THE UNIVERSITY OF MEMPHIS | The Center for Research in Educational Policy (CREP)

Smithsonian Science for North and South Carolina Classrooms: *Improving Student Achievement across State Borders and State Standards*

Initial Summative and Main Findings Report

Todd Zoblotsky, Ed.D.

Robert McKinney, M.S.

Carolyn Kaldon, Ph.D.

Mojtaba Khajeloo, Ph.D.

Yu Wu, Ph.D.

Ivysmeralys Morales, M.A.

July 23, 2024

This document represents the initial Summative and main findings Report incorporating available data across the four years of the study prepared by the Center for Research in Educational Policy (CREP) at the University of Memphis for the Smithsonian Science Education Center (SSEC) to provide a summative evaluation for *Smithsonian Science for North and South Carolina Classrooms.* The evaluation is led by Dr. Todd Zoblotsky, Principal Investigator, Dr. Carolyn Kaldon, Co-PI, and by Project Manager Robert McKinney. The statistical analysis for this project was designed by Dr. Todd Zoblotsky, with support from Dr. Brenda McSparrin-Gallagher. Additional support was provided by Chelsea Jones and Ivysmeralys Morales, CREP Graduate Assistants, and Research Assistant Professors Dr. Mojtaba Khajeloo and Dr. Yu Wu. The evaluation team would also like to recognize the leadership and contributions of Dr. Christine Bertz. Data collection for this project is ongoing, and results reported here are preliminary.

Table of Contents

List of Tables

List of Figures

Executive Summary

The goal of *Smithsonian Science for North and South Carolina Classrooms* is to provide teachers with ongoing, differentiated professional development and research-based curricular materials to improve elementary student achievement in science, mathematics, and reading. This work used an inquiry-based science curriculum, *Smithsonian Science for the Classroom,* which is aligned with the Next Generation Science Standards (NGSS), and was implemented in Grades 3-5 within predominantly rural North and South Carolina school districts serving high-needs students. The Smithsonian Science Education Center (SSEC) partnered with the North Carolina Science, Mathematics, and Technology Education Center and South Carolina's Coalition for Mathematics and Science for this project, which is also being supported by Carolina Biological Supply Company, the manufacturer of the curricular modules. For this study, 37 schools (including one split cohort being treated as one school) in seven districts within North and South Carolina were randomly assigned to treatment and comparison groups. The study cohort, followed over three academic years, is composed of more than 1,600 third grade (2020-21) students in these schools.

The current document represents the initial Summative and main findings Report incorporating and summarizing available data across the four years of the study, which includes the first year of site recruitment (2019-20 school year) and three years of implementation (2020-21 through 2022-23 school years). The final summative report will include qualitative data collected during the fifth and final year of the grant (2023-24 school year).

Project implementation largely occurred as planned during the 2022-23 school year (the final year of implementation), and in-person activities resumed after several years of disruptions from the COVID-19 pandemic. Third through fifth grade teachers from treatment schools received in-person content-focused professional development (PD) for the first time in summer 2023, after two sessions of virtual PD (spring and summer 2021) and one session of hybrid PD (summer 2022). Teachers then implemented the physical science curriculum for the first time and engineering modules they received previously throughout the 2022-23 school year. CREP collected PD Evaluation Surveys and Module Logs following these PD activities.

As the **pretests** for achievement outcomes, students took the Abbreviated Battery of the Stanford Achievement Test Series, Tenth Edition (Stanford-10) in Reading and Math in spring of 2021 (the cohort's third grade year). The **main findings and posttests** for the study included the combined sample (i.e., both states combined) for (a) the Stanford-10 Science subtest, which was administered in the spring of 2023 (the cohort's fifth grade year), and (b) spring 2023 standardized state assessment scores in Reading and Math. The effect of *Smithsonian Science* for the three main findings was (a) **positive** (g = 0.18) **and statistically significant** (p = .032) on the **Stanford-10 science assessment** and (b) **positive** for both **Reading** (g = 10) and **Math** (g = 16), meaning students in the treatment group performed better than the control group, but the effects were **not statistically significant**.

Based on **fidelity standards**, professional development and curricular support were provided to schools with **high fidelity for both components in Year 1 and Year 3**, while **neither component was implemented with fidelity in Year 2**. Readers should note that in Year 2 (the 2021-22 school year), instruction in schools was severely disrupted due to the COVID-19 pandemic.

Only six items were common across all four **Professional Development (PD) surveys** and could be compared across all module (Engineering or Physical Science), training (Introductory or Intermediate), and delivery (virtual, hybrid, or fully in-person) types. Readers should note that the PD for each combination of module and type (for example, introductory Engineering) was only delivered once. Additionally, PD delivery progressed from fully virtual at the beginning of the project, to hybrid in the middle, to fully in-person at the end. These results, therefore, cannot be interpreted as being caused by any one specific factor.

Across the five questions about **facilitator performance**, the intermediate (i.e., content-focused) PD responses were slightly *less favorable* than the introductory (curriculum-focused) PD responses. However, this effect was not consistent across all modules. For the question about the **science content of the unit**, the percentage of positive responses was, on average, stable between introductory and intermediate PD. When averaged across PD type (introductory and intermediate), responses to the **Physical Science** PD were slightly *more positive* than responses to the **Engineering** PD. However, these differences are relatively small, and their causal factor is unclear. Across both PDs and grade levels, responses were primarily positive for all question categories. In addition, many questions had close to 100% positive responses, but with a very small sample size.

On the **module logs**, across all three years, teachers consistently reported they (a) had all necessary materials, (b) felt comfortable with the science content of the modules, and (c) had sufficient training to teach the modules. Over time, teachers became less likely to report that (a) they taught lessons in the suggested sequence, (b) modules were easy to use, (c) modules could comfortably fit in a class period, and (d) they taught the module during instructional time not intended for science. Meanwhile, teachers became more likely to report that they supplemented the lessons with materials from other sources.

On the **School Observation Measure (SOM)**, for **treatment and control schools** who participated in classroom observations during the 2021-22 and 2022-23 academic years, the most prevalent strategy observed both years was "Direct Instruction." Additionally, the overall classroom environments were similar between treatment and control schools. However, **treatment and control schools had a large divergence on "Experiential hands-on learning "** in the last year of the study, which was "*Extensively/Frequently observed*" over 60% more often in treatment vs. control schools (37% vs. 23% of the time, respectively). On the **Rubric for Inquiry Based Assessment (RIBA)**, two of the three activities most frequently observed over both years were the same for treatment and control schools: "Students engaged in experimentation" and "Students gathering or recording evidence". Meanwhile, over both years, "Prepared science kits or modules in use" was observed more frequently in treatment schools while "Students hypothesizing or making predictions" was observed more frequently in control schools. The **level of class time dedicated to inquiry-based science** was rated as "high" four times as often in control schools in the first year of the study, but over 50% higher in treatment classroom observations in the last year of the study.

Findings from the **teacher focus groups** should be interpreted with caution as they only represent 19 teachers across both years from six of the seven districts and 20 out of 36 schools, limiting the representativeness of their responses. Participants mentioned lingering impacts from COVID on the general loss of students' reading and mathematics comprehension and skills, and lack of prior science knowledge. Since the pandemic, multiple teachers in the **treatment group** also agreed there was an

impact on increasing teacher's confidence in teaching science. In addition, several **treatment teachers** agreed that engaging students in inquiry-based learning or hands-on activities helped students fully comprehend the content and increased inquiries, helping retain the knowledge of the concepts. Several teachers in both states and both years spoke about the state-tested nature of science in their school being a determining factor for the structure of their science teaching. Teachers in both years also mentioned issues with alignment between the modules and their state standards and having difficulty fitting the module in the limited time allotted for science instruction.

