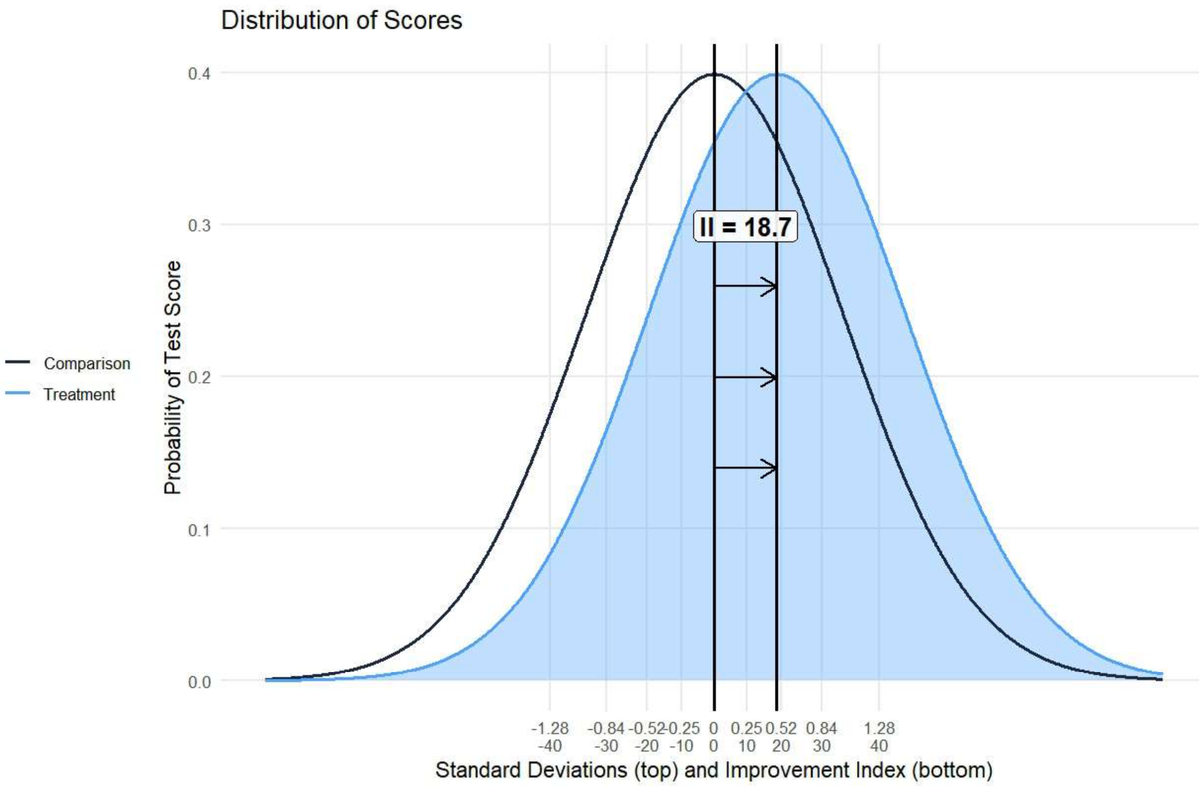


It is worth noting that the ITT analysis for the state science test outcome includes seven more schools and approximately 1,500 more students than did the analogous impact analysis for the study-specific science assessment. This is because even schools that withdrew from the study, postrandomization, contributed state test outcome scores and could be included in the ITT analysis. Consequently, the treatment condition includes a notable number of untreated students who were enrolled in withdrawn schools.

Teacher Content Knowledge

Exhibit 27 shows the main effects results of the multilevel model of STeLLA impact on teacher content knowledge, scaled from 0 to 100 within the Rasch model. The impact estimate suggests that teachers in the STeLLA treatment condition outperformed teachers in the BaU control group by 2.49 Rasch-estimated ability units, on average (B = 2.49, p = .005). This effect is statistically significant at the .05 level, and the corresponding Hedges' g effect size is 0.486 *SD*. The corresponding WWC improvement index is 18.7 percentile points, indicating that the average STeLLA teacher would fall between the 68th and 69th percentiles of the BaU control group distribution, scoring higher than approximately 69% of the BaU control group (see Exhibit 27).

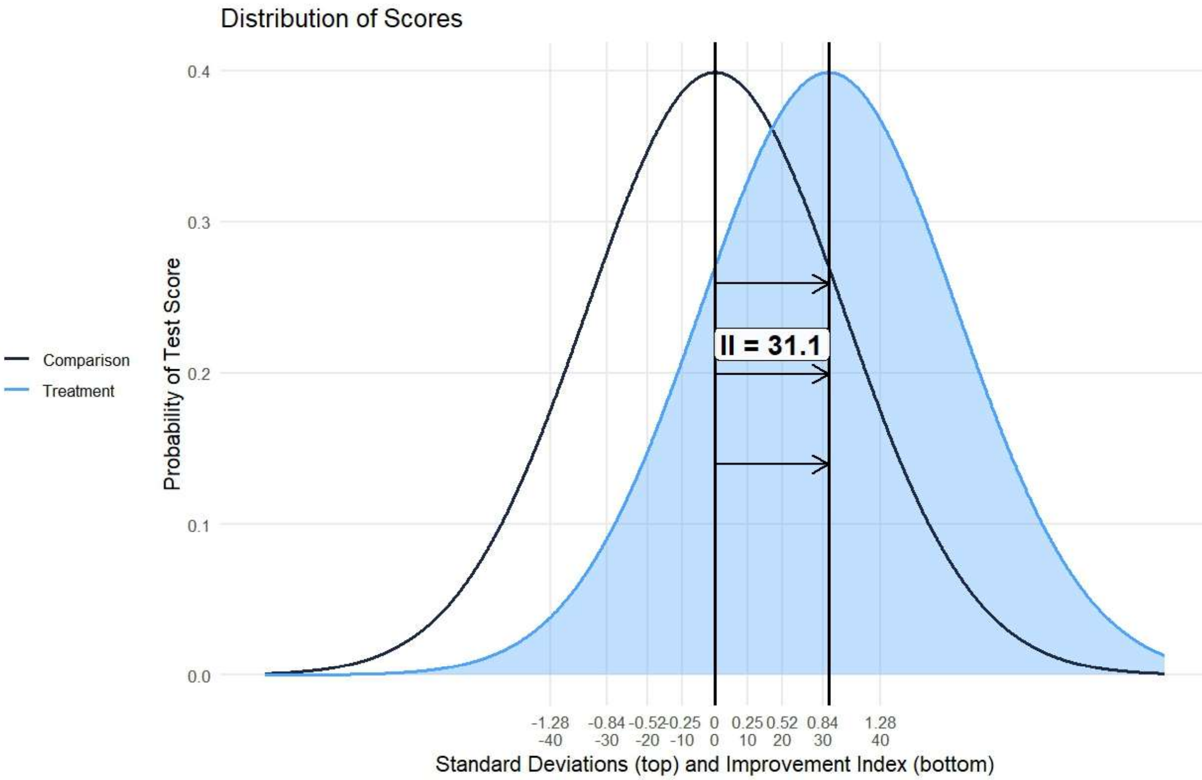
Exhibit 27. Improvement Index for STeLLA Impact on Teacher Content Knowledge



Teacher Pedagogical Content Knowledge

Exhibit 28 shows the main effects results of the multilevel model of STeLLA impact on teacher PCK, scaled from 0 to 100 within the Rasch model. The impact estimate suggests that teachers in the STeLLA treatment condition outperformed teachers in the BaU control group by 7.76 Rasch-estimated ability units, on average ($B = 7.76$, $p < .001$). This effect is statistically significant at the .05 level, and the corresponding Hedges' g effect size is 0.888 *SD*. The corresponding WWC improvement index is 31.3 percentile points, indicating that the average STeLLA teacher would fall between the 81st and 82nd percentiles of the BaU control group distribution, scoring higher than approximately 81% of the BaU control group (see Exhibit 28).

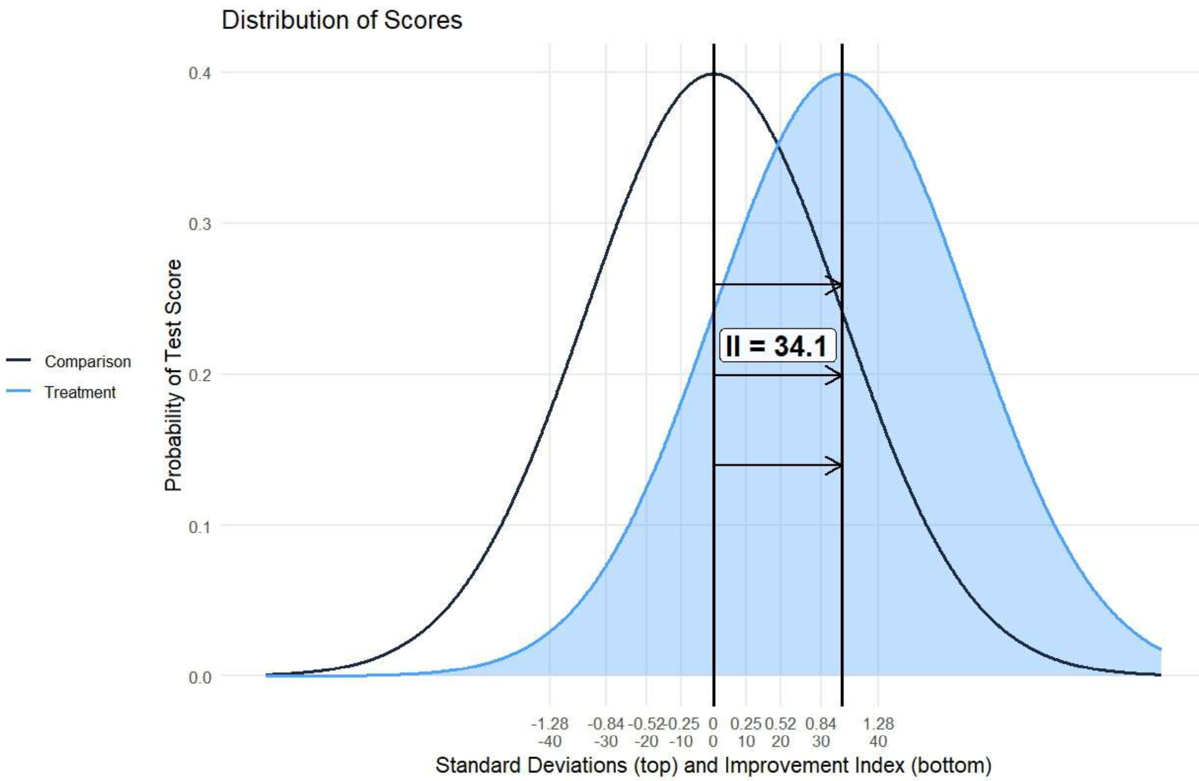
Exhibit 28. Improvement Index for STeLLA Impact on Teacher Pedagogical Content Knowledge



Teacher Instructional Practice

Exhibit 29 shows the main effects results of the multilevel model of STeLLA impact on teachers' instructional practice score, scaled from 0 to 100 within the Rasch model. The impact estimate suggests that teachers in the STeLLA treatment condition outperformed teachers in the BaU control group by 14.51 Rasch-estimated ability units, on average (B = 14.51, p = <.001). This effect is statistically significant at the .05 level, and the corresponding Hedges' g effect size is 0.999 *SD*. The corresponding WWC improvement index is 34.11 percentile points, indicating that the average STeLLA teacher would fall between the 84th and 85th percentiles of the BaU control group distribution, scoring higher than approximately 84% of the BaU control group (see Exhibit 29).

Exhibit 29. Improvement Index for STeLLA Impact on Teacher Instructional Practice



Moderation Effects

Moderation of Student Outcomes

The preregistered analytic plan also specified that we intended to test moderation (i.e., cross-level interactions) between student demographic characteristics and both teacher outcomes and treatment condition. We also tested outcome moderation by teacher-level variables (e.g., content knowledge, teaching experience). For this exploratory analysis, we used a sample frame consisting only of students with complete data on outcomes, baseline measures, demographics, teacher outcomes, and treatment condition. For the study-specific science assessment, this sample included 1,965 students, and for the state science achievement test, 2,521. In Exhibit 30 we present point estimates for all statistically significant moderation effects. Note that though statistically significant interactions were observed, none involved the treatment indicator. This suggests that STeLLA treatment effects did not differ significantly across key student subgroups.