Introduction

In 2019, the Smithsonian Science Education Center (SSEC), a division of the Smithsonian Institution, received a five-year, $4 million Early-Phase Education Innovation and Research (EIR) grant (PR# U411C190055) from the U.S. Department of Education, Office of Elementary and Secondary Education, for *Smithsonian Science for North and South Carolina Classrooms* (*Smithsonian Science*). The goal of the project is to provide teachers with ongoing, differentiated professional development (PD) and research-based curricular materials to improve elementary student achievement in science, mathematics, and reading. This work utilizes an inquiry-based science curriculum aligned with the Next Generation Science Standards (NGSS), *Smithsonian Science for the Classroom*, and is being implemented in Grades 3-5 within predominantly rural North and South Carolina school districts serving high-needs students. To implement Smithsonian Science, the SSEC has partnered with the North Carolina Science, Mathematics, and Technology Education Center (NC SMT) and South Carolina's Coalition for Mathematics and Science (SCCMS). The *Smithsonian Science for the Classroom* curriculum was developed by the SSEC and is manufactured and distributed by Carolina Biological Supply Company.

The Center for Research in Educational Policy (CREP) at the University of Memphis, a State of Tennessee Center of Excellence, is the independent third-party evaluator for this study. The Center's role is to implement a Randomized Controlled Trial (RCT) to collect and analyze data using a rigorous, mixed-methods approach, which will allow CREP to provide formative and summative feedback to the SSEC and to answer the following main and supplemental evaluation questions (What Works Clearinghouse, 2022):

1. Does the intervention improve student achievement, particularly achievement of high needs students, in science, math, and reading to a statistically significant extent, relative to controls (main)?
2. Is adoption of NGSS or NGSS-like standards (i.e., in South Carolina, but not North Carolina) at the state level associated with a difference in the effect of the intervention on student outcomes (supplemental)?
3. To what extent does the PD meet (a) teachers' perceived needs in North Carolina vs. South Carolina, and (b) SSEC's stated goals (supplemental)?
4. To what extent are teachers who receive the PD implementing key program components with fidelity in the classroom? Does fidelity of implementation vary with the type of underlying state standards (NGSS-like vs. not NGSS-like) (supplemental)?
5. To what extent do teachers participating in the overall intervention feel it has been effective? What teacher needs still remain? Do teacher impressions of the intervention vary with the type of underlying state standards (NGSS-like vs. not NGSS-like) (supplemental)?

This evaluation is supported by technical advising from Abt Associates, and included establishment of an Evaluation Design Plan that underwent review by the EIR Evaluation technical assistance team to maximize the effectiveness of the program evaluation. In addition to answering the above research questions, CREP's randomized controlled trial has been designed so that the results of the student achievement analysis have the potential to meet the U.S. Department of Education's *What Works Clearinghouse* Standards without reservations.

The current document represents the initial Summative and main findings Report incorporating and summarizing available data across the four years of the study, which includes the first year of site recruitment (2019-20 school year) and three years of implementation (2020-21 through 2022-23 school years). The final summative report will include qualitative data collected during the fifth and final year of the grant (2023-24 school year).

Background

According to the National Science Foundation's Science and Engineering indicators (2022), STEM (Science, Technology, Engineering, and Mathematics) job opportunities in the United States have grown faster than the overall job market since 2010, and further STEM job openings are projected for future years. With this increasing need, it becomes even more critical for early engagement of students in STEM education, as fostering STEM awareness at a younger age can help maintain interest in STEM programs at both the high school and college level (Brown & Browning, 2021). As a result, inquiry-based science education (IBSE) has become increasingly important as researchers recognize it as an engaging approach to science learning that centers on students' personal interests and promotes active learning (van Uum et. al., 2016). IBSE is defined as a process where students employ critical thinking to explore and understand the natural world through asking questions, conducting investigations, interpreting data, constructing arguments, building models, and communicating results (Crawford, 2014). This definition emphasizes scientific practices in the learning process and necessitates that learners actively engage in understanding scientific concepts, processes, and the nature of science (Strat et al., 2023).

Early STEM education experiences such as inquiry-based learning have demonstrated positive outcomes on students' learning. For instance, Krajcik et al. (2023) investigated the effect of a Project-Based Learning science intervention on students' academic, social, and emotional outcomes. The study involved 2,371 third graders across 46 Michigan urban and rural public schools with a focus on recruiting schools having racial and ethnic minorities as well as students receiving free and reduced lunch. The study was evaluated through a cluster randomized control trial. Results indicated that students participating in the intervention had higher standardized science test scores (0.28 standard deviations (SD)) and reported higher levels of self-reflection and collaboration during science activities. Another study using a pretest-posttest control group experimental design on sixth-grade students in Türkiye found that inquiry-based learning substantially improved students critical-thinking skills (Duran & Dökme, 2016).

When considering the benefits of inquiry-based activities, it is also important to consider the development of teachers' skills and knowledge for effective implementation. Programs designed to enhance teachers' capabilities generally employ two main strategies: Teacher professional development (PD) and the introduction of new curriculum materials. PD aims to modify aspects of teachers' instructional techniques, while new curricula are designed to influence both teaching methods and the

content delivered. These programs can operate independently (e.g., primarily focusing on PD) or in conjunction, where the use of new curricular materials is supported through PD. Both strategies provide direct instructional guidance and aim to shape the daily interactions between teachers and students with structured lesson plans and teaching strategies (Ball & Cohen, 1996).

Meta-analysis studies on the effect of PD have shown generally positive impacts on student learning. For instance, Fletcher-Wood and Zuccollo (2020), in a meta-analysis of 53 randomized controlled trials, observed an average effect size of 0.09 on student learning, alongside potential benefits for student self-efficacy and teacher retention. Additionally, a study by Gonzalez et al. (2022) reviewed 37 experimental studies on preK-12 STEM PD and curriculum interventions, revealing statistically significant enhancements in teachers' content and pedagogical knowledge with an average impact of +0.56 SD. These findings collectively affirm the profound impact of professional development on educational outcomes.

Regarding effective PD practices, Kennedy (2016) analyzed 28 studies and found that strategies such as coaching and curriculum-based programs that assist teachers in gaining insights into their practices were more effective in predicting positive outcomes for students. Specifically in science, Slavin et al. (2014) found that PD programs that supported inquiry-based learning, but did not include kits, led to average improvements of 0.36 SD in students' outcome on science achievement measures. These findings align with the literature recommending best practices for PD, which emphasize (a) enhancing teacher content knowledge, (b) providing models such as lesson plans, (c) hands-on activities and instructional materials, and (d) engaging teachers in STEM practices (Lo, 2021).

Research shows that using only curriculum materials leads to modest improvements in student outcomes. In a systematic review of elementary science teaching approaches, Slavin and colleagues (2014) observed that inquiry-based programs with primary focus on professional development for teachers and science kits did not substantially enhance science achievement (effect size +0.02 from seven studies). Similarly, Cheung et al. (2017), in their review of science programs for grades 6-12, found that programs relying only on science kits and innovative textbooks had lower effects compared to other science programs. These findings underscore the limited impact of standalone curriculum materials without additional instructional support. Moreover, Lynch et al. (2019) conducted a meta-analysis of 95 pre-K–12 STEM PD and curriculum programs, contending that examining PD and curriculum studies *separately* poses conceptual and practical challenges, as most curriculum programs include PD for teachers; likewise, PD programs provide classroom materials. Their analysis revealed an average effect size of +0.21 SD, emphasizing the advantages of combining curriculum materials with consistent professional support to enhance teaching quality and student academic performance in STEM fields.

When exploring the benefits of IBSE and strategies promoting its use, it is crucial to consider the challenges faced in practicing this approach, particularly in rural and low-income areas. In general, these challenges include (a) teachers' instructional beliefs shaped by prior experiences, (b) a lack of science knowledge and skills, (c) insufficient instructional time, (d) curricula that do not align with state science standards, (e) difficulties in managing hands-on activities, and (f) inadequate administrative support in the classroom (Herb, 2022; Lee & Houseal, 2003; Yoon et al., 2012). These challenges are intensified in rural and low-income schools, where studies indicate K-6 teachers face issues such as (a) limited PD, (b) isolation, (c) teaching multiple subjects, (d) inadequate materials and curriculum, (e) low prioritization of science, and (f) low self-efficacy in teaching science (Zinger et al., 2020).

The COVID-19 pandemic exacerbated the challenges in practicing IBSE as students shifted to remote learning. Key issues included limited access to technological resources, as well as teachers' and students' unfamiliarity with online education. A study surveying elementary school teachers in the U.S. and Canada revealed that those in low-income schools felt ineffective, were unable to cover the expected curriculum fully, and reported their students were less prepared for remote education and experienced higher dropout rates compared to their counterparts in higher-income schools (Alvarez-Rivero et al., 2023). Additionally, adapting STEM education to an online environment was challenging, particularly in converting hands-on and small group activities to virtual formats (Moreno et al., 2021). These types of barriers and challenges brought by the pandemic were present in the implementation of *Smithsonian Science*, underscoring the need for further professional development to help teachers utilize STEM programs in the classroom. To address this need, *Smithsonian Science* provided teachers with direct instruction and modeling in both the content and techniques necessary for teaching inquiry-based STEM modules.

Methods

On October 8, 2019, the University of Memphis Institutional Review Board (IRB) conducted a review of CREP's evaluation plan, and determined that activities associated with the evaluation did not meet the Office of Human Subjects Research Protections definition of human subjects research (IRB ID: PRO-FY2020-179).

The elements of CREP's evaluation included (a) assessment of student achievement in both treatment and comparison schools for a single cohort of students who received the intervention for a total of three years (from third to fifth grade), (b) anonymous teacher PD evaluation surveys, (c) anonymous teacher module logs, (d) classroom observations by trained observers, and (e) teacher focus groups. See *Instruments*, below, for details on each of the elements implemented during the three past years.

Impact of COVID-19

The original evaluation plan for this study included baseline data collection for student achievement during early Fall 2020, at the beginning of the study cohort's third grade year. Prior to that, teachers were scheduled to receive PD in Summer 2020, with planned implementation of the first *Smithsonian Science for the Classroom* module in treatment schools initiated in Fall 2020. With implementation, classroom observations and other data collection activities would begin. However, on March 11, 2020, the World Health Organization declared COVID-19, the disease caused by the novel coronavirus SARS-CoV-2, a global pandemic (World Health Organization, 2020).

With states of emergency declared beginning in March 2020 by federal, state, and city leaders, schools in North and South Carolina closed for the remainder of the 2019-20 school year, and staff at the SSEC and CREP were instructed to restrict program implementation and research activities to telework. Schools in North and South Carolina instructed students through distance learning for most of the 2020-21 academic year (the first year of implementation). This complicated recruitment of school districts and significantly altered both the program implementation and evaluation timeline during Project Years 1 and 2. Principally, these changes were: (a) extension of school recruitment into winter 2020, with randomization occurring in January 2021; (b) redesign of SSEC's first and second teacher PD

workshops for delivery to teachers in a virtual setting (i.e., over Zoom), with the first PD session delayed until late spring 2021; (c) shift from paper-and-pencil to online assessment, so that teachers could assess students still learning from home during the delayed baseline assessment period of late spring 2021; (d) only pilot implementation of the engineering curriculum in late spring 2021, with minimal Module Log data collection occurring; and (e) no classroom observations during the 2020-21 school year. For the 2021-22 school year (the second year of implementation), the major impacts from COVID included (a) no classroom observations could be conducted in Burke County Schools in North Carolina as they declined to allow visitors due to COVID protocols, and (b) summer 2021 teacher PD was once again conducted over Zoom instead of in person.

For the 2022-23 school year, project implementation largely occurred as planned, and in-person activities resumed. Teachers received hybrid professional development on the physical science curriculum during summer 2022, where the teachers attended in person and the trainers provided instruction remotely. Teachers then implemented the physical science and engineering units throughout the 2022-23 school year. Observations resumed in person with all districts. Only two schools in Orangeburg County were unable to be observed. Teachers provided module log data as planned, and in summer 2023, professional development was conducted in-person as planned. Modifications to project implementation and evaluation are summarized in **Table 1**.

Table 1: Modifications to Project Implementation and Evaluation as a result of COVID-19

	2019-20 (PY1)			2020-21 (PY2)			2021-22 (PY3)	2022-23 (PY4)
	Fall	Spring	Summer	Fall	Spring	Summer	Summer	Summer
Y1-Y4 Implementation Timeline:								
Original	<---PD development--->	Recruitment completed	PD 1	<---Teacher implementation--->		PD 2	PD 3	PD 4
Modified	<---PD development--->		<--PD revision for online delivery-->	Recruitment completed	PD 1 (Virtual) Pilot Implementation	PD 2 (Virtual)	PD 3 (Hybrid)	PD 4 (In-Person)
Y1-Y4 Evaluation Timeline:								
Original		Random assignment	PD 1 Evaluation	Baseline assessment Observations Module Logs		PD 2 Evaluation	PD 3 Evaluation	PD 4 Evaluation
Modified			<--Baseline assessment revision for online delivery-->		Random assignment Baseline assessment PD 1 evaluation Module Logs	PD 2 Evaluation	PD 3 Evaluation	PD 4 Evaluation

Participants

Schools within seven school districts in North and South Carolina were recruited by the Smithsonian Science Education Center in collaboration with regional partners: NC SMT in North Carolina, and SCCMS in South Carolina. In accordance with the standards for a Randomized Controlled Trial (RCT), schools within districts agreeing to participate did not know at recruitment whether they would be assigned as a treatment or comparison school. Treatment schools in this project received four professional development (PD) workshops and materials support beginning in spring 2021 (with PD

originally scheduled to begin in summer 2020); comparison schools served as controls through the 2022-23 school year, and received one PD workshop in summer 2024 and one curricular module to implement during 2024-25.