Study-Specific Science Achievement Measure

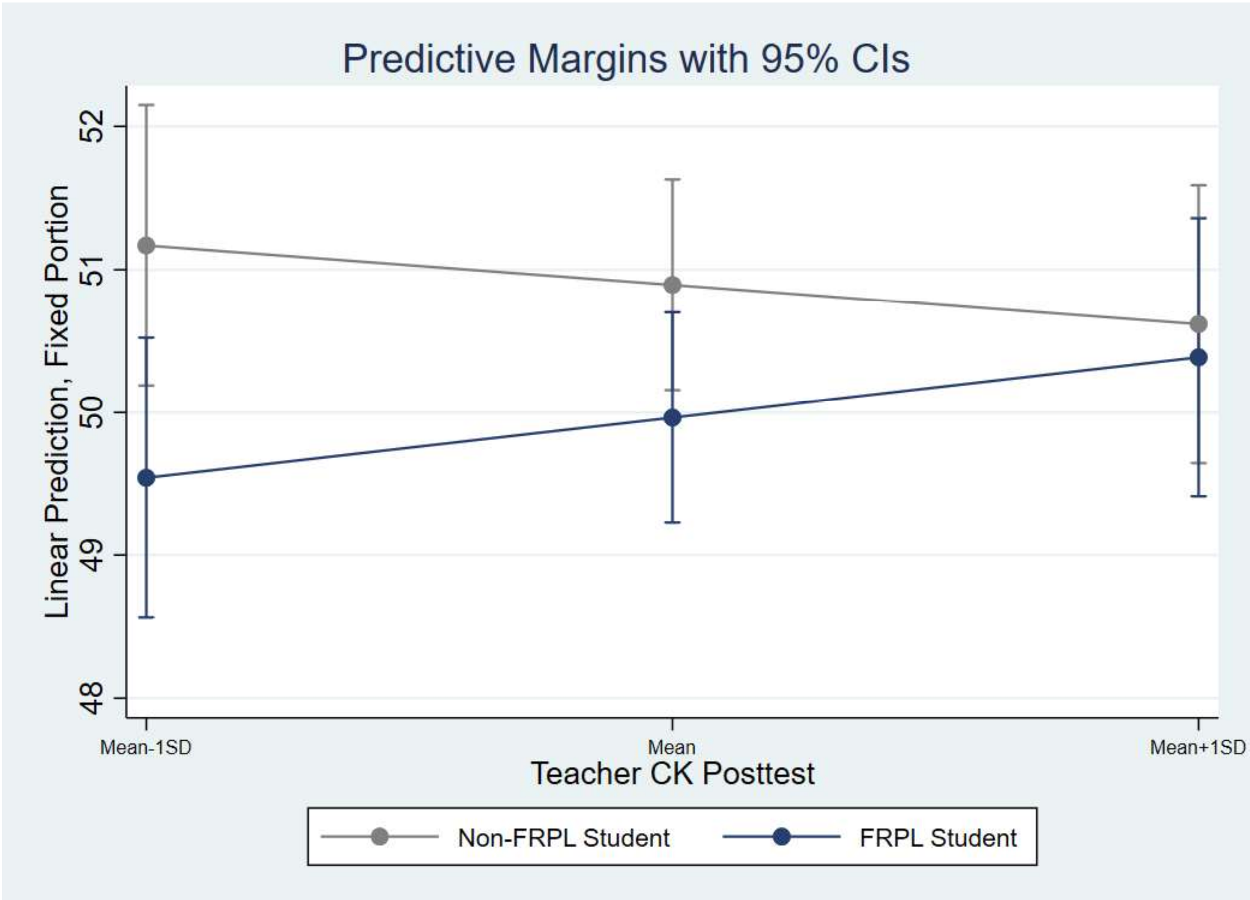
For the study-specific achievement measure, the model estimated that the effect of teacher content knowledge on student achievement is significantly larger for students who receive free or reduced-price lunch (Exhibit 31). In addition, the effect of class size on student achievement is significantly smaller if the student has a special education designation (Exhibit 32). Finally, the relationship between teacher content knowledge and student posttest is significantly larger when students' baseline knowledge is higher (Exhibit 33). For students with relatively higher pretest scores, the relationship between teacher content knowledge and student posttest is positive, whereas that same relationship is negative for students with relatively lower pretest scores.

Exhibit 30. Estimates of Moderating Effects on Student Achievement Outcomes

Interaction	Estimate (*SE*)	*N*
Study-specific science assessment		
Student FRPL × Teacher content knowledge	0.252	1,965
	(0.119)	
Student SPED × Class size	−0.207	1,965
	(0.082)	
Student pretest x Teacher content knowledge	0.023	1,965
	(0.007)	
State science achievement test		
Student SPED x Teacher content knowledge	−0.036	2,521
	(0.012)	

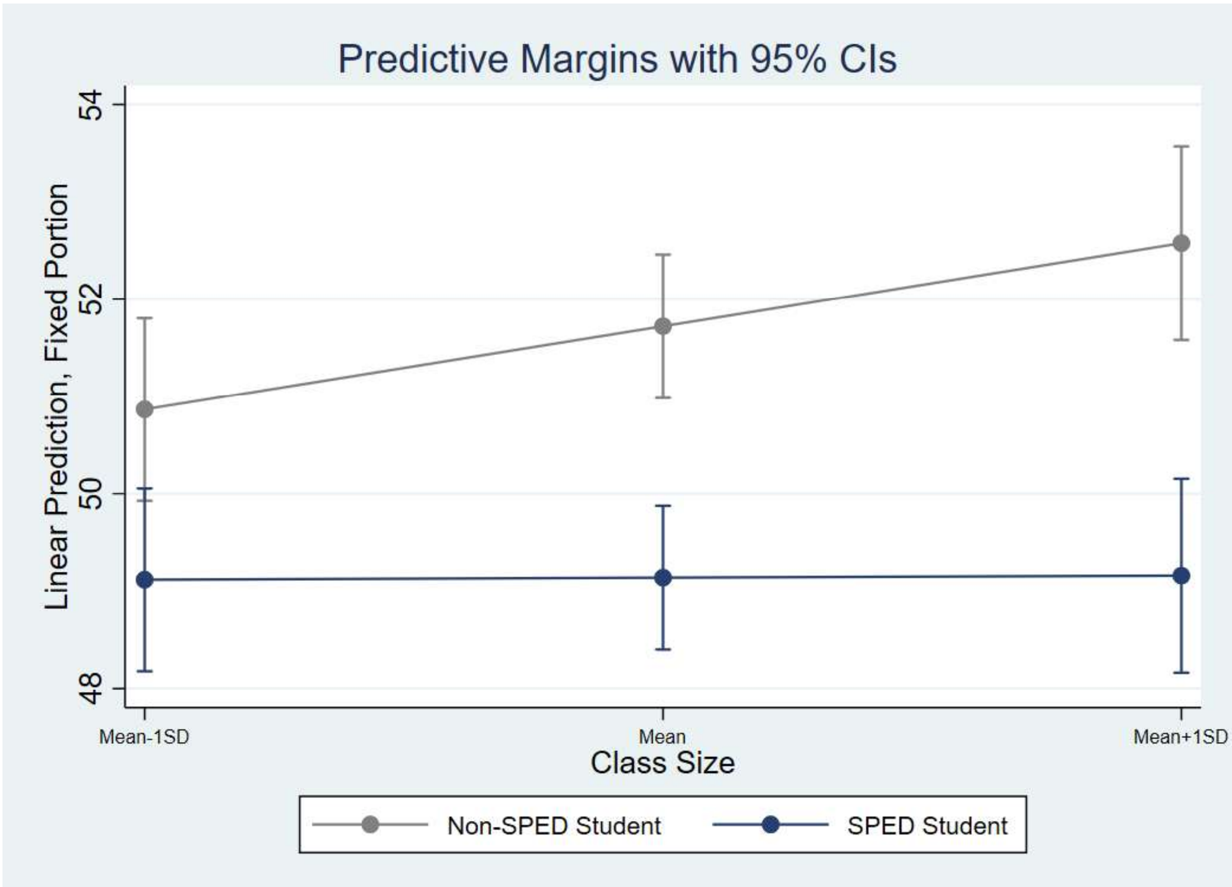
Note. *SE* = standard error; FRPL = free or reduced-price lunch; SPED = special education.

Exhibit 31. Effect of Teacher Content Knowledge on Study-Specific Science Assessment Scores by Free or Reduced-Price Lunch Status



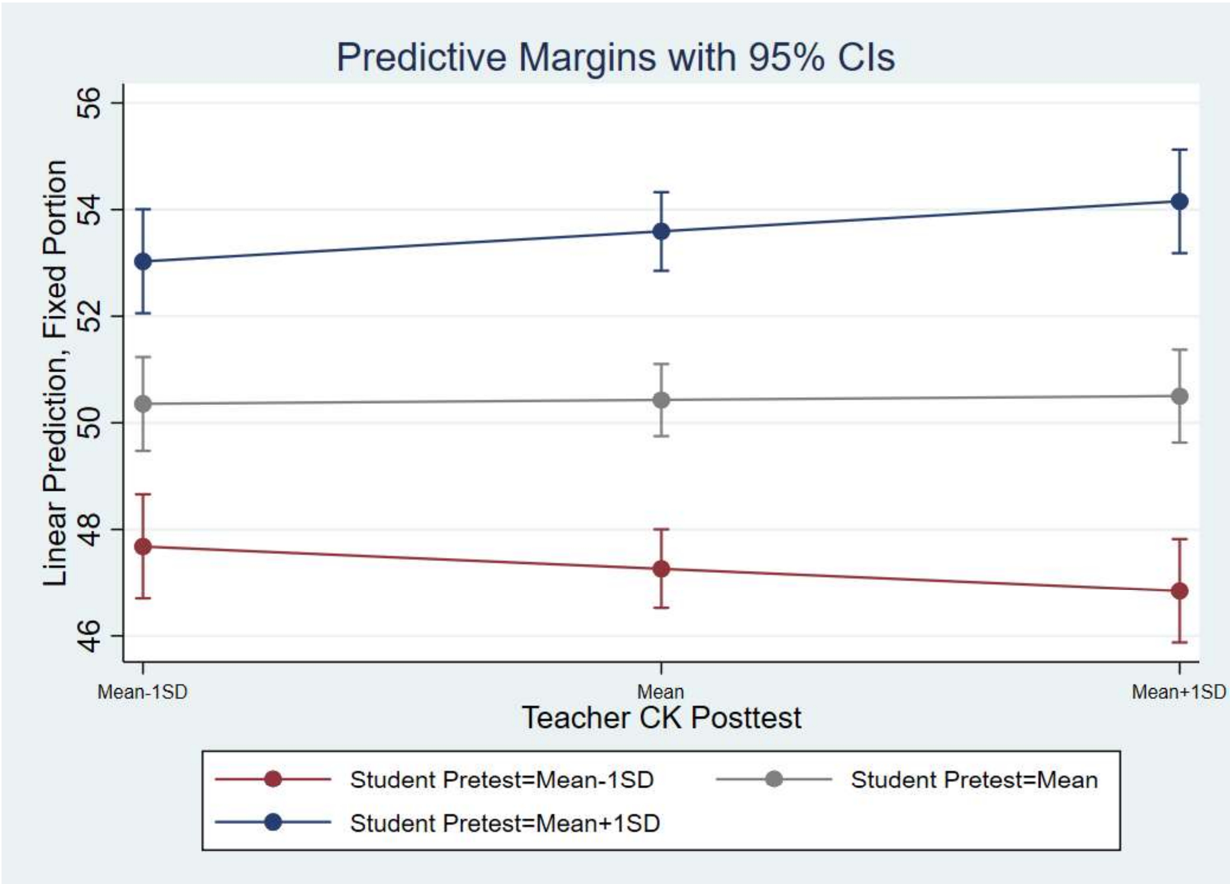
Note. CI = confidence interval; CK = content knowledge; FRPL = free or reduced-price lunch.

Exhibit 32. Effect of Class Size on Study-Specific Science Assessment Scores by Special Education Designation Status



Note. CI = confidence interval; SPED = special education.

Exhibit 33. Effect of Teacher Content Knowledge on Study-Specific Science Assessment Scores by Pretest

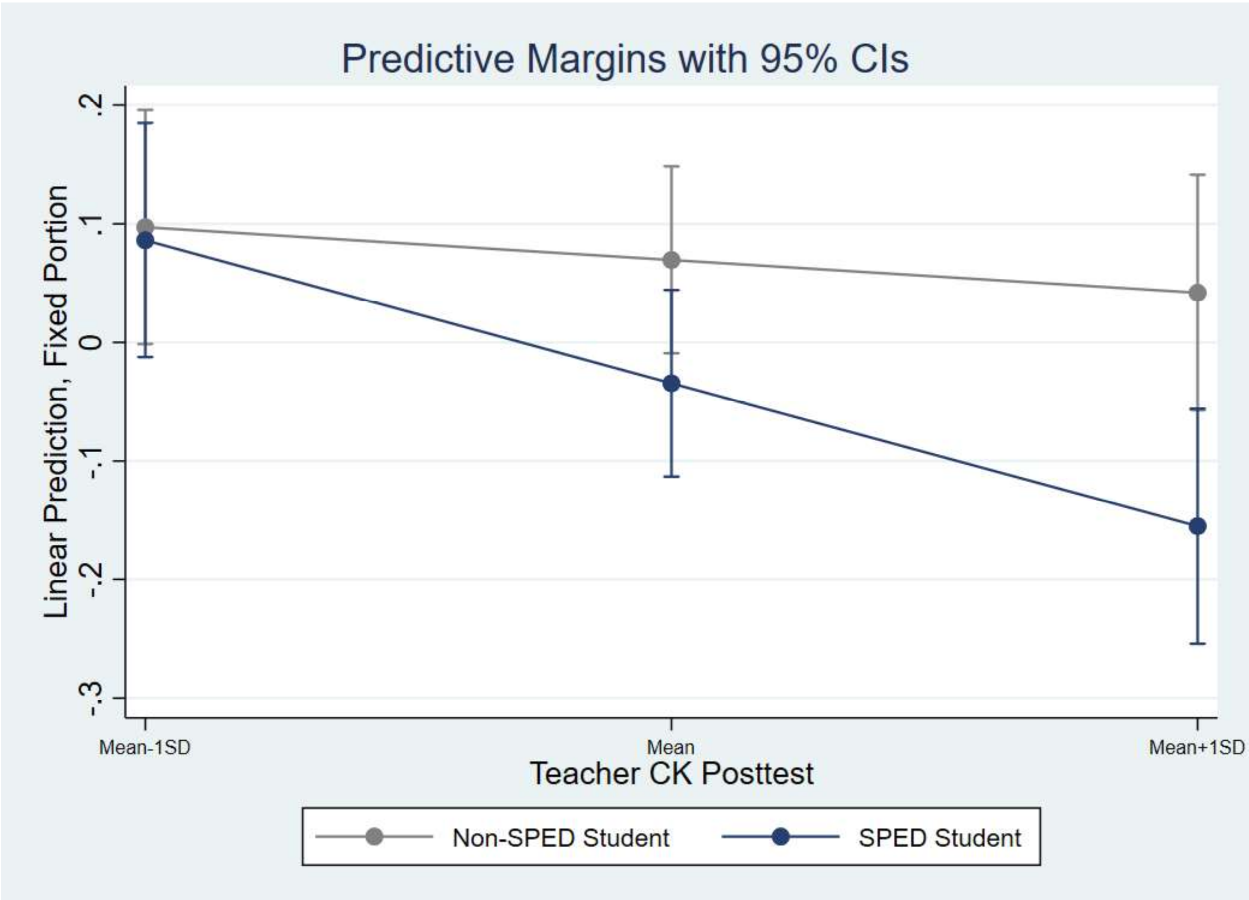


Note. CI = confidence interval; CK = content knowledge.