School Randomization

In January 2021, after receiving student rosters from all participating schools, CREP randomly assigned schools to the treatment or comparison group. Prior to randomization, three schools from one district in North Carolina (Alexander County) were dropped because they had mixed grade classrooms (i.e., all grades 3-5 are in one class), and therefore students in the third grade cohort followed for the study would not receive the full treatment (i.e., three consecutive years of grade-specific instruction). In addition, two other schools in North Carolina were dropped because they served special populations: In Burke County, one Special Education school, and in Caldwell County, one Alternative Education school were dropped. The number of available schools randomly assigned was limited by the number of students and teachers for which the grant budget would allow materials support. Ten randomization blocks were created across the two states: One block for each of the seven districts, and an additional performance block within three of the larger districts that contained a greater number of schools with more varied student achievement. In other words, the four smaller districts (with more uniform student achievement) were each a block, and the three larger districts each had two blocks with schools grouped by higher or lower performance based on publicly available third grade state assessment data from the 2018-19 school year.

Two sets of random numbers were then assigned to schools: A random number for the entire block, and a separate random number for each school within every block. The list of schools was then sorted by block random number (low to high), then by school random number (low to high) within each block. Where available, the first two pairs of schools (n = 4) in each of the 10 blocks were selected, for a total of 18 randomization pairs (36 schools). Each school in each pair was then randomly assigned by flipping a coin, with the first school on the randomly sorted pair list assigned to the treatment group for heads or the comparison group for tails. The study cohort, followed over three school years, was composed of more than 1,600 third grade students in these schools.

To gain access to state student rosters and standardized test data for participating schools, CREP established data agreements with the Departments of Education in North Carolina and South Carolina. CREP also asked each participating treatment and comparison school to identify a Stanford-10 assessment coordinator to serve as a point-of-contact and to organize administration of the baseline assessment in spring 2021 and the final assessment in spring 2023. Additionally, the SSEC asked treatment schools to designate a school site coordinator, who would manage the logistics for receiving and distributing *Smithsonian Science* modules, organize school observations, and provide teacher lists for the professional development.

School Demographics

Demographics across all treatment and comparison schools in each participating district at the time of randomization are summarized in **Table 2.** Demographics of individual treatment and comparison schools at the time of randomization are presented in **Table 3**.

Table 2: Treatment and Comparison School Demographics by State and District

State	District	Schools	Total Students	% FRL[1]	% Rural[2]
NC	Alexander	4	1,293	41.4%	100%
NC	Burke	8	2,972	64.2%	25%
NC	Caldwell	8	2,934	59.8%	50%
NC	Polk	4	1,017	88.9%	75%
SC	Marion	2	899	100%	100%
SC	Marlboro	5[3]	2,253	100%	60%
SC	Orangeburg	6	1,922	100%	83%

[1]Percentage of students who qualify for free and reduced lunch according to the Common Core of Data (CCD), 2019-2020. [2]Percentage of project schools classified as Rural by the CCD. [3]One cohort of students in this district transferred from Bennettsville Primary to Bennettsville Intermediate partway through the project. For most purposes of this study, including randomization, these two schools are treated as a single school.

Table 3: Demographics of Treatment and Comparison Schools

State	District	School	Total Students	% FRL[1]	Urbanicity[2]	Status[3]
NC	Alexander	Ellendale Elementary	285	40.3%	42-Rural: Distant	T
NC	Alexander	Stony Point Elementary	266	57.5%	41-Rural: Fringe	T
NC	Alexander	Bethlehem Elementary	461	31.0%	41-Rural: Fringe	C
NC	Alexander	Wittenburg Elementary	281	44.1%	41-Rural: Fringe	C
NC	Burke	Forest Hill Elementary	305	77.4%	13-City: Small	T
NC	Burke	George Hildebrand Elementary	327	64.2%	42-Rural: Distant	T
NC	Burke	Icard Elementary	304	63.8%	22-Suburb: Mid-size	T
NC	Burke	Mull Elementary	302	61.2%	41-Rural: Fringe	T
NC	Burke	Glen Alpine Elementary	379	69.0%	22-Suburb: Mid-size	C
NC	Burke	Hildebran Elementary	379	62.5%	22-Suburb: Mid-size	C
NC	Burke	Valdese Elementary	580	57.9%	22-Suburb: Mid-size	C
NC	Burke	W A Young Elementary	396	62.6%	22-Suburb: Mid-size	C
NC	Caldwell	Baton Elementary	378	58.2%	41-Rural: Fringe	T
NC	Caldwell	Collettsville School	352	61.1%	41-Rural: Fringe	T
NC	Caldwell	Hudson Elementary	702	61.2%	22-Suburb: Mid-size	T
NC	Caldwell	Sawmills Elementary	340	57.6%	22-Suburb: Mid-size	T
NC	Caldwell	Dudley Shoals Elementary	463	60.7%	41-Rural: Fringe	C
NC	Caldwell	Kings Creek Elementary	172	66.3%	42-Rural: Distant	C
NC	Caldwell	Lower Creek Elementary	391	42.4%	13-City: Small	C
NC	Caldwell	West Lenoir Elementary[4]	136	97.8%	13-City: Small	C
NC	Polk	Polk Central Elementary	347	98.5%	42-Rural: Distant	T
NC	Polk	Tryon Elementary	416	99.3%	31-Town: Fringe	T
NC	Polk	Saluda Elementary	163	98.1%	41-Rural: Fringe	C
NC	Polk	Sunny View Elementary	127	97.6%	42-Rural: Distant	C
SC	Marion	Marion Intermediate	579	100%	41-Rural: Fringe	C
SC	Marion	Mccormick Elementary	320	100%	41-Rural: Fringe	T
SC	Marlboro	Bennettsville Primary	439	100%	32-Town: Distant	T
SC	Marlboro	Bennettsville Intermediate	441	100%	32-Town: Distant	T
SC	Marlboro	McColl Elementary/Middle	754	100%	42-Rural: Distant	T
SC	Marlboro	Clio Elementary	145	100%	42-Rural: Distant	C
SC	Marlboro	Wallace Elementary/Middle	474	100%	41-Rural: Fringe	C
SC	Orangeburg	Holly Hill Elementary	430	100%	42-Rural: Distant	T
SC	Orangeburg	St. James-Gaillard Elementary	345	100%	42-Rural: Distant	T
SC	Orangeburg	Mellichamp Elementary	305	100%	32-Town: Distant	T
SC	Orangeburg	Bethune-Bowman Elementary	318	100%	42-Rural: Distant	C
SC	Orangeburg	Lockett Elementary	281	100%	43-Rural: Remote	C
SC	Orangeburg	Vance-Providence Elementary	243	100%	43-Rural: Remote	C

[1] Percentage of students who qualify for free and reduced lunch according to the Common Core of Data (CCD), 2019-2020

[2] School Urban-Centric Locale code according to the Common Core of Data (CCD), 2019-2020

[3] School membership in the Treatment (T) or Comparison (C) group

[4] West Lenoir Elementary closed and combined with Valmead Elementary at the end of SY21-22.

Professional Development

In March 2021, the SSEC offered three days (12 hours) of synchronous virtual PD to teachers in treatment schools, which introduced participants to a grade-specific **Engineering** curricular module to be piloted during the remainder of the 2020-21 school year. Teachers who were unable to attend or wished to review the PD also had the opportunity to view six recordings (two per day) of the online PD sessions. Staff from the SSEC recorded attendance by tracking the number of hours each teacher was logged into the PD session, and by recording the number of videos each teacher clicked to view.

During August 2021, the SSEC offered two days (12 hours) of professional development (PD) to teachers in treatment schools, which provided a content-focused look at the **Engineering** modules. The PD took place over Zoom, with teacher attendance determined by (a) the number of hours each teacher spent logged into the PD, and (b) recording the number of videos each teacher clicked to view. Teachers who were unable to attend or who wished to review the PD had the opportunity to view four recordings (two per day) of the online PD sessions.

During Summer 2022, the SSEC offered two days (11 hours) of PD to teachers in treatment schools, which introduced participants to a grade-specific **Physical Science** curricular module they would receive during the 2022-23 school year. The trainers delivered instructional content over Zoom, while teachers were in-person at training sites so they could experience the hands-on materials. While previous PDs took place over Zoom due to COVID-19 restrictions, trainers were on Zoom for this PD because each needed to cover several training sites simultaneously. Staff from the SSEC recorded attendance using teacher sign-in sheets. Teachers who were unable to attend or who wished to review the PD had the opportunity to view four recordings (two per day) of the PD sessions or attend the PD at another location on alternate dates.

During Summer 2023, the SSEC offered two days (10 hours) of PD to teachers in treatment schools, which provided a content-focused look at the **Physical Science** modules. The PD took place in person at the training sites with each state, as originally planned in project implementation.

Design Summary and Fidelity Tracking

During the 2020-21 project year (Project Year 1), CREP completed an Evaluation Design Plan for the U.S. Department of Education, Education Innovation and Research (EIR) program, to specify in advance the evaluation design, research questions, and data collection and analysis strategy. This plan was reviewed by Abt Associates, CREP's technical advisor for the evaluation, and was submitted in December 2021 to the Registry of Efficacy and Effectiveness Studies (REES), a database of causal inference studies in education and other related fields.

As part of the Design Summary, CREP developed standards for estimating fidelity of implementation for two relevant aspects of the SSEC's intervention: Professional development and curricular support (i.e., access to materials, in this case the *Smithsonian Science for the Classroom* modules). Data to evaluate fidelity of implementation based on the rubrics is provided by the SSEC (for PD attendance) and Carolina Biological Supply Company (for curricular support). The key components of fidelity and measurable indicators within each key component are specified by implementation year in **Table 31**, **Table 32**, and **Table 33** in **Appendix A**.

High fidelity for **professional development** for all three years was defined as having (a) at least 75% of teachers (b) attend at least 80% of the available hours of PD (c) in at least 75% of treatment schools. High fidelity of implementation for **curricular support** in all three years was based on shipment of needed curricular materials (i.e., modules) by an appropriate target date and in sufficient amount to allow participating treatment schools to serve all students in grades 3-5. Therefore, based on the established fidelity standards, professional development and curricular support were provided to schools with **high fidelity for both components in Year 1 and Year 3**, while **neither component was implemented with fidelity in Year 2**.

However, readers should note that in **Year 2** (the 2021-22 school year), instruction in schools was severely disrupted due to the COVID-19 pandemic. Module deliveries for treatment schools were originally planned for *fall 2020 and 2022*, following summer professional development. But due to delays in recruitment, randomization, and subsequently professional development because of the pandemic, the Engineering curriculum instead shipped to schools *between late February and March 2021*, to begin implementation that spring. As a result of high mobility of students and teachers during the COVID-19 pandemic, site coordinators were asked at the start of the 2021-22 school year to update the SSEC with their latest student numbers and any additional curriculum needs resulting from fluctuation in enrollment and staffing. One of the largest schools in the study, Hudson Elementary in Caldwell County, was the only school to request two additional modules to meet their needs. While the order was placed on August 30, 2021, they were not delivered until late October and early November due to supply chain-related delays brought on by the pandemic, thus resulting in missing the October 1 target date for shipping.

Table 4: Fidelity of Implementation Summary

Year	PD Session Content	PD Type	Professional Development	Curricular Support
			75% of teachers attended 80% of total available hours of PD in 75% of schools	90% of ordered modules shipped by target date
Year 1 (2020-21)	SP 21: Introductory Engineering (Curriculum-Focused)	Virtual	Yes	Yes
Year 2 (2021-22)	SU 21: Intermediate Engineering (Content-Focused)	Virtual	No	No[1]
Year 3 (2022-23)	SU 22: Introductory Physical Science (Curriculum-Focused)	Hybrid	Yes	Yes

[1] Due to delays in shipping materials because of the COVID-19 pandemic

Instruments

The Abbreviated Stanford-10

The main (i.e., confirmatory) findings for this evaluation are student achievement in science, math, and reading for all students in a single cohort participating in three years of curricular implementation from third through fifth grade, and whom CREP followed longitudinally. To maximize the opportunity to conduct a study that meets What Works Clearinghouse (WWC) standards without reservations and receive the highest possible rating, CREP administered a baseline assessment of student achievement in math and reading in spring 2021, near the end of the cohort's third grade year. The instrument used for baseline assessment was the Abbreviated Battery of the Stanford Achievement Test Series, Tenth Edition (Pearson Education, 2018).

To minimize loss of learning time for students during an already COVID-disrupted school year, assessment was reduced to the greatest extent possible. Third grade students in both treatment and comparison schools completed the (a) Mathematics: Problem Solving, (b) Mathematics: Procedures, and (c) Reading Comprehension subscales of the Stanford-10 during March and April 2021. According to the WWC Primary Science Review Protocol Version 4.0 (WWC, 2019), at the elementary level, reading comprehension and general mathematics achievement may be used as a proxy for science achievement, allowing fewer Stanford-10 subtests to be administered as the baseline. As many students were still attending school remotely due to the COVID-19 pandemic at the time of testing, CREP provided instructions, materials, and technical support to allow teachers to administer the Stanford-10 online using Pearson's testing system. Students whose parents declined to provide passive parental consent (i.e., completed and submitted a Denial of Consent form to CREP) were not assessed, and if assessed accidentally, scores for these students were not used.

Third grade students in 35 of the 36 treatment and comparison schools completed the designated subtests on the Abbreviated Stanford-10 during March and April 2021. One school, Valdese Elementary (Burke County Schools, North Carolina), reported insurmountable technical difficulties and was permitted to halt attempts at assessment to resume student learning. As a result, this school had baseline data for Reading Comprehension, but not Mathematics. Some third grade classes in certain schools were also still Virtual Academy classes with limited student contact time. As a result, many Virtual Academy teachers were unable to complete baseline assessment. In addition, several school districts with a small number of students enrolled in virtual learning had assigned all virtual students in the district to a single teacher, and not all these teachers were associated with a school that was part of the study. In these cases, CREP identified the teacher where possible, and requested they administer the pre-assessment to students from participating schools.

In total, over 1,600 students completed at least one subtest of the Abbreviated Stanford-10. This included 1,522 students who completed the Mathematics: Problem Solving subtest, 1,495 students who completed the Mathematics: Procedures subtest, and 1,519 students who completed the Reading Comprehension subtest. Because some teachers administered subtests on different days and/or did not have students in class five days a week, not all students completed all subtests. Students missing one or more subtests may have been absent, or may have encountered technical difficulties with access to online assessment materials.

Table 5 displays the testing location (in person or remote) of students who completed at least one section of the Stanford-10 baseline assessment. Students who completed in-person assessment took the Stanford-10 at school under the supervision of a teacher proctor. Students who completed remote assessment took the Stanford-10 at home, proctored by a teacher, with another adult requested to be in the room with the student to keep them on-task. Assessment location data were not available for all classrooms, so the total number of students with location data (Total Reported) is smaller than the total number of students assessed. All students were assessed online regardless of assessment location.