State Science Achievement Test

For the state science achievement test, the negative effect of teacher content knowledge is more negative for students with a special education designation (Exhibit 34).

Exhibit 34. Effect of Teacher Content Knowledge on State Science Scores by Special Education Designation Status



Note. CI = confidence interval; CK = content knowledge; SPED = special education.

Moderation of Teacher Outcomes

As with student outcomes, we explored the moderating effects of teacher baseline measures, years of teaching experience, and school locale for teacher outcomes. We used a sample including only teachers with complete data on outcomes, baseline measures, demographics, and treatment condition. The samples for teacher content knowledge, PCK, and practice outcomes included 77, 79, and 75 teachers, respectively. Our analyses did not find any statistically significant differential treatment effects across subgroups, indicating that impacts on teacher content knowledge, PCK, and practice were largely consistent across teachers with different baseline scores, various years of teaching experience, and rural or nonrural settings.

Mediation Effects

In the following tables, we provide mediation results for the study-specific science assessment and the state science achievement test. As exploratory analyses, the mediation tests used complete case (i.e., complete data on all covariates and mediators) analytic samples of 1,965 and 2,521, respectively. Please see Exhibit 17 in conjunction with these results.

Study-Specific Science Assessment

The *a* paths are the unstandardized treatment effects presented earlier for teacher content knowledge, PCK, and instructional practice. The *c′* and *b* paths were estimated from a three-level model nesting students within teachers and teachers within schools (see Appendix B). From that model, the mediated direct effect (*c′*) is 1.86 (p = .010).

To complete mediation analyses for the study-specific science assessment, the same complete case analytic sample of 1,965 was used to estimate an unmediated direct effect (*c*) of STeLLA (B = 1.92, p < .001), using the same covariates as the main effect model for the full sample. Each *ab* product is expressed as a percentage of the unmediated direct effect.

For the study-specific science assessment, we observed that mediation (indirect) effects are relatively weak and nonsignificant. This was largely due to the weak relationship between teacher and student outcomes (*b* paths). This result suggests that other constructs related to teachers' knowledge or practice, aside from what was measured by our three teacher outcomes, could be contributing to STeLLA's effect on student outcomes. See Exhibit 35 for full results.

Exhibit 35. Mediation Results for the Study-Specific Science Assessment

Indirect path	*a*	SE_a	*b*	SE_b	*ab* product	*p*
Through teacher CK (a_1,b_1)	2.49	0.89	−.003	.050	−0.007	.947
Through teacher PCK (a_2,b_2)	7.76	1.71	.026	.029	0.202	.371
Through teacher instructional practice (a_3,b_3)	14.51	2.99	−.010	.017	−0.145	.545

Note. *SE* = standard error; CK = content knowledge; PCK = pedagogical content knowledge.

State Science Achievement Test

The *a* paths are the unstandardized treatment effects presented earlier for teacher content knowledge, PCK, and instructional practice. The *c′* and *b* paths were estimated from a three-level model nesting students within teachers and teachers within schools (see Appendix B). From that model, the mediated direct effect (*c′*) is −0.15 (p = .011).

To complete mediation analyses for the study-specific assessment, the same complete case analytic sample of 2,521 was used to estimate an unmediated direct effect (*c*) of STeLLA (B = −0.04, p = .284), using the same covariates as the main effect model for the full sample. Each *ab* product is expressed as a percentage of the unmediated direct effect.

For the state science assessment, we observed that mediation (indirect) effects for content knowledge and practice are relatively weak and nonsignificant. This was largely due to the weak relationship between teacher and student outcomes (*b* paths). This result suggests that other constructs related to teachers' knowledge or practice, aside from what was measured by our three teacher outcomes, could be contributing to STeLLA's effect on student outcomes. The notable exception is the indirect path through teacher PCK. This indirect path was stronger and detected as a statistically significant mediation effect at the 5% level. That is, despite the overall effect of STeLLA being negative in direction, as teacher PCK scores increased, student scores also increased significantly. See Exhibit 36 for full results.

Exhibit 36. Mediation Results for the State Science Achievement Test

Indirect path	*a*	SE_a	*b*	SE_b	*ab* product	*p*
Through teacher CK (a_1,b_1)	2.49	0.89	−.002	.004	−0.005	.672
Through teacher PCK (a_2,b_2)	7.76	1.71	.008	.003	0.062	.002
Through teacher instructional practice (a_3,b_3)	14.51	2.99	.001	.002	0.015	.386

Note. *SE* = standard error; CK = content knowledge; PCK = pedagogical content knowledge.

Impact Evaluation Limitations

There are a few key limitations to note for the impact study. First, random assignment of the STeLLA condition results in comparing outcomes for treatment schools that are using STeLLA strategies for the very first time to those for comparison schools that are using extant curriculum and instruction, perhaps for an extended time. Thus, a limitation of the impact evaluation is that while STeLLA implementation for students spanned portions of one year, that may not be enough time for its impacts to manifest on distal outcome measures such as state standardized tests.

A second limitation is attrition. We were able to detect nearly all focal impacts as statistically significant. However, attrition at the cluster level, in particular, left us with a smaller sample size than anticipated and, consequently, less precision in impact estimates (i.e., wider confidence intervals). We also note that though student and teacher sample characteristics mirror those of the communities in which the study was conducted (Tennessee and Kentucky), they do not mirror the demographics of the nation more broadly. This reality limits the generalizability of the findings.

Finally, our mediation results suggest that teacher outcomes explain only a small portion of variation in student outcomes. As a result, we posit that other factors could be driving outcomes, such as the extant curriculum materials used in treatment and BaU control classrooms for topics within and outside STeLLA influence. We have sufficient information on the curriculum materials used for some research sites but not all. This leaves us with a potential confounding factor that could reduce our confidence in attributing treatment effects entirely to the STeLLA program.

Scale-Up Evaluation Study

Strategy to Scale

At the beginning of the project, BSCS developed a strategic plan to achieve the following three goals:

1. Reduce the cost of the STeLLA program by developing local leaders and ensuring that they are equipped to lead high-quality implementation at scale.
2. Increase flexibility in the format and timing of STeLLA to reduce cost and expand participation, especially in rural areas.
3. Develop the capacity of state and district partners to build infrastructure to sustain the STeLLA program beyond the life of the grant.

Next, we describe these goals and the associated scaling strategies in greater detail.

Goal 1. Build Local Leadership Capacity and Reduce Program Costs

A significant barrier to scaling STeLLA in the past was the program's reliance on BSCS staff to deliver the program. This approach was difficult to scale because of the limited number of BSCS staff available and geographic constraints. To address these barriers, BSCS implemented a local STeLLA PL leader development program whereby BSCS leadership development experts prepared and supported local PL leaders, who then led STeLLA PL for teachers. During the leadership development years of the study, local leaders completed a summer Leadership Institute and Learning Team meetings during the academic year. The focus of the leadership program was twofold. Participants were able to experience the STeLLA PL program as learners with a focus on the program experiences and then shift focus to the knowledge and skills needed to lead the program. The new leaders were then mentored by experienced BSCS staff and implemented the STeLLA program with their own groups of teachers.

Goal 2. Increase Flexibility and Expand Participation, Especially in Rural Areas

Teacher PL that relies on in-person delivery presents a significant barrier to scaling because it requires considerable travel from both teachers and PL leaders, demands scheduling at times when all participants are available, and limits the number of teachers a PL leader can support. BSCS's strategy for addressing this barrier was to implement a **blended in-person and virtual version of the STeLLA program**. This blended approach replicated the STeLLA in-person program with the same learning goals, key program features (Roth et al., 2017), and total program hours and took advantage of virtual delivery to allow for flexibility of location (especially for rural teachers), variability in the timing of individual participation through asynchronous opportunities, and lower costs of delivery and participation. The synchronous virtual learning opportunities (i.e., study group sessions) provided participants opportunities to learn from analysis of videos from each other's teaching of STeLLA lessons and from analysis of student work. The asynchronous learning opportunities (i.e., homework tasks) consisted of analyzing classroom videos to enhance participants' understanding of science content, student ideas about the content, and strategies to respond to student ideas and move student thinking forward. The virtual delivery drew on emerging design principles for virtual and blended PL in STEM education, including motivating and sustaining engagement that builds knowledge and advances professional goals; creating opportunities for teachers to collaborate as learners; and supporting reflection on content and practice (Kowalski et al., 2017; Roth et al., 2017).

Goal 3. Develop Capacity for Sustainability and Continued Scaling of the Program

A third challenge for scaling a PL program is implementing structures and supports to ensure that the program continues beyond the life of the initial source of funding (in this case, the EIR grant program). To develop capacity for sustainability and continued scaling of the STeLLA program, BSCS aimed to hold regular meetings with district leads, state leads, and project partners to develop plans for sustainability during and beyond the years of the study. Sustainability plans included considerations for ensuring that teachers who did not participate in the study and/or who were part of the comparison schools were able to access STeLLA PL opportunities after data collection concluded. By developing local leaders to serve as STeLLA program PL leaders, the likelihood of the program sustaining beyond the life of the grant also increased.

Mechanisms to Scale

When developing the scale-up strategy plan, BSCS and AIR identified several potential challenges, including limited local expertise, offering PL that can be easily accessed by teachers in rural areas and on an as-needed schedule, and the necessity for accessible support materials for local leaders and teachers, as well as state and district partners. To address these anticipated challenges, the scale-up strategy outlined mechanisms associated with each of the scale-up goals. These mechanisms included

- working with district leads to identify and recruit local leaders, implementing a leadership development program that prepares the leaders to implement the STeLLA program, and ensuring that leadership development materials are accessible in online format to all local leaders (Goal 1);
- holding virtual, synchronous study groups for teachers in the fall and spring semesters, while also providing asynchronous learning activities (i.e., homework tasks) in the fall and spring to provide extended learning opportunities, and ensuring that teacher materials are accessible in online format (Goal 2); and
- working with district and state partners to ensure that sustainability plans are in place, providing STeLLA PL opportunities to all teachers in comparison schools as well as to teachers in treatment schools who did not participate in the study, and ensuring that program materials are accessible to district and state partners in online format (Goal 3).

Together, these mechanisms formed a cohesive strategy to scale up the STeLLA PL program sustainably, ensuring quality and support while reducing costs.

Scale-Up Evaluation Study Design

The study addressed the following research question:

- Were the scale-up strategies implemented as intended?

To address the research question, AIR and BSCS identified key indicators and implementation thresholds associated with each aforementioned goal (Exhibit 37). In the sections that follow, we describe the measurement plan in greater detail.

Exhibit 37. STeLLA Scale-Up Goals, Challenges, Mechanisms for Scaling, and Measurement

Goal	Challenge	Mechanism	Indicator	Threshold	Data source(s)
Goal 1. Reduce cost of program by replacing program staff with local leaders who can deliver the STeLLA PL to teachers.	Local expertise and capacity might be limited. The COVID-19 pandemic also posed a challenge.	Work with regional partners and district leads to identify and recruit local leaders.	Number of local leaders recruited and developed.	Recruit 20 leaders and retain 15 for Cohort 1 and retain 10 for Cohort 2.	Program attendance records (Implementation Data—Component 1)
	Leading STeLLA effectively requires experiencing STeLLA as a learner and participating in significant leadership development.	Implement a leadership development program that prepares leaders to implement the STeLLA program.	Participation in leadership development activities (from attendance records).	Seventy-five percent of all leaders attend 75% of all leader events.	Program attendance records (Implementation Data—Component 1)
	Supporting local leaders requires comprehensive yet accessible PL leader guides and materials.	Develop online resources for leaders to make materials easily accessible.	Leader materials are made available (online or hard copy).	Online hub for leader materials is in place, and hard copies of leader materials are distributed.	Program materials on BSCS website
Goal 2. Increase flexibility in the format and timing and reduce cost of STeLLA PL to expand participation, particularly in rural areas.	Offering PL that can be easily accessed by teachers in rural areas and on an as-needed schedule.	Hold fall and spring study groups virtually and provide asynchronous learning activities.	Number of and participation in virtual sessions.	Seventy-five percent of all teachers attend 75% of all teacher events.	Program attendance records (Implementation Data - Component 1)
			Participants' reports on the effectiveness of virtual learning.	Average score on fall and spring survey items between 3 and 4.	Fall and spring teacher survey, Cohorts 1–2
		Develop online resources for teachers to make program	Online hub containing all necessary materials for teachers.	Online hub for teacher materials is in place.	Program materials on BSCS website

Goal	Challenge	Mechanism	Indicator	Threshold	Data source(s)
		materials easily accessible.			
Goal 3. Develop capacity for sustainability and continued scaling of the program.	Districts, states, and partner organizations need support in developing strategies and plans to sustain the program.	Hold regular meetings with district leads, state leads, and project partners to develop plans for sustainability.	Frequency of meetings with district leads, state leads, and project partners.	Sustainability-focused meetings occur at least twice a year.	BSCS meeting records, including conversations with partners
			Each district, state, and partner has a sustainability plan in place by the end of the project.	Plans are in place.	BSCS meeting records, including conversations with partners
	Provide access to STeLLA PL and materials to teachers who did not receive the STeLLA program as part of the treatment.	Work with local organizations and local PL leaders to provide STeLLA PL experiences to teachers assigned to the comparison group and to teachers in treatment schools who were not able to participate in the study.	STeLLA PL is offered to all teachers in the control group and to teachers in treatment schools who did not participate in the study.	Fifty percent of teachers who did not receive STeLLA programming as part of the study receive it prior to closing of the grant.	BSCS program records, including conversations with partners

Note. PL = professional learning.

Measurement of Implementation of Scaling Strategies

To assess whether the identified scale-up strategies were implemented as planned, BSCS and AIR gathered program data from 2020–24 (i.e., Local Leader Development; Cohort 1; Cohort 2) and administered surveys in the fall and spring of the 2022–23 and 2023–24 school years (Cohorts 1 and 2, respectively). For Goal 1, *Reduce cost of program by replacing program staff with local leaders who can deliver the STeLLA PL to teachers*, BSCS gathered program data on the number of local leaders who were recruited and developed as intended (i.e., completed at least 75% of all development activities), and who went on to serve as local leaders for either one or both cohorts of teachers.

For Goal 2, *Increase flexibility in the format and timing and reduce cost of STeLLA PL to expand participation, particularly in rural areas*, BSCS gathered program data on the number of Cohort 1 and 2 teachers who participated in virtual activities during the fall and spring semesters (i.e., 8 hours per semester, 16 hours total). Via survey, AIR asked teachers to report on the amount of time spent completing asynchronous homework tasks (i.e., four tasks per semester, eight tasks total, with "completion" assessed according to whether teachers reported spending time on the task). The survey also asked participants to report on the effectiveness of virtual learning activities. Specifically, participants responded to a series of seven items that asked them to rate the extent to which virtual study groups and/or homework tasks were helpful in developing their science knowledge and science teaching abilities. Each item was rated on a 4-point scale (1 = *not at all*, 2 = *small extent*, 3 = *moderate extent*, 4 = *large extent*). Survey participants also were asked how likely they were to recommend STeLLA to other teachers based on a 4-point scale (1 = *very unlikely*, 2 = *somewhat unlikely*, 3 = *somewhat likely*, 4 = *very likely*).

For Goal 3, *Develop capacity for sustainability and continued scaling of the program*, BSCS held recurring meetings with partners and district leads and kept a record of meeting dates and notes. BSCS also kept a record of when comparison group teachers were offered the opportunity to participate in STeLLA PL experiences and whether these teachers chose to participate.

For Goals 1–3, BSCS also developed an online hub that provided ongoing, free access to all materials associated with the project, including PL leader guides, lesson plans, classroom videos, and video analysis tools. In the section that follows, we describe the degree to which the scale-up strategy was implemented as intended.

Findings on Implementation of Scaling Strategies

By the end of the project, BSCS was able to implement the majority of the mechanisms outlined in the scale-up plan in order to achieve the following three goals:

- Build local leadership capacity and reduce program costs.
- Increase flexibility and expand access to STeLLA PL opportunities.
- Develop capacity for sustainability and continued scaling.

Next, we describe these findings in greater detail.

Goal 1. Build Local Leadership Capacity and Reduce Program Costs

The project aimed to reduce costs by recruiting and developing local leaders to deliver STeLLA PL instead of relying solely on BSCS program staff. The threshold was to recruit 20 leaders and retain 15 for Cohort 1 and 10 for Cohort 2. This threshold was met and even exceeded: 15 local leaders served Cohort 1, and 14 local leaders served Cohort 2.

For leadership development, the expectation was that 75% of all leaders would attend at least 75% of leader events. This expectation was achieved overall. For all 12 leadership development activities, participation exceeded the 75% threshold except for the Leadership Kickoff Series, for which only 11/16 local leaders (68%) met the expectation (see Exhibit A1 in Appendix A).

Finally, BSCS achieved the goal of making program materials easily accessible to program leaders. Hard copies of all materials were distributed to local leaders during the leadership development activities, and an online version was also available.

In summary, BSCS fully implemented all scale-up strategies for Goal 1.

Goal 2. Increase Flexibility and Expand Participation, Especially in Rural Areas

To expand STeLLA participation, especially in rural areas, the project offered virtual and asynchronous learning activities to all treatment teachers. BSCS aimed for at least 75% of all treatment teachers to attend 75% of all virtual activities offered. This goal was either met or exceeded:

- **Fall Virtual Study Group Activities:** 85% of teachers met the goal.
- **Fall Asynchronous Homework Tasks:** 96% of teachers met the goal.
- **Spring Virtual Study Group Activities:** 83% of teachers met the goal.
- **Spring Asynchronous Homework Tasks:** 81% of teachers met the goal.

Teachers' feedback on the effectiveness of virtual learning was positive, with average ratings indicating moderate to large impacts on science knowledge and teaching abilities, and a high likelihood of recommending the STeLLA program to others.

BSCS's goal of making program materials easily accessible for teachers was also met, with all materials available via Google Drive during the study and later through a permanent website.

In summary, BSCS fully implemented all strategies for Goal 2.

Goal 3. Develop Capacity for Sustainability and Continued Scaling

Goal 3 focused on sustainability and ongoing access to STeLLA PL opportunities. BSCS aimed to hold at least two sustainability-focused meetings per year with district and state leads and partners. This goal was achieved, with recurring monthly meetings and multiple formal meetings documented. BSCS also intended for each district, state, and partner to have a sustainability plan in place by the end of the project, and BSCS has confirmed that these plans were developed in most districts.

Another key indicator associated with Goal 3 was to offer STeLLA PL opportunities to at least 50% of teachers who did not receive it as part of the study (i.e., comparison teachers). This threshold was nearly met: 100% of BaU control teachers were offered STeLLA, and 46.97% participated. In addition, STeLLA was organically scaled to nonstudy teachers in participating districts, even though this opportunity was not part of the original plan. Finally, ongoing, free access to all project materials was provided to all study participants, and a website containing all curriculum created during the study is available as open access.

In summary, BSCS nearly implemented all strategies for Goal 3.

Exhibit 38 presents the overall findings on the implementation of scale-up strategies by the end of the project.

Exhibit 38. Overall Findings on Implementation of Scale-Up Strategies by End of Project

Goal	Scale-up strategy	Threshold for successful implementation	Findings on actual level of implementation	Implementation of strategy met or exceeded threshold (Yes/No)	If implementation of strategy did not meet threshold, possible reasons
Goal 1: Reduce cost of program by replacing program staff with local leaders who can deliver the STeLLA PL to teachers.	Work with district leads and partner organizations to identify and recruit local leaders.	Recruit 20 leaders and retain 15 for Cohort 1 and retain 10 for Cohort 2.	Sixteen local leaders participated. Of those, 15/16 local leaders served Cohort 1, and 14/16 local leaders served Cohort 2.	Yes	NA
	Implement a leadership development program that prepares leaders to implement the STeLLA program.	Seventy-five percent of all leaders attend 75% of all leader events.	Seventy-five percent of all leaders attended 75% of all activities. For all 12 activities, participation exceeded the 75% threshold except for the Leadership Kickoff Series, for which only 11/16 local leaders met the expectation.	Yes	NA
	Develop and provide online access to leader materials.	Materials are distributed.	Materials provided in person and available online.	Yes	NA
Goal 2: Increase flexibility in the format and timing and reduce cost of STeLLA PL to expand participation, particularly in rural areas.	Hold fall and spring virtual study groups and provide asynchronous learning activities.	Seventy-five percent of all teachers attend 75% of all teacher events.	Fall virtual activities: 85% of all teachers met the goal. Of teachers, 28/34 Cohort 1 teachers (82%) and 18/20 Cohort 2 teachers (90%) completed at least 75% of the fall virtual activities.	Yes	NA

Goal	Scale-up strategy	Threshold for successful implementation	Findings on actual level of implementation	Implementation of strategy met or exceeded threshold (Yes/No)	If implementation of strategy did not meet threshold, possible reasons
			Fall asynchronous homework tasks: 96% of all teachers met the goal. Of teachers, 34/34 Cohort 1 teachers (100%) and 18/20 Cohort 2 teachers (90%) completed at least 75% of the fall asynchronous homework activities. Spring virtual activities: 83% of all teachers met the goal. Of teachers, 28/34 Cohort 1 teachers (82%) and 17/20 Cohort 2 teachers (85%) completed at least 75% of the spring virtual activities. Spring asynchronous homework tasks: 81% of all teachers met the goal. Of teachers, 25/34 Cohort 1 teachers (74%) and 19/20 Cohort 2 teachers (95%) completed at least 75% of the spring asynchronous homework activities.		
		Average score on fall and spring survey items between 3 and 4.	On average, teachers reported that the fall and spring virtual activities impacted their science knowledge and science teaching abilities to a	Yes	NA

Goal	Scale-up strategy	Threshold for successful implementation	Findings on actual level of implementation	Implementation of strategy met or exceeded threshold (Yes/No)	If implementation of strategy did not meet threshold, possible reasons
			moderate to large extent (Cohort 1 average: 3.23; Cohort 2 average: 3.53; 4-point scale). On average, teachers reported that the fall and spring asynchronous homework activities impacted their science knowledge and science teaching abilities to a moderate extent (Cohort 1 average: 3.0; Cohort 2 average: 3.26; 4-point scale). On average, teachers reported that they were very likely to recommend STeLLA to other teachers (Cohort 1 average: 3.8; Cohort 2 average: 3.72; 4-point scale).		
	Develop online resources to make program materials easily accessible to teachers.	Online hub for teacher materials is in place.	Teachers have access to online materials.	Yes	NA

Goal	Scale-up strategy	Threshold for successful implementation	Findings on actual level of implementation	Implementation of strategy met or exceeded threshold (Yes/No)	If implementation of strategy did not meet threshold, possible reasons
Goal 3: Develop capacity for sustainability and continued scaling of the program.	Hold regular meetings with district leads, state leads, and project partners to develop plans for sustainability.	Sustainability-focused meetings occur at least twice a year, and sustainability plans are in place by the end of the project.	BSCS documented recurring monthly meetings and multiple formal meetings.	Yes	NA
	Work with local organizations and local PL leaders to provide STeLLA PL experiences to teachers assigned to the comparison group and to teachers in treatment schools who were not able to participate in the study.	Of teachers who did not receive STeLLA programming as part of the study, 50% receive it prior to closing of the grant.	Of control teachers, 100% were offered the option to receive STeLLA; 46.97% participated (31/66); 53.03% opted not to participate (35/66).	No	The study took place during the onset of the COVID-19 pandemic and the years immediately following. During this time, teachers may have been more inclined to refuse PL that was not required by their school or district.
	Develop online resources to make program materials easily accessible.	Online hub for districts/states is in place.	States/districts have access to online materials.	Yes	NA

Cost-Effectiveness Study

Study Description

The study team aimed to answer three cost-related research questions: (a) What are the costs associated with the STeLLA program as it was implemented for this particular study? (b) What are the costs under scaled implementation conditions in which the program operates at a higher capacity? (c) What is the cost-effectiveness of the STeLLA program in improving outcomes including student performance on science assessments and teacher content knowledge and practice?

To address these questions, we gathered data from multiple sources, including administrative and program records documenting the in-program personnel costs of BSCS staff, local leaders, and teachers engaged in STeLLA PL activities. We also developed surveys for local leaders and teachers to capture additional personnel costs associated with time spent on preparation and communication, as well as nonpersonnel costs such as travel, lodging, meals, and equipment (e.g., video recording devices). Information on student enrollment was collected from school-level records and teacher rosters.

To be included in the per student cost estimates, teachers must have had at least one student tested on the student assessment and responded to at least one of the three surveys administered during the course of each implementation cohort.[8] In order for teachers to be accounted for in per-teacher cost estimates, teachers must have completed the teacher assessment and at least one of the three cost surveys. Local leaders needed to complete the end-of-year survey in order to be included in the cost analysis.[9] Exhibit 39 summarizes the number of schools, local leaders, teachers, and students included. In Cohort 1, 16 schools met these criteria, and 12 schools in Cohort 2 met them. A total of 12 local leaders facilitated PL sessions across cohorts, though fewer local leaders participated in 2023–24 compared to 2022–23. Specifically, Cohort 2 had four fewer local leaders, and four of those participating lacked co-leaders, possibly contributing to higher individual workloads and costs.

[8] The team administered three surveys throughout the school year to teachers in each cohort. Surveys were conducted at the end of the summer institute and at the end of the fall and spring semesters. The survey questions asked about hours spent on after-session assignments for the two institutes; preparation for the small group sessions; time spent meeting or communicating with school administrators, BSCS staff, local leaders, and other STeLLA teachers about the program; and all out-of-pocket costs associated with attending the two institutes.

[9] The end-of-year local leader survey collected information on leaders' highest level of education; hours spent preparing for the two institutes and the small group sessions; and hours spent meeting or communicating with school administrators, BSCS staff, local leaders, and STeLLA teachers about the program.