Table 5: Stanford-10 Pre-Assessment Location by District

State	District	Known In Person[1]	Known Remote[2]	Total Reported[3]	Total Unreported[4]	Total Assessed[5]
NC	Alexander	120	1	121	11	132
NC	Burke	144	7	151	149	300
NC	Caldwell	72	0	72	7	79
NC	Polk	63	2	65	76	141
SC	Marion	50	30	80	145	225
SC	Marlboro	137	26	163	52	215
SC	Orangeburg	64	99	163	44	207

[1] All students who took the assessment in the classroom
[2] All students who took the assessment while not in the classroom environment
[3] Number of students whose teacher reported their testing location
[4] Number of students whose teacher did not report their testing location
[5] All assessed students reported by teachers, regardless of testing location data

In Spring 2023, fifth grade students in all 36 participating *Smithsonian Science* schools completed the online Stanford-10 Science assessment for the posttest. The Stanford-10, along with the state assessment data in reading and math, determine whether students in treatment and comparison schools demonstrate similar achievement levels at the end of the project. The Stanford-10 is not a timed assessment. The recommended time for testing is 25 minutes. All usual accommodations for students with IEPs and/or who have limited English proficiency were permitted. Each district selected a testing window within a timeframe from March 20, 2023 to April 17, 2023 for all students to complete the assessment (see **Table 6**). Students could complete the subtest any time within the district's testing window.

Table 6: Stanford-10 Post-Assessment Testing Windows by District

North Carolina	
Alexander	March 20-24, 2023
Burke	March 27-31, 2023
Caldwell	March 20-24, 2023
Polk	April 3-7, 2023
South Carolina	
Marion	March 27-31, 2023
Marlboro	April 17-21, 2023
Orangeburg	March 20-24, 2023

To help ensure as many fifth grade students as possible were tested, CREP collected student rosters from each school in fall/winter 2022. Doing so also allowed CREP to confirm parental consent and set up student IDs in the Pearson system, as only students with an assigned ID can complete the Stanford-10. CREP then returned the updated rosters with student IDs to schools prior to their testing window. Test administrators could register for one of two online training sessions held over Zoom on March 2, 2023 (*N* = 40) and March 7, 2023 (*N* = 32) to cover test administration procedures. After completing testing, schools were asked to submit a post-administration online checkout to answer a few short questions about assessment conditions and report any disruptions experienced.

State Assessment Data

CREP obtained spring 2023 posttest state assessment data for both North Carolina (reading, math, science) and South Carolina (reading and math) through data agreements with the respective state Department of Education (DOE). CREP researchers requested the same variables (e.g., student demographics) from both state DOE offices, helping ensure the data are as similar as possible, and recoded any of the original variables to ensure consistency in coding between the states. The main finding student outcome measures are listed in **Table 7**.

Table 7: Main Finding Student Outcome Measures

Outcome Measure	Outcome Domain	Unit of Measurement	Baseline Measure	Time Periods Represented	Other Covariates
SAT-10 Science subtest	General Science Achievement (Primary Science, version 4.0)	Student	SAT-10 Reading Comprehension subtest	spring 2023	**Student Level**: FRL, gender, EL, IEP; **School Level**: State, district, and other blocking variables (TBD)
North Carolina End-of-Grade (EOG) Math	General Mathematics Achievement (Primary Mathematics, version 4.0)	Student	SAT-10 Total math subtest (Problem Solving and Procedures)	spring 2023	**Student Level**: FRL, gender, EL, IEP; **School Level**: State, district, and other blocking variables (TBD)
South Carolina College-and Career-Ready Assessments (SC READY) Math	General Mathematics Achievement (Primary Mathematics, version 4.0)	Student	SAT-10 Total math subtest (Problem Solving and Procedures)	spring 2023	**Student Level**: FRL, gender, EL, IEP; **School Level**: State, district, and other blocking variables (TBD)
North Carolina End-of-Grade (EOG) Reading	Comprehension (Adolescent Literacy, version 4.0)	Student	SAT-10 Reading Comprehension subtest	spring 2023	**Student Level**: FRL, gender, EL, IEP; **School Level**: State, district, and other blocking variables (TBD)

Outcome Measure	Outcome Domain	Unit of Measurement	Baseline Measure	Time Periods Represented	Other Covariates
South Carolina College-and Career-Ready Assessments (SC READY) English Language Arts (ELA)	Comprehension (Adolescent Literacy, version 4.0)	Student	SAT-10 Reading Comprehension subtest	spring 2023	**Student Level**: FRL, gender, EL, IEP; **School Level**: State, district, and other blocking variables (TBD)

PD Evaluation Surveys

Teachers attending Intermediate Physical Science PD in summer 2023 were asked to complete an anonymous follow-up survey providing feedback about aspects of the PD. During each year's PD, a link to the online survey was distributed to teachers by SSEC during the last day of training. Additionally, paper copies were made available at the end of the Summer 2022 PD, which had a hybrid delivery format. Open-ended comments were reviewed and cleaned of identifying data before reporting.

Teacher Module Logs

Teachers were also asked to complete an anonymous Module Log after they finished implementing each *Smithsonian Science* module during the 2022-23 school year. Each year, CREP sent a link to the Module Log survey to the Site Coordinators at each school and asked them to distribute it to their teachers. This short online survey asked teachers for feedback on various aspects of the modules, such as the extent to which they aligned with state science education standards. It also asked teachers for details regarding their implementation of module content, even if they were only able to implement a portion of the module.

Classroom Observations

School Observation Measure (SOM).

The *School Observation Measure (SOM)* was developed to determine the extent to which different common and alternative teaching practices are used throughout an entire school (Ross, Smith, & Alberg, 1998). For **targeted observations**, one teacher is observed for 45-60 minutes. The observer examines classroom events and activities descriptively, not judgmentally. Notes are taken relative to the use or nonuse of 24 target strategies, and the frequency recorded via a 5-point rubric that ranges from (0) Not Observed to (4) Extensively. Two global items are used to rate, respectively, (a) the level of academically focused instructional time and (b) degree of student attention and interest. The notes forms are completed every 15 minutes of the lesson, then summarized on a SOM Data Summary Form.

The SOM strategies include (a) traditional practices (e.g., direct instruction and independent seatwork) and (b) alternative, predominantly student-centered methods associated with educational reforms (e.g., cooperative learning, project-based learning, inquiry, discussion, using technology as a learning tool). The strategies were identified through surveys and discussions involving policy makers, researchers, administrators, and teachers, as those most useful in providing indicators of schools'

instructional philosophies and implementations of commonly used reform designs (Ross, Smith, Alberg, & Lowther, 2004).

In a reliability study (Lewis, Ross, & Alberg, 1999), pairs of trained observers selected the identical overall response on the five-category rubric on 67% of the items, and were within one category on 95% of the items. Further results establishing the reliability and validity of *SOM©* are provided in the Lewis et al. (1999) report. In a reliability study using Generalizability Theory, Sterbinsky and Ross (2003) found reliability at the .74 level for five SOMs conducted at a school. Reliability increased to .82 with eight SOMs and to .85 with 10 SOMs conducted at a school.

To ensure the reliability of data, observers received (a) training over Zoom, (b) a manual providing definitions of terms, (c) examples and explanations of the strategies, and (d) a description of procedures for completing the instrument. After receiving the manual and instruction in a group session, each observer participated in several practice exercises. Inter-rater reliability data were later collected and used to compare each observer's ratings with ratings from experienced observers. The trained observers met acceptable thresholds for inter-rater reliability in this study.