Exhibit 39. Number of Schools, Local Leaders, STeLLA Teachers, and Students Included in the Cost Analysis by Cohort

Category	Cohort 1 (2022–23)	Cohort 2 (2023–24)
Schools	16	12
Local leaders	12	8
STeLLA teachers	31 [a]	19 [a]
Students	1,379	919

[a] The number of STeLLA teachers used to estimate per-teacher cost were 29 for Cohort 1 and 18 for Cohort 2.

The analytic approach employs the ingredients method (Levin et al., 2018), which systematically identifies and quantifies resources, assigns unit costs to the resources, and combines this information to calculate total program costs. Personnel resources included the time devoted by BSCS staff, local leaders, and teachers to summer and winter institutes, fall and spring study groups, classroom implementation, informal communication, and administrative tasks. Non-personnel resources included direct expenses such as travel, lodging, and meals to attend the summer and winter institutes, as well as video equipment for small group sessions.

For each program staff member involved (BSCS staff, local leaders, and teachers), we estimated the personnel cost by first calculating the share of annual working hours spent on the project's STeLLA activities. We monetized this time by multiplying the share of hours by the staff member's annual compensation (salary plus benefits), yielding the individual personnel cost of the program. For each program staff member, we added personnel and nonpersonnel costs to calculate the total individual program cost. Dividing the total cost by the number of students produced program per-pupil estimates, while dividing by the number of teachers participating in the STeLLA program produced per-teacher estimates. This calculation was done to capture the value of the resource effort used in both Pilot Years 1 and 2 in which the local leaders did leadership development (2020–21 and 2021–22), and the Implementation Cohorts 1 and 2 (2022–23 and 2023–24) in which teachers were developed and implemented what they learned in the classroom while engaging in ongoing support activities such as the small group sessions.

Because these teacher PL investments remain useful across a number of years, their economic costs are not immediately incurred but instead are realized during the period in which they are used. Therefore, the personnel and nonpersonnel costs were amortized to determine annualized costs across a set implementation period. This approach essentially spreads out these program costs across prespecified learning cycles or periods representing the "shelf life" during which these skills and equipment are assumed to remain useful.[10] We then computed the pupil-weighted average of the annualized costs of the teacher PL and small group sessions occurring in Implementation Cohorts 1 and 2. To obtain overall program costs, we added the annualized local leader development costs from Pilot Years 1 and 2 to the calculated pupil-weighted average cost across the two implementation cohorts.

The program as implemented was marked by significant underuse of local leaders' capacity in terms of the number of teachers that each led. As a result, local leaders supported fewer teachers (and consequently fewer students) than anticipated. We therefore conducted a simulation in which program costs included a higher caseload of teachers served by local leaders. Specifically, we adjusted observed local leader costs using a "full-capacity" ratio of teachers to local leader of 6:1, rather than the observed ratio of about 2.5:1.[11] We then compared the observed and simulated program costs to the traditional BSCS direct-delivery model across multiple cohorts to identify the break-even point at which the local leader delivery approach becomes less costly.

Finally, we calculated cost-effectiveness ratios associated with all the cost estimates by dividing both the observed and simulated costs by the student- and teacher-level impact estimates presented earlier. To assess the cost-effectiveness of the current study in context, we then compared these ratios to those reported in prior cost-effectiveness analyses of math achievement and teacher PL interventions.

Findings

The annualized program costs associated with Implementation Cohorts 1 and 2 broken out by resource type are reported in Exhibit 40. The first two columns show that the costs for the two Implementation Cohorts are quite similar, with the largest difference being the cost of local leaders, which is higher for Cohort 2. The third column lists the pupil-weighted average of the program costs across the two Implementation Cohorts, where the results suggest that the majority of the costs are associated with BSCS and teacher staff efforts, each of which is more than double the cost of local leaders. It is notable that the cost associated with local leaders increased in Implementation Cohort 2 (2023–24). This increase is attributed to a roughly 54% rise in reported local leader time outside planned program hours, including time spent preparing for the institutes and small group sessions, as well as additional communication with

[10] For STeLLA teachers, we assumed a 5-year learning cycle across which the first-year PL costs were amortized. For local leaders and BSCS staff, we used an 8-year cycle for costs incurred in the kickoff year and a 7-year cycle for costs from their development year, which occurred the following year. The longer learning cycle for local leaders reflects the assumption that their leadership development would remain useful for the full 5-year cycle of Cohort 2 STeLLA teachers, whose learning cycle ends 8 years after the kickoff year. The two participant groups' (Cohort 1 and Cohort 2) yearly costs were amortized using the 5-year learning cycle.
[11] The assumed full-capacity ratio was determined in conversation with BSCS program leadership.

STeLLA teachers. The results further suggest that the costs associated with program equipment are negligible. The average costs of the program across the Implementation Cohorts equal $104.61 per student or $4,874.62 per teacher.

Exhibit 40. STeLLA Program Costs Across Cohorts 1 and 2 Broken Out by Resource Type

Costs	Cohort 1 (2022–23)	Cohort 2 (2023–24)	Average across Cohorts 1 and 2
Per-pupil costs			
BSCS staff	$39.32	$36.14	$38.05
Local leaders	$14.22	$28.13	$19.78
STeLLA teachers	$43.99	$50.70	$46.67
Equipment	$0.09	$0.14	$0.11
Overall per-pupil cost	**$97.62**	**$115.10**	**$104.61**
Per-teacher costs			
BSCS staff	$1,807.66	$1,815.39	$1,810.75
Local leaders	$632.70	$1,360.94	$924.01
STeLLA teachers	$1,985.62	$2,358.30	$2,134.70
Equipment	$4.12	$6.72	$5.16
Overall per-teacher cost	**$4,430.10**	**$5,541.35**	**$4,874.62**

Note. The costs are inflation adjusted and are presented in 2023–24 dollars.

Exhibit 41 shows the annualized per-pupil and per-teacher program costs by staff type across the 4 study years (2020–21 through 2023–24). The first two columns report costs associated with Pilot Years 1 and 2, while the third column lists the average cost across Cohorts 1 and 2 presented earlier. The fourth column presents the sum of the costs across the pilot years and the pupil-weighted average across the two cohorts and represents the overall estimates of the program costs.

The first panel in the table presents the program costs per pupil. The overall cost of implementing the STeLLA program is $151.80 per pupil, with the largest share of the cost coming from BSCS staff (52%), followed by STeLLA teachers (31%), local leaders (18%), and equipment (< 1%). This overall per-pupil cost represents approximately 1% of the annual per-pupil expenditure in both Tennessee and Kentucky.[12] The second panel shows per-teacher costs, with the overall program cost amounting to $7,086.79 per teacher. As with the per-pupil

[12] We used per pupil expenditures for Kentucky and Tennessee from the National Center for Education Statistics' *Digest of Education Statistics: 2022*. The figures are based on the 2020–21 school year and have been standardized to the national average using CWIFT (Comparable Wage Index for Teachers) rates, then adjusted for inflation to reflect 2023–24 dollars. The per pupil expenditure was $16,176 for Kentucky and $13,613 for Tennessee.

results, the largest share of the per-teacher cost comes from BSCS staff (52%), followed by STeLLA teachers (30%), local leaders (18%), and equipment (< 1%).

Exhibit 41. STeLLA Program Costs Across Pilot Years and Implementation Cohorts as Implemented for Study, Broken Out by Resource Type

Costs	Pilot Year 1 2020–21	Pilot Year 2 2021–22	Average across Cohorts 1 and 2	Overall program cost (sum of pilot years and average across cohorts)
Per-pupil costs				
BSCS staff	$16.48	$23.77	$38.05	$78.30
Local leaders	$2.17	$4.71	$19.78	$26.27
STeLLA teachers	N/A	N/A	$46.67	$46.67
Equipment	N/A	$0.06	$0.11	$0.17
Overall per-pupil cost	**$18.65**	**$28.54**	**$104.61**	**$151.80**
Per-teacher costs				
BSCS staff	$775.39	$1,117.89	$1,810.75	$3,704.03
Local leaders	$99.61	$216.73	$924.01	$1,240.35
STeLLA teachers	N/A	N/A	$2,134.70	$2,134.70
Equipment	N/A	$2.55	$5.16	$7.72
Overall per-teacher cost	**$875.00**	**$1,337.18**	**$4,874.62**	**$7,086.79**

Note. The costs are inflation adjusted and are presented in 2023–24 dollars.

Exhibit 42 presents results of the sensitivity analysis that simulates costs under the alternative implementation scenario in which there is full utilization of local leaders' capacity. Whereas the figures presented in Exhibit 41 are based on the observed numbers of teachers served by local leaders at an average ratio of 2.5:1, the simulated alternative assumes a full-capacity utilization to be 6:1. The local leader costs across all years of implementation drop considerably after full capacity is reached, resulting in a decrease in the cost associated with local leader effort from $26.27 to $11.09 per pupil and from $1,240.35 to $547.08 per teacher (58% and 56% decline, respectively). This drop in local leader costs translates into an overall decrease in the cost of the program from $151.80 to $136.22 per pupil or $7,086.79 to $6,393.52 per teacher (10% decline in both). These results suggest that running the program at higher capacity with respect to the number of teachers served by local leaders could yield substantial savings in costs.

Exhibit 42. Estimated STeLLA Program Costs Across Pilot Years and Implementation Cohorts Under Full-Capacity Simulation Broken Out by Resource Type

Costs	Pilot Year 1 2020–21	Pilot Year 2 2021–22	Average across Cohorts 1 and 2	Overall program cost (sum of pilot years and average across cohorts)
Per-pupil costs				
BSCS staff	$16.48	$23.77	$38.05	$78.30
Local leaders	$0.93	$2.03	$8.13	$11.09
STeLLA teachers	N/A	N/A	$46.67	$46.67
Equipment	N/A	$0.06	$0.11	$0.17
Overall per-pupil cost	**$17.41**	**$25.85**	**$92.96**	**$136.22**
Per-teacher costs				
BSCS staff	$775.39	$1,117.89	$1,810.75	$3,704.03
Local leaders	$45.63	$99.27	$402.18	$547.08
STeLLA teachers	N/A	N/A	$2,134.70	$2,134.70
Equipment	N/A	$2.55	$5.16	$7.72
Overall per-teacher cost	**$821.01**	**$1,219.72**	**$4,352.79**	**$6,393.52**

Note. The costs are inflation adjusted and are presented in 2023–24 dollars.

To assess how and when costs can be saved through the scale-up of STeLLA, we compared the estimated per-pupil costs of the traditional BSCS direct-delivery model, whereby PL is delivered directly by BSCS staff and does not involve PL of local leaders,[13] with the per-pupil costs under the current implementation for the study reported earlier. The estimated cost of the BSCS direct-delivery model is $76.92 per pupil and is assumed to remain constant across STeLLA teacher cohorts. The cost reflects recurring personnel and nonpersonnel costs associated with BSCS instructors and participating teachers for each cohort served.

Exhibit 43 projects the observed costs from the current implementation, the estimated costs under the full-capacity model, and the estimated costs of the BSCS direct-delivery model across 10 STeLLA teacher cohorts. Estimated costs for both the current implementation and full-capacity local leader models decline across cohorts as fixed start-up costs associated with local leader PL are amortized across an increasing number of teachers and students served.[14]

[13] Estimated costs for the BSCS direct-delivery model are based on an assumed full-utilization capacity of six teachers per BSCS instructor and assume that participating teachers incur the same costs as those observed in the program implementation that is the focus of this current study.
[14] Note that the costs of teacher participation in STeLLA PL (resource costs associated with developing a new cohort of STeLLA teachers) recur and remain constant as part of the estimated per-pupil cost estimates.

For these local leader models, higher early cohort costs also reflect unplanned delays in program initiation and additional orientation activities during the COVID-19 period. These additional orientation activities increased personnel time and coordination demands for the local leaders and BSCS staff in the initial years, but these costs disperse and smooth out across subsequent cohorts. Additional cost reductions occur after the first two cohorts, under the assumption that BSCS staff are no longer involved in PL and administrative support beyond that point.

Relative to the BSCS direct-delivery model, the current study model and the full-capacity model initially exhibit higher per-pupil costs but rapidly decline across cohorts. As shown in Exhibit 43, the break-even point, defined as the number of cohorts at which the local leader delivery model becomes less costly than the BSCS direct-delivery model, occurs by the 10th cohort for the current implementation model and by the fourth cohort for the full-capacity model. The shorter time to reach the break-even point by the full-capacity model highlights the importance of fully utilizing local leaders to achieve cost efficiency.

Exhibit 43. STeLLA Program Per-Pupil Cost Estimates by Delivery Models Across Numbers of Cohorts

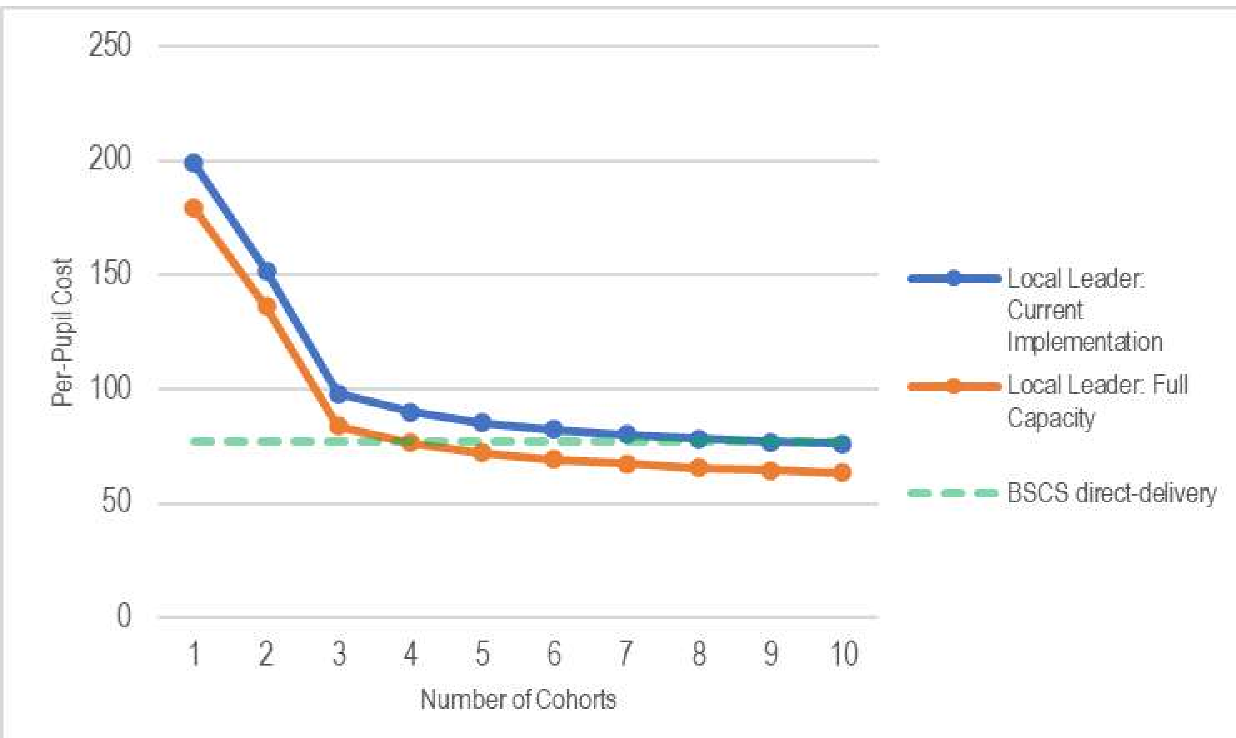


Exhibit 44 reports cost-effectiveness ratios using the observed cost estimates and impact estimates. At the student level, the program is estimated to cost $774.49 per pupil to produce a one standard deviation increase in science achievement scores. We did not calculate the cost-effectiveness ratio for the student state assessment scores because a negative impact estimate makes the ratio difficult to interpret in a meaningful way. At the teacher level, the program costs $14,581.88 to raise teacher content knowledge by one standard deviation, $7,980.62 to raise teacher PCK by one standard deviation, and $6,927.46 to raise teacher instructional practice quality by one standard deviation.

Exhibit 44. Cost-Effectiveness Ratios by Impact Estimates

Analysis Level/Outcome Measure	Cost per unit of analysis	Impact estimate	Cost-effectiveness ratio (cost per *SD*)
Student-level analysis			
Study-specific science assessment	$151.80	0.196	$774.49
Student state science assessment	NA	−0.130	$774.49
Teacher-level analysis			
Teacher content knowledge assessment	$7,086.79	0.486	$14,581.88
Teacher PCK assessment	$7,086.79	0.888	$7,980.62
Teacher instructional practice	$7,086.79	1.023	$6,927.46

Note. PCK = pedagogical content knowledge.

To contextualize these estimates, we compared STeLLA's student-level cost-effectiveness ratio to those reported in prior studies of interventions in math and teacher PL (Barrett & VanDerHeyden, 2020; Foster et al., 2013; Xu et al., 2023; Yeh, 2010). Exhibit 45 summarizes comparable cost-effectiveness ratios, with all cost figures inflation adjusted to match the year for which costs are estimated in the current study. Although the comparison impact estimates are based on math rather than science achievement outcomes, we present this comparison because, to our knowledge, no published cost-effectiveness analyses of science achievement exist.

Compared to prior interventions, STeLLA exhibits the lowest cost per standard deviation increase in student achievement at $774.49. Even when compared to teacher PL interventions such as the Appalachian Math and Science Partnership, which has the lowest per student implementation cost, STeLLA remains more cost-effective because of its proportionately larger impacts on student outcomes.

Exhibit 45. Cost-Effectiveness of STeLLA Program Compared to Interventions in Math Achievement and Teacher Professional Learning

Intervention	Cost per unit of analysis	Impact estimate	Cost-effectiveness ratio (cost per *SD*)
STeLLA program	$151.80 *(2)*	0.20 *(2)*	$774.49
Classwide math intervention	$308.57 *(3)*	0.18 *(3)*	$1,714.29
AMSP training: Partnership training of K–12 teachers by local college instructors	$68.67 *(1)*	0.03 *(5)*	$2,289.03
Obtaining a master's degree for teachers	$1,088.42 *(5)*	0.22 *(1)*	$4,947.36
MAF	$908.92 *(4)*	0.09 *(4)*	$9,773.31

Note. The costs are inflation adjusted and are presented in 2023–24 dollars. Values in parentheses indicate relative rankings across interventions based on effectiveness and efficiency. For the impact estimate, rankings reflect effectiveness and are ordered from largest to smallest effect size (e.g., 1 = largest impact). For cost per unit of analysis, rankings reflect efficiency and are ordered from lowest to highest cost (e.g., 1 = lowest cost).
AMSP = Appalachian Math and Science Partnership; MAF = Math Achievement Fund.

Conclusion

In this study we observed evidence that confirms portions of the STeLLA theory of action. Specifically, we observed positive and statistically significant impacts for teacher outcomes of content knowledge, PCK, and instructional practice, as well as for the study-specific student achievement test that was aligned with the focal content of the STeLLA spring module (Matter; Earth's Changing Surface). These positive effects mirror those of Taylor et al. (2017), although the magnitude of the positive effects was smaller for the study-specific student achievement test, perhaps a consequence of STeLLA delivery from less experienced facilitators.

Although STeLLA impact on the state science achievement test was framed as confirmatory, it had not been tested previously, so the negative result is challenging to interpret. There is much that could be producing a negative effect outside the causal effect of STeLLA. For example, the ITT student sample for the state test analysis included students in withdrawn sites, as well as students whose teachers did not teach with adequate fidelity. There are also differences in extant instructional materials across treatment conditions. The influence of these materials on student outcomes is not possible to estimate in this study. Finally, though we have confidence that most STeLLA teachers taught selected topical areas such as Matter and Earth's Changing Surface in a STeLLA-informed fashion (i.e., through a content storyline and student thinking lens), we have little knowledge of what happened for the remainder of the school year related to many other topics that were on the Tennessee and Kentucky state tests.

In terms of moderation, differential effects of STeLLA and the interplay of student characteristics and teacher outcomes were tested. For both the study-specific and state science assessments, the vast majority of these effects were nonsignificant, and those that were significant involved student characteristics and teacher outcomes. These interactions should be explored more deeply with subsequent analyses. However, our primary concern was interactions between treatment condition and student characteristics. For the state test outcome, we found one such interaction between STeLLA assignment and being a non-White student. This interaction suggested that the negative effect of STeLLA on students' state test scores was yet more negative if students were non-White. This result is concerning and should be investigated more deeply.

Regarding mediation, the results are mixed toward confirming the mediational aspects of the logic model. That is, the effects of STeLLA on teacher outcomes are consistently positive, and the effects of STeLLA on the study-specific assessment outcome are consistently positive. However, the correlations between the teacher and student outcomes remain weak, suggesting that the positive causal effects of STeLLA are happening through mechanisms (e.g., teacher knowledge or abilities) that we are not yet directly measuring.

The scaling and implementation study concluded that implementation by local leaders and teachers reached desired levels of fidelity (i.e., composite fidelity score of 75% or better) for all three fidelity components—Component 1: BSCS identifies, develops, and supports local leaders; Component 2: Local leaders deliver the STeLLA two-lens framework blended through a virtual and in-person PL program; and Component 3: Treatment teachers implement STeLLA in their classrooms. By the end of the project, BSCS fully implemented all scaling strategies, successfully building local leadership capacity, expanding flexible and accessible PL opportunities (especially for rural teachers), and establishing sustainability plans and open access to program materials. All key participation and implementation thresholds were met or exceeded for each scale-up goal, with the exception of providing STeLLA PL opportunities to 50% or more of the nontreatment teachers (100% of nontreatment teachers were offered the PL, but only 46.97% participated).

The cost analysis indicates that program costs were similar across Cohorts 1 and 2, with the main difference being higher local leader costs in Cohort 2 due to increased reported time effort. On average, the program cost $104.61 per student or $4,352.79 per teacher across the two implementation cohorts, and $151.80 per pupil or $6,393.52 per teacher when local leader development costs from the pilot years are added. Most costs were associated with BSCS staff and STeLLA teachers, with local leaders accounting for a smaller share and equipment costs being negligible. Sensitivity analysis shows that operating at full local leader capacity would reduce both per-pupil and per-teacher costs by 10%. Cost-effectiveness analysis indicates that the program cost $774.49 per pupil for a one standard deviation increase in student achievement, and at the teacher level, $14,581.88 for content knowledge, $7,980.62 for PCK, and $6,927.46 for instructional practice quality.

In sum, this study found that although STeLLA was implemented by nonexperts at a reasonable cost, high levels of implementation fidelity were achieved and significant positive impacts were observed for the teacher and student study-specific science outcome measures used in both the current study and previous STeLLA studies. The effects in the current study on those outcome measures are generally consistent with those of previous studies, providing evidence of effect replication. This study also observed a negative and significant impact of STeLLA on state science test scores. This finding is perplexing and one that warrants deeper analyses of potential treatment effect confounds. Further, notable heterogeneity of block-specific treatment effects on both student outcomes suggests that a deeper examination of site-specific STeLLA implementation could provide important insight into the relationship between implementation and outcomes. The study results also show that though STeLLA had strong effects on teacher outcomes, those teacher outcome scores have generally weak correlations with student outcomes, suggesting further study of STeLLA's mechanisms of impact.

References

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). Report of the 2018 NSSME+. *Horizon Research, Inc.*

Barrett, C. A., & VanDerHeyden, A. M. (2020). A cost-effectiveness analysis of classwide math intervention. *Journal of School Psychology*, 80, 54-65.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). Effective teacher professional development. *Learning policy institute*.

Foster, J. M., Toma, E. F., & Troske, S. P. (2013). Does teacher professional development improve math and science outcomes and is it cost effective?. *Journal of Education Finance*, 255-275.

Kowalski, S., Busey, A., Goldsmith, L., Bates, M., Beilstein, S., & Perry, M. (2017). *Emerging design principles for online and blended teacher professional development in K–12 STEM*. Community for Advancing Discovery Research in Education.

Levin, H., McEwan, P., Belfield, C., Bowden, A., & Shand, R. (2018). *Economic evaluation in education*. SAGE Publications, Inc, https://doi.org/10.4135/9781483396514

National Research Council, Division of Behavioral, Social Sciences, Board on Science Education, & Committee on a Conceptual Framework for New K-12 Science Education Standards. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education*, *25*(2), 177-196.

Roth, K. J., Bintz, J., Wickler, N. I., Hvidsten, C., Taylor, J., Beardsley, P. M., Caine, A., & Wilson, C. D. (2017). Design principles for effective video-based professional development. *International Journal of STEM Education*, *4*(1), 31.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312.

Taylor, J., Roth, K., Wilson, C., Stuhlsatz, M., & Tipton, E. (2017). The effect of an analysis-of-practice, videocase-based, teacher professional development program on elementary students' science achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 241–271.

What Works Clearinghouse (WWC). (2022). *What Works Clearinghouse procedures and standards handbook, Version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Handbooks

Wilson, S. M. (2013). Professional development for science teachers. *Science*, *340*(6130), 310-313.

Yeh, S. S. (2010). The cost effectiveness of 22 approaches for raising student achievement. *Journal of Education Finance*, 38-75.

Xu, Z., Özek, U., Levin, J., & Hoon Lee, D. (2023). Effects of Large-Scale Early Math Interventions on Student Outcomes: Evidence From Kentucky's Math Achievement Fund. *Educational Evaluation and Policy Analysis*, 01623737241303142.