Rubric for Inquiry-Based Assessment (RIBA).

The Rubric for Inquiry-Based Assessment (RIBA) is designed to record evidence of inquiry-based science activities in the classroom, as well as to rate the overall level of class time dedicated to inquiry-based science. With **targeted** RIBA observations, one rubric is completed for the entire science class period for each classroom observed. The RIBA has 10 Student-Centered Activity items rated as either "Not Observed" or "Observed", in addition to one summary item ("Level of class time dedicated to inquiry-based science") rated as either "Low", "Moderate", or "High". In this study, the RIBA was used as an addendum with each targeted SOM observation.

Teacher and Site Coordinator Focus Groups

CREP developed a focus group protocol with a series of questions for teacher and site coordinator focus groups, which were conducted in spring and summer 2023. CREP asked for teacher and site coordinator volunteers across all schools to participate, and 14 treatment and control teachers and 15 site coordinators participated. Those who participated represented all districts in the study except (a) Polk in North Carolina (teachers) and (b) Marion in South Carolina (teachers and site coordinators). CREP held virtual focus groups over Zoom for control points of contact and teachers in spring 2023. Multiple times were offered during the school day to support participation. CREP held in-person focus groups with treatment site coordinators and teachers in July and August 2023 during on-site professional development. All focus groups lasted between 30 minutes and 1 hour.

Results

Abbreviated Stanford-10 Combined States Analyses

The **main findings** (i.e., confirmatory impact analyses) are the effects of *Smithsonian Science* on student achievement at the end of **fifth grade** for the 2020-21 cohort of third grade students in treatment vs. control schools after three years of implementation (i.e., in 2022-23, or Grant Year 4) on

(a) two state assessments (math and reading) and (b) the Science subtest of the Abbreviated Battery of the **Stanford Achievement Test Series, Tenth Edition®** (SAT-10; Pearson Education, 2018) for the **overall sample** (all students combined across both states) (**Cell I of Table 8**). The state assessment in North Carolina is the End-of-Grade test (reading, math, and science), while in South Carolina the state assessment is SC Ready for reading and math and SCPASS for science.

Table 8: Achievement Outcomes for the Study Cohort

State	2019-20	2020-21			2021-22		2022-23		
	Grant Year 1	Grant Year 2			Grant Year 3		Grant Year 4		
		Implementation Year 1			Implementation Year 2		Implementation Year 3		
	A	B	C	D	E	F	G	H	I
	Grade 2	Grade 3 (Winter: Baseline)	Grade 3 (Spring: Exploratory)	Grade 3 (Spring: Exploratory)	Grade 4 (Spring: Exploratory)	Grade 4 (Spring: Exploratory)	Grade 5 (Spring: Exploratory)	Grade 5 (Spring: Exploratory)	Grade 5 (Spring: Confirmatory)
NC	N/A	Stanford-10 (total math & reading comp. as proxy for science)	Standardized state assessment scores (reading & math): Combined (NC & SC) full sample (all students) and subgroups	End-of-Grade (reading & math): NC full sample and subgroups	Standardized state assessment scores (reading & math): Combined full sample and subgroups	End-of-Grade (reading & math): NC full sample and subgroups	End-of-Grade (reading, math, science) & Stanford-10 (science): NC full sample and subgroups	Standardized state assessment scores (reading & math) & Stanford-10 (science): Combined sample subgroups	Standardized state assessment scores (reading & math) & Stanford-10 (science): Combined full sample
SC				SC-READY (reading & math): SC full sample and subgroups		SC READY (reading & math) SCPASS (science): SC full sample and subgroups	SC READY (reading & math) & Stanford-10 (science): SC full sample and subgroups		

Stanford-10 Combined States Sample

A total of 1,952 fifth grade students were on the enrollment rosters collected from schools in fall 2022. Of those, 1,754 students (90%) were tested. Two of those tested students were extreme outliers that each had as much influence on the statistical model as an average school. The removal of those two records resulted in the final analysis sample of 1,752 students. Individual-level demographics for the analysis sample are given in **Table 9** below.

Table 9: Stanford-10 Combined States Analysis Sample Demographics (N = 1,752)

Variable	Control	Treatment
Female	49.6%	51.4%
Black, Indigenous, and People of Color (BIPOC)	40.7%	43.8%
Economically Disadvantaged	65.4%	62.9%
English Learner (EL)	3.2%	7.0%
Special Education/IEP	13.2%	11.9%
Sample Size	**838**	**914**

Stanford-10 Combined States Attrition

All schools remained in the project for all three years of implementation (i.e., school-level attrition was 0%). At the individual (student) level, attrition was calculated according to the WWC Handbook's (2022) reference sample 3 (i.e., individuals present in clusters (schools) at follow-up). Overall sample attrition was 10.2%, and differential attrition (i.e., the difference in attrition rates for the treatment vs. comparison group) was 2.1%, which meets the WWC low attrition standards.

Stanford-10 Combined States Outcomes

The spring 2023 fifth grade Stanford-10 Science assessment was the posttest for the analysis, with the spring 2021 third grade Stanford-10 Reading Comprehension and Mathematics Problem Solving subtests as the pretests. The treatment and control groups met WWC (2022) baseline equivalence standards for both Reading Comprehension (g = 0.12) and Mathematics Problem Solving (g = -0.20): 0.05 < |Effect size at baseline| ≤ 0.25. Stanford-10 outcome data were analyzed with a hierarchical linear model (HLM), using school as a random effect. (Technical details of the analysis may be found in **Appendix B**.

When the analysis was originally completed in late 2023, Common Core of Data (CCD) data, which contains important school-level covariates such as the student-to-teacher ratio, were not yet available for the 2022-2023 school year. School data from 2021-2022 were therefore used instead. The Stanford-10 analysis was updated in April 2024 to use the newly-released 2022-2023 school year CCD data (that is, CCD data from the year in which the posttest occurred). For that reason, the results below may be **slightly different** than in previous reports.

As shown in **Table 10**, the effect of the treatment was **positive** (g = 0.18) and **statistically significant** (p = .032). The HLM coefficient shows the treatment group scored 5.6 points **higher** than the control group. The percentile rank in **Table 10**, based on the effect size (g = 0.18), indicates that a student who scored at the median (50th percentile) in the treatment group would rank at the **57th percentile** in the comparison group (or better than 57% of students in the comparison group). This is reflected in the improvement index, which demonstrates that participation in *Smithsonian Science* improved outcomes in the treatment group by **7 percentile points**. As an indicator of the impact or "practical significance" of the treatment, the effect size (calculated as Hedges' g) is a descriptive statistic that indicates the magnitude of the difference (in standard deviation units) between the treatment and control groups. To aid interpretation of p-values, CREP calculated the minimum detectable effect size (MDES) for each analysis, using the lower bound of the 95% confidence interval for the treatment estimate. The MDES (.083) shows the smallest effect size that would be required for a statistically significant result, i.e., the effect size for which p = 0.05.