Appendix A. Implementation Fidelity Indicator Scores

Exhibit A–1. Local Leaders' Participation in Leadership Development Activities

Leadership development activity	Intended dosage	Priority category	Average percentage of hours completed (*n* = 16 local leaders)
Kickoff leadership development series	40 hours	Medium	68
Summer orientation meeting	3 hours	Low	94
Summer institute	40 hours	High	94
Summer institute debrief	5 hours	High	88
Summer leadership institute	36 hours	High	81
Fall lesson analysis study groups	8 hours	High	88
Fall leader reflection sessions	4 hours	Medium	88
Winter institute	16 hours	High	88
Winter institute debrief	2 hours	Medium	88
Spring lesson analysis study groups	4 hours	High	84
Spring leader reflection sessions	2 hours	Medium	94
Spring leader planning session(s)	24 hours	High	94

Exhibit A–2. Component 2 Average Indicator Percentages by Cohort and Study Group

Cohort and study group number	Indicator 1: Summer institute (40 hours) *Percentage of total possible hours*	Indicator 2: Fall Homework tasks (4 tasks) *Percentage of total possible tasks*	Indicator 3: Fall lesson analysis (8 hours) *Percentage of total possible hours*	Indicator 4: Winter institute (16 hours) *Percentage of total possible hours*	Indicator 5: Spring lesson analysis (8 hours) *Percentage of total possible hours*	Indicator 6: Spring HW tasks (4 tasks) *Percentage of total possible tasks*	Indicator 7: Quality of implementation *Avg. percentage of points possible for fall and spring*
1-1	100	57	93	86	86	36	95
1-2	100	100	92	100	83	33	71
1-3	100	80	100	100	90	80	62
1-4	100	94	100	100	97	100	94
1-5	100	100	100	100	86	100	65
1-6	100	67	100	100	75	67	85
1-7	100	100	100	100	100	100	79
Cohort 1 average	**100**	**85**	**98**	**98**	**88**	**74**	**79**
2-3	100	100	100	100	100	100	81
2-5	100	100	100	100	100	100	85
2-6	100	100	100	100	88	100	70
2-7	100	80	95	100	85	95	55
2-8	100	100	81	100	63	88	85
2-9	100	67	83	100	67	100	84
Cohort 2 average	**100**	**91**	**93**	**100**	**84**	**97**	**77**
Average of averages	**100**	**88**	**96**	**99**	**86**	**85**	**78**

Exhibit A–3. Component 3 Average Indicator Percentages by Cohort and Study Group

Cohort and study group number	Indicator 1: Fall lesson sequence	Indicator 2: Spring lesson sequence	Indicator 3: Implementation quality—Spring lesson sequence
	Percentage of total possible points	*Percentage of total possible points*	*Percentage of total possible points*
1-1	80.49	81.45	100
1-2	84.80	84.80	100
1-3	77.12	86.36	100
1-4	81.05	85.11	87.50
1-5	88.55	85.45	100
1-6	75.45	86.97	100
1-7	86.36	90.00	100
Cohort 1 average	**81.98**	**85.74**	**98**
2-3	70.00	81.36	100
2-5	85.45	83.74	75
2-6	83.64	87.27	50
2-7	81.27	82.64	100
2-8	82.50	83.64	100
2-9	84.24	82.12	100
Cohort 2 average	**81.18**	**83.46**	**87.50**
Average of averages	**81.58**	**84.60**	**92.86**

Appendix B. Description of Analytic Approaches

Model Specifications

B.1. Main Effects for Student Outcomes

We used a design with a purposive sample of district blocks, with generalization of effects to the current sample of blocks, and random assignment of schools within each block. As a result, we ran a model that produced a fixed and precision-weighted average treatment effect and fixed block effects. The model, presented below, is a two-level hierarchical linear model, in which the effect of treatment on the outcome measure is estimated at the cluster (school) level, controlling for pretest, a set of other baseline student/teacher covariates, and baseline school covariates. The *B*-1 dummy variables representing the nonreference blocks produced fixed intercepts for blocks defined by cohort and school district, as well as number of teachers per school, urbanicity, and/or other school characteristics. The model produced an overall average estimate of the treatment effect as a precision-weighted mean of the within-block T-C differences. The overall treatment effect estimate (β_1) is a precision-weighted estimate because the weights used in the calculation are inversely proportional to variance of the treatment effect in each block. Note that the following model applies to estimating impacts on both the study-specific student science outcome measure and the state science test outcome measure.

$$Y_{ik} = \beta_0 + \beta_1 T_k + \beta_2 X_{ik} + \beta_3 Z_k + \sum_{b=2}^{B} \delta_b \, Block_{bk} + u_k + \varepsilon_{ik}$$

Y_{ik} is the student outcome measure (e.g., student study-specific science assessment posttest score or state science achievement in the intervention year) for student *i* in school *k*;

β_0 is the intercept (for school k), or average outcome for the control school students in the reference level of student and school covariates;

β_1 represents the average difference in outcomes between treatment schools and control schools, conditional on student covariates, school covariates, and block fixed effects;

T_k is a school-level indicator for treatment assignment, equal to 1 for treatment schools and 0 for control schools;

β_2 is a vector of coefficients associated with student-level covariates;

X_{ik} is a vector of student-level characteristics, including student pretest measure (e.g., student study-specific science assessment pretest score or state science achievement in the prior year) and demographic characteristics (e.g., grade level, gender, race/ethnicity, eligibility for free or reduced-price lunch, English learner status, eligibility for special education);

β_3 is a vector of coefficients associated with school-level covariates;

Z_k is a vector of school-level characteristics, including student pretest scores aggregated to the school level, the percentage of non-White students, the percentage of students eligible for free or reduced-price lunch, the percentage of English learners, the percentage of students eligible for special education, the percentage of students in Grade 4, and total school enrollment (school size);

δ_b represents the fixed effect associated with block *b*;

$Block_{bk}$ is an indicator equal to 1 if school *k* belongs to block *b* and 0 otherwise; block fixed effects are included for all blocks except the reference block;

u_k and ε_{ik} are the school- and student-level random error terms, respectively.

B.2. Main Effects for Teacher Outcomes

$$Y_{ik} = \beta_0 + \beta_1 T_k + \beta_2 X_{ik} + \beta_3 Z_k + \sum_{b=2}^{B} \delta_b \, Block_{bk} + u_k + \varepsilon_{ik}$$

Y_{ik} is the teacher outcome measure (e.g., teacher content knowledge assessment posttest score, PCK assessment posttest score, instructional practice posttest score) for teacher *i* in school *k*;

β_0 is the intercept (for school k), or average outcome for the control school teachers in the reference level of teacher and school covariates;

$\beta_1$represents the average difference in outcomes between treatment schools and control schools, conditional on teacher covariates, school covariates, and block fixed effects;

T_k is a school-level indicator for treatment assignment, equal to 1 for treatment schools and 0 for control schools;

β_2 is a vector of coefficients associated with teacher-level covariates;

X_{ik} is a vector of teacher-level characteristics, including teacher pretest measure (e.g., teacher content knowledge assessment pretest scores, PCK assessment pretest scores, or instructional practice pretest scores), years of teaching experience, and class size, as well as student demographic characteristics and pretest scores aggregated to the teacher level (e.g., the percentage of female students, the percentage of non-White students, the percentage of students eligible for free or reduced-price lunch, and average student science achievement in the prior year);

β_3 is a vector of coefficients associated with school-level covariates;

Z_k is a vector of school-level characteristics, including teacher pretest scores aggregated to the school level, average years of teaching experience, the percentages of non-White and female students, the percentage of students eligible for free or reduced-price lunch, average student science achievement in the prior year, total school enrollment (school size), and a rural-school indicator;

δ_b represents the fixed effect associated with block *b*;

$Block_{bk}$ is an indicator equal to 1 if school *k* belongs to block *b* and 0 otherwise; block fixed effects are included for all blocks except the reference block;

u_k and ε_{ik} are the school- and teacher-level random error terms, respectively.

B.3. Moderation Effects for Student Outcomes

$$Y_{ijk} = \beta_0 + \beta_1 T_k + \beta_2 X_{ijk} + \beta_3 Z_k + \beta_4 A_{jk} + \sum_{n=1}^{30} M_{\mathrm{n}}(A_{jk} * X_{ijk}) + \sum_{n=1}^{6} P_{\mathrm{n}}(T_k * X_{ijk}) + \sum_{b=2}^{B} \delta_b \, Block_{bk} + u_k + r_{jk} + \varepsilon_{ijk}$$

Y_{ijk} is the student outcome measure (e.g., student study-specific science assessment posttest score or state science achievement in the intervention year) for student *i* of teacher *j* in school *k*;

β_0 is the intercept (for school k), or average outcome for the control school students in the reference level of student and school covariates;

β_1 represents the average difference in outcomes between treatment schools and control schools, conditional on student covariates, teacher covariates, school covariates, and block fixed effects;

T_k is a school-level indicator for treatment assignment, equal to 1 for treatment schools and 0 for control schools;

β_2 is a vector of coefficients associated with student-level covariates;

X_{ijk} is a vector of student-level characteristics, including student pretest measure (e.g., student study-specific science assessment pretest score or state science achievement in the prior year) and demographic characteristics (e.g., grade level, gender, race/ethnicity, eligibility for free or reduced-price lunch, English learner status, eligibility for special education);

β_3 is a vector of coefficients associated with school-level covariates;

Z_k is a vector of school-level characteristics, including student pretest scores aggregated to the school level, the percentage of non-White students, the percentage of students eligible for free or reduced-price lunch, the percentage of English learners, the percentage of students eligible for special education, the percentage of students in Grade 4, and total school enrollment (school size);

β_4 is a vector of coefficients associated with teacher-level covariates;

A_{jk} is a vector of teacher-level characteristics, including teacher posttest outcome measures (e.g., teacher content knowledge posttest scores, PCK posttest scores, or instructional practice posttest scores), years of teaching experience, and class size;

M_n represents the interaction effect associated with 30 multiplicative combinations of student- and teacher-level covariates ($A_{jk} * X_{ijk}$);

$A_{jk} * X_{ijk}$ represent the 30 multiplicative combinations of student- and teacher-level covariates;

P_n represents the interaction effect associated with six multiplicative combinations of treatment and student-level covariates ($T_k * X_{ijk}$);

$T_k * X_{ijk}$ represent the six multiplicative combinations of treatment and student-level covariates (race, sex, language learner status, special education status, free or reduced-price lunch status, baseline achievement);

δ_b represents the fixed effect associated with block *b*;

$Block_{bk}$ is an indicator equal to 1 if school *k* belongs to block *b* and 0 otherwise; block fixed effects are included for all blocks except the reference block;

u_k, r_{jk}, and ε_{ijk} are the school-, teacher-, and student-level random error terms, respectively.

B.4. Moderation Effects for Teacher Outcomes

$$Y_{ik} = \beta_0 + \beta_1 T_k + \beta_2 X_{ik} + \beta_3 Z_k + \beta_4(T_k * Pretest_{ik}) + \beta_5(T_k * Experience_{ik}) + \beta_6(T_k * Rural_k) + \sum_{b=2}^{B} \delta_b\, Block_{bk} + u_k + \varepsilon_{ik}$$

Y_{ik} is the teacher outcome measure (e.g., teacher content knowledge assessment posttest score, PCK assessment posttest score, instructional practice posttest score) for teacher *i* in school *k*;

β_0 is the intercept (for school k), or average outcome for the control school teachers in the reference level of teacher and school covariates;

β_1 represents the average difference in outcomes between treatment schools and control schools, conditional on teacher covariates, school covariates, and block fixed effects;

T_k is a school-level indicator for treatment assignment, equal to 1 for treatment schools and 0 for control schools;

β_2 is a vector of coefficients associated with teacher-level covariates;

X_{ik} is a vector of teacher-level characteristics, including teacher pretest measure (e.g., teacher content knowledge assessment pretest scores, PCK assessment pretest scores, or instructional practice pretest scores), years of teaching experience, and class size, as well as student demographic characteristics and pretest scores aggregated to the classroom level (e.g., the percentage of female students, the percentage of non-White students, the percentage of students eligible for free or reduced-price lunch, and average student science achievement in the prior year);

β_3 is a vector of coefficients associated with school-level covariates;

Z_k is a vector of school-level characteristics, including teacher pretest scores aggregated to the school level, average years of teaching experience, the percentages of non-White and female students, the percentage of students eligible for free or reduced-price lunch, average student science achievement in the prior year, total school enrollment (school size), and a rural-school indicator;

β_4 represents the interaction effect associated with the interaction between treatment and teacher pretest measure;

$T_k * \ Pretest_{ik}$ represents the interaction between treatment and teacher pretest measure;

β_5 represents the interaction effect associated with the interaction between treatment and teachers' years of teaching experience;

$T_k * Experience_{ik}$ represents the interaction between treatment and teachers' years of teaching experience;

β_6 represents the interaction effect associated with the interaction between treatment and the rural-school indicator;

$T_k * Rural_k$ represent the interaction between treatment and the rural-school indicator;

δ_b represents the fixed effect associated with block *b*;

$Block_{bk}$ is an indicator equal to 1 if school *k* belongs to block *b* and 0 otherwise; block fixed effects are included for all blocks except the reference block;

u_k and ε_{ik} are the school- and teacher-level random error terms, respectively.

B.5. Mediation Effects for Student Outcomes

Mediation was tested using a total of four models. The first three models estimated the impact of STeLLA on three proposed mediators/teacher outcomes of content knowledge, PCK, and instructional practice. The model specifications for these "*a* paths" are given earlier under **Main Effects for Teacher Outcomes**. The fourth model, which follows, is the three-level model used to estimate the b_1, b_2, b_3, and *c′* paths—that is, the effect of STeLLA on student outcomes, controlling for the mediator effects, and the correlations between the mediator values and student outcomes.

$$Y_{ijk} = \beta_0 + \beta_1 T_k + \beta_2 X_{ijk} + \beta_3 Z_k + \beta_4 A_{jk} + \sum_{b=2}^{B} \delta_b \, Block_{bk} + u_k + r_{jk} + \varepsilon_{ijk}$$

Y_{ijk} is the student outcome measure (e.g., student study-specific science assessment posttest score or state science achievement in the intervention year) for student *i* of teacher *j* in school *k*;

β_0 is the intercept (for school k), or average outcome for the control school students in the reference level of student and school covariates;

β_1 represents the average difference in outcomes between treatment schools and control schools, conditional on student covariates, teacher covariates, school covariates, and block fixed effects (i.e., the *c′* path);

T_k is a school-level indicator for treatment assignment, equal to 1 for treatment schools and 0 for control schools;

β_2 is a vector of coefficients associated with student-level covariates;

X_{ijk} is a vector of student-level characteristics, including student pretest measure (e.g., student study-specific science assessment pretest score or state science achievement in the prior year) and demographic characteristics (e.g., grade level, gender, race/ethnicity, eligibility for free or reduced-price lunch, English learner status, eligibility for special education);

β_3 is a vector of coefficients associated with school-level covariates;

Z_k is a vector of school-level characteristics, including student pretest scores aggregated to the school level, the percentage of non-White students, the percentage of students eligible for free or reduced-price lunch, the percentage of English learners, the percentage of students in Grade 4, and total school enrollment (school size);

β_4 is a vector of coefficients associated with teacher-level mediators (i.e., the *b* paths) and additional covariates;

A_{jk} is a vector of teacher-level characteristics, including the hypothesized mediators of teacher content knowledge posttest scores, PCK posttest scores, and instructional practice posttest scores, as well as additional covariates of years of teaching experience, and class size;

δ_b represents the fixed effect associated with block *b*;

$Block_{bk}$ is an indicator equal to 1 if school *k* belongs to block *b* and 0 otherwise; block fixed effects are included for all blocks except the reference block;

u_k, r_{jk}, and ε_{ijk} are the school-, teacher-, and student-level random error terms, respectively.

Appendix C. Supplemental Results

Exhibit C–1. Baseline Equivalence Assessment, Student Study-Specific Assessment

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Student level										
Student pretest	1,640	0.19	6.50	1,458	−0.21	6.84	−0.61	0.61	0.12	−0.091
Percent female students	1,640	0.01	0.49	1,458	−0.01	0.49	−0.01	0.02	0.00	−0.020
Percent AIAN students	1,640	0.00	0.07	1,458	0.00	0.05	0.00	0.00	0.00	0.000
Percent Asian students	1,640	0.00	0.13	1,458	0.00	0.15	0.01	0.01	0.01	0.072
Percent Black students	1,640	−0.02	0.28	1,458	0.02	0.32	0.04	0.02	0.15	0.133
Percent Hispanic students	1,640	−0.04	0.27	1,458	0.04	0.36	0.05	0.03	0.19	0.158
Percent Pacific Islander students	1,640	0.00	0.00	1,458	0.00	0.06	0.00	0.00	0.00	0.000
Percent White students	1,640	0.05	0.41	1,458	−0.06	0.46	-0.10	0.04	0.27	−0.229
Percent multiracial students	1,640	0.01	0.15	1,458	−0.01	0.10	0.00	0.00	0.04	0.000
Percent fourth graders	1,640	0.01	0.50	1,458	−0.02	0.50	0.03	0.09	0.51	0.059
Percent students eligible for ELL	1,640	−0.02	0.18	1,458	0.02	0.27	0.04	0.02	0.09	0.176
Percent students eligible for SPED	1,640	−0.01	0.33	1,458	0.01	0.34	0.02	0.01	0.02	0.060

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Percent students eligible for FRPL	1,640	-0.01	0.46	1,458	0.01	0.47	0.03	0.03	0.26	0.064
Cluster level										
Student pretest aggregated to school level	21	0.26	2.04	24	−0.23	3.19	−0.83	0.79	−0.305	−0.300
Percent non-White students aggregated to school level	21	−0.05	0.15	24	0.05	0.28	0.09	0.05	0.384	0.377
Percent ELL students aggregated to school level	21	−0.02	0.03	24	0.01	0.09	0.04	0.02	0.532	0.523
Percent SPED students aggregated to school level	21	−0.01	0.04	24	0.01	0.09	0.03	0.02	0.420	0.413
Percent FRPL students aggregated to school level	21	−0.01	0.27	24	0.01	0.26	0.03	0.04	0.129	0.127
Percent fourth graders aggregated to school level	21	0.01	0.35	24	−0.01	0.45	0.03	0.11	0.075	0.074
School size	21	26.84	216.08	24	−23.49	180.82	−53.28	52.87	−0.269	−0.264

Note. AMD = adjusted mean difference; *SE* = standard error; ICC = intraclass correlation coefficient; AIAN = American Indian and Alaska Native; ELL = English language learner; SPED = special education; FRPL = free or reduced-price lunch.

Exhibit C–2. Baseline Equivalence Assessment, Student State Achievement

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Student level										
Student pretest	2,103	0.05	0.96	2,500	−0.04	1.03	−0.12	0.08	0.09	−0.120
Percent female students	2,103	0.00	0.49	2,500	0.00	0.49	−0.01	0.02	0.00	−0.020
Percent AIAN students	2,103	0.00	0.08	2,500	0.00	0.05	−0.01	0.00	0.00	−0.153
Percent Asian students	2,103	−0.01	0.13	2,500	0.00	0.16	0.01	0.01	0.02	0.068
Percent Black students	2,103	−0.03	0.28	2,500	0.03	0.35	0.04	0.02	0.15	0.125
Percent Hispanic students	2,103	−0.03	0.29	2,500	0.03	0.36	0.05	0.03	0.17	0.151
Percent Pacific Islander students	2,103	0.00	0.00	2,500	0.00	0.07	0.00	0.00	0.00	0.000
Percent White students	2,103	0.06	0.42	2,500	−0.05	0.47	−0.09	0.04	0.25	−0.200
Percent multiracial students	2,103	0.01	0.14	2,500	−0.01	0.08	0.00	0.00	0.04	0.000
Percent fourth graders	2,103	0.00	0.50	2,500	0.00	0.50	0.01	0.07	0.49	0.020
Percent students eligible for ELL	2,103	−0.02	0.20	2,500	0.02	0.27	0.04	0.02	0.09	0.166
Percent students eligible for SPED	2,103	−0.01	0.35	2,500	0.01	0.37	0.02	0.01	0.02	0.055
Percent students eligible for FRPL	2,103	−0.02	0.46	2,500	0.01	0.47	0.05	0.03	0.25	0.107

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Cluster level										
Student pretest aggregated to school level	23	0.06	0.32	29	-0.05	0.36	−0.12	0.10	−0.360	−0.354
Percent non-White students aggregated to school level	23	−0.05	0.14	29	0.04	0.28	0.08	0.05	0.362	0.357
Percent fourth graders aggregated to school level	23	0.01	0.36	29	−0.01	0.42	0.02	0.09	0.043	0.043
Percent ELL students aggregated to school level	23	−0.02	0.04	29	0.01	0.09	0.04	0.02	0.473	0.466
Percent SPED students aggregated to school level	23	−0.02	0.07	29	0.01	0.08	0.03	0.02	0.411	0.404
Percent FRPL students aggregated to school level	23	−0.02	0.28	29	0.01	0.25	0.06	0.04	0.235	0.231
School size	23	13.55	216.45	29	−10.75	216.18	−18.02	53.46	−0.083	−0.082

Note. AMD = adjusted mean difference; *SE* = standard error; ICC = intraclass correlation coefficient; AIAN = American Indian and Alaska Native; ELL = English language learner; SPED = special education; FRPL = free or reduced-price lunch.

Exhibit C–3. Baseline Equivalence Assessment, Teacher Content Knowledge Assessment

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Teacher level										
Teacher pretest	44	0.36	5.03	40	−0.40	6.38	−1.63	1.09	0.12	−0.285
Teacher years of experience	44	0.86	7.78	40	−0.95	6.14	−2.16	1.39	0.05	−0.306
Class size	44	−1.53	10.98	40	1.69	18.04	3.75	3.07	0.72	0.251
Percent female students aggregated to teacher level	44	0.00	0.08	40	0.00	0.08	0.01	0.02	0.27	0.125
Percent non-White students aggregated to teacher level	44	−0.06	0.16	40	0.06	0.31	0.08	0.05	0.91	0.324
Percent FRPL students aggregated to teacher level	44	−0.02	0.22	40	0.02	0.25	0.03	0.04	0.87	0.126
Student pretest aggregated to teacher level	44	0.04	0.24	40	−0.04	0.38	−0.09	0.07	0.65	−0.283
Cluster level										
Teacher pretest aggregated to school level	17	0.89	3.48	20	−0.76	5.17	−1.97	1.25	−0.440	−0.430
Teacher years of experience aggregated to school level	17	1.29	6.51	20	−1.10	5.19	−2.58	1.94	−0.441	−0.432
School size	17	29.45	217.73	20	−25.03	164.36	−56.35	62.51	−0.296	−0.289
Percent non-White students aggregated to school level	17	−0.06	0.15	20	0.05	0.29	0.09	0.06	0.390	0.382

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Percent FRPL students aggregated to school level	17	0.00	0.28	20	0.00	0.27	0.02	0.04	0.071	0.069
Student state achievement pretest score aggregated to school level	17	0.02	0.23	20	−0.02	0.36	−0.08	0.10	−0.246	−0.241
Percent female students aggregated to school level	17	0.00	0.06	20	0.00	0.08	0.00	0.02	0.060	0.058
Percent rural schools	17	0.02	0.51	20	−0.01	0.51	0.01	0.12	0.018	0.018

Note. AMD = adjusted mean difference; *SE* = standard error; ICC = intraclass correlation coefficient; AIAN = American Indian and Alaska Native; ELL = English language learner; SPED = special education; FRPL = free or reduced-price lunch.

Exhibit C–4. Baseline Equivalence Assessment, Teacher Pedagogical Content Knowledge Assessment

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Teacher level										
Teacher pretest	44	−0.68	8.04	41	0.74	8.05	1.48	1.68	0.26	0.183
Teacher years of experience	44	0.86	7.88	41	−0.92	6.19	−2.19	1.40	0.06	−0.308
Class size	44	−1.53	11.09	41	1.64	17.90	3.64	3.05	0.70	0.244
Percent female students aggregated to teacher level	44	0.00	0.07	41	0.00	0.08	0.00	0.02	0.37	0.000
Percent non-White students aggregated to teacher level	44	−0.06	0.15	41	0.07	0.31	0.08	0.05	0.92	0.328
Percent FRPL students aggregated to teacher level	44	−0.02	0.22	41	0.02	0.25	0.03	0.04	0.88	0.126
Student pretest aggregated to teacher level	44	0.05	0.24	41	−0.05	0.38	−0.10	0.07	0.68	−0.314
Cluster level										
Teacher pretest aggregated to school level	18	−0.41	6.55	21	0.35	7.55	0.59	2.44	0.083	0.081
Teacher years of experience aggregated to school level	18	1.28	6.55	21	−1.10	5.38	−2.62	1.85	−0.440	−0.431
School size	18	23.50	219.89	21	−20.14	161.03	−46.45	60.16	−0.244	−0.239
Percent non-White students aggregated to school level	18	−0.06	0.15	21	0.05	0.29	0.09	0.06	0.369	0.361

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Percent FRPL students aggregated to school level	18	−0.01	0.28	21	0.01	0.27	0.03	0.04	0.121	0.119
Student state achievement pretest score aggregated to school level	18	0.03	0.22	21	−0.03	0.36	−0.10	0.10	−0.317	−0.311
Percent female students aggregated to school level	18	0.00	0.06	21	0.00	0.08	0.00	0.02	0.056	0.055
Percent rural schools	18	0.02	0.51	21	−0.01	0.51	0.01	0.11	0.017	0.017

Note. AMD = adjusted mean difference; *SE* = standard error; ICC = intraclass correlation coefficient; FRPL = free or reduced-price lunch.

Exhibit C–5. Baseline Equivalence Assessment, Teacher Instructional Practice

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Teacher level										
Teacher prescore	41	−3.99	18.11	40	4.09	15.97	7.50	3.32	0.25	0.437
Teacher years of experience	41	0.55	7.44	40	−0.56	6.22	−1.80	1.41	0.07	−0.262
Class size	41	−1.56	11.40	40	1.60	17.95	3.05	3.16	0.69	0.202
Percent female students aggregated to teacher level	41	0.00	0.07	40	0.00	0.08	0.01	0.02	0.18	0.133
Percent non-White students aggregated to teacher level	41	−0.06	0.16	40	0.07	0.31	0.08	0.05	0.90	0.322
Percent FRPL students aggregated to teacher level	41	−0.01	0.23	40	0.01	0.24	0.02	0.04	0.88	0.084
Student pretest aggregated to teacher level	41	0.03	0.23	40	−0.03	0.38	−0.08	0.07	0.68	−0.253
Cluster level										
Teacher prescore aggregated to school level	18	−4.04	16.75	21	3.47	10.29	7.06	3.86	0.517	0.507
Teacher years of experience aggregated to school level	18	0.88	6.37	21	−0.75	5.36	−1.86	1.84	−0.319	−0.312
School size	18	23.50	219.89	21	−20.14	161.03	−46.45	60.16	−0.244	−0.239
Percent non-White students aggregated to school level	18	−0.06	0.15	21	0.05	0.29	0.09	0.06	0.369	0.361

Measure	Comparison group			Treatment group			AMD	*SE* for AMD	ICC	Effect size
	Sample size	*M*	*SD*	Sample size	*M*	*SD*				
Percent FRPL students aggregated to school level	18	−0.01	0.28	21	0.01	0.27	0.03	0.04	0.121	0.119
Student state achievement pretest score aggregated to school level	18	0.03	0.22	21	−0.03	0.36	−0.10	0.10	−0.317	−0.311
Percent female students aggregated to school level	18	0.00	0.06	21	0.00	0.08	0.00	0.02	0.056	0.055
Percent rural schools	18	0.02	0.51	21	−0.01	0.51	0.01	0.11	0.017	0.017

Note. AMD = adjusted mean difference; *SE* = standard error; ICC = intraclass correlation coefficient; FRPL = free or reduced-price lunch.

About the American Institutes for Research®

Established in 1946, the American Institutes for Research (AIR) is a nonpartisan, not-for-profit institution that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit AIR.ORG.

AIR Headquarters
1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | **AIR.ORG**

29646_01/26