Table 10: Stanford-10 Spring 2023 Combined States Results

HLM Coefficient	95% CI	*p* value	Hedges' *g*	Percentile Rank	Imp. Index	MDES
5.59	2.52, 9.21	.032*	0.18	57	7	.083

* p < .05

Table 11, below, shows estimated marginal means (EMMs) for the treatment and control groups. EMMs are average scores on the Stanford-10 Science that have been adjusted to control for differences between the treatment and control groups on other variables (e.g., pretest score, free lunch status).

Table 11: Stanford-10 Spring 2023 Combined States Science SDs and Estimated Marginal Means by Group

Group	*N*	EMM	SD
Treatment	914	640.02	30.8
Control	838	634.44	30.6

As shown in **Figure 1,** in addition to being statistically significant, the magnitude of the difference (g = 0.18) is larger than most effect sizes reported in the research literature for similar interventions. While it was equivalent to the average effect found in a meta-analysis by Lynch et al. (2019) of effect sizes from approximately 35 studies in the research literature on STEM professional development and curriculum programs for science outcomes, they also found that effects were **larger** for studies of programs using many of the components used by the SSEC, including:

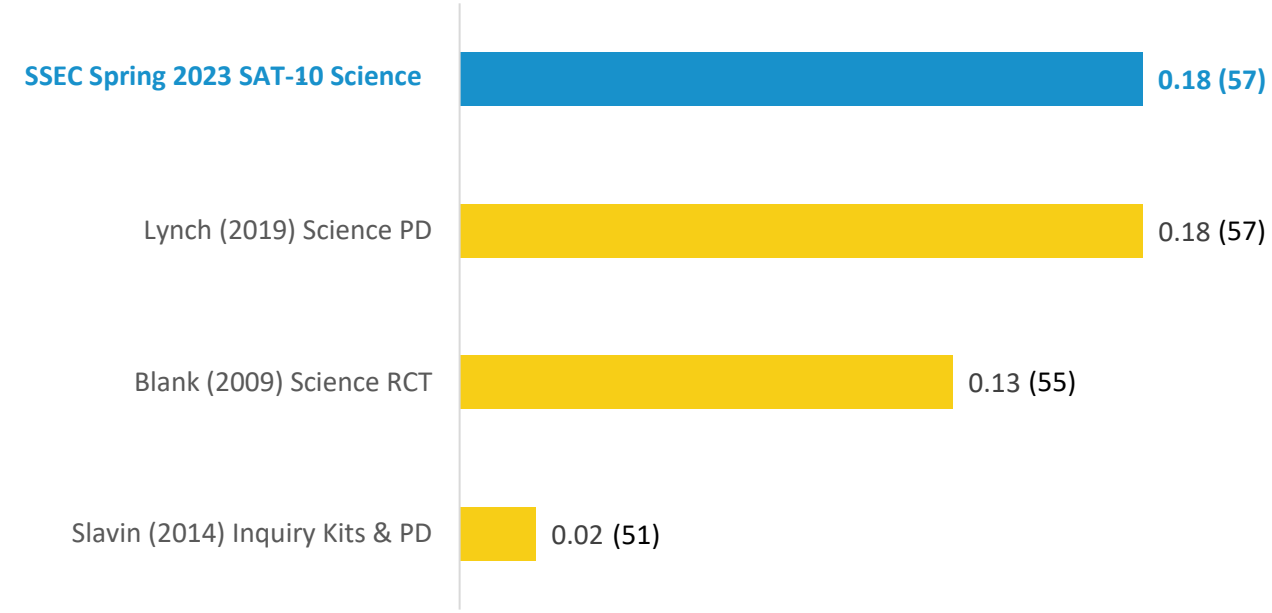
- Incorporating **both professional development** and **new curriculum materials** (vs. just one or the other)
- Focusing on **how to use curriculum materials** and improving teachers' content and pedagogical content **knowledge** and/or how students **learned** the content
- Having teachers participate alongside **other teachers in their school**, including meeting with their fellow teachers to discuss enacting the intended practices
- A **summer workshop**

In addition, Lynch et al. (2019) found that on average, programs providing any component of the professional development **online** had **worse** outcomes. Therefore, readers should keep in mind the impacts of COVID-19 on the delay in implementation (i.e., less student exposure to the modules) and the format for teacher professional development (virtual/hybrid vs. in person) when interpreting student achievement outcomes (i.e., impacts may have been larger without the impacts of COVID-19 on implementation).

Another systematic review (Slavin et al., 2014) of 23 randomized or matched control group studies on the achievement outcomes of all types of approaches to teaching science in elementary schools found that among studies evaluating **inquiry-based** teaching approaches, programs that **used science kits and accompanying professional development** did not show positive outcomes on science achievement measures (weighted ES = +0.02 in 7 studies, or essentially zero effect). In terms of research design, Blank & de las Alas (2009), in their meta review of 16 studies, found an average effect size of g = 0.13 for the four science teacher professional development studies in their review using an RCT design.

The effect size of fifth grade itself (the full academic year of study) based on average annual gains in effect size from nationally normed tests in science is approximately g = 0.40 (Bloom et. al, 2008). Therefore, the impacts of *Smithsonian Science* are not only positive, but are positive relative to findings in the research literature.

Figure 1: Spring 2023 Stanford-10 Combined States Science Effect Size (Percentile) Compared to the Research Literature



Spring 2023 Combined North and South Carolina State Assessment Analyses

A total of 1,952 fifth grade students were on the enrollment rosters collected from 36 participating schools in North Carolina and South Carolina. Of those, 984 took the North Carolina End-of-Grade (EOG) tests in Reading and Math, and 727 took the South Carolina College-and Career-Ready Assessments (SC READY) in English Language Arts (ELA) and Math. Thus, 1,711 students (87.6%) were tested in both Reading and Math with valid scores. Eleven of those tested students were extreme outliers (i.e., had scores that differed considerably from other students) such that each had as much influence on the statistical model as an average school. The removal of those eleven records resulted in the final analysis sample of 1,700 students. 937 students were in the treatment group, and 763 were in the control group. Individual-level demographics for the analysis samples are given in **Table 12**.

Table 12. Analysis Sample Demographics for North and South Carolina Students

Variable	Treatment		Control		Total	
	N	%	*N*	%	*N*	%
Female	441	47.1%	347	45.5%	788	46.4%
BIPOC (Non-White)	430	45.9%	334	43.8%	764	44.9%
Economically Disadvantaged (ED)	573	61.2%	482	63.2%	1,055	62.1%
Special Education/IEP	80	8.5%	72	9.4%	152	8.9%
English Learner (EL)	45	4.8%	15	2.0%	60	3.5%
Sample Size (Final)	**937**	100.0%	**763**	100.0%	**1,700**	100.0%

Spring 2023 Combined State Assessments Attrition

All schools remained in the project for all three years of implementation (i.e., school-level attrition was 0%). At the individual (student) level, attrition was calculated according to the WWC Handbook's (2022) reference sample 3 (i.e., individuals present in clusters (schools) at follow-up). Overall sample attrition was 12.9%, and differential attrition (i.e., the difference in attrition rates for the treatment vs. comparison group) was 3.4%, which met the WWC low attrition standards.

Spring 2023 Combined State Assessments Outcomes

Since North and South Carolina have different state assessments with different scales for both reading and math achievement, these outcomes were standardized before analyzing the treatment effect. This allowed scores from the two states to be put on the same scale and combined for analysis. Details on the standardization are provided in the **Technical Note** section at the end of this report.

The 2023 fifth grade combined state assessment standardized scores in Reading and Math were the posttests for the analysis, with the spring 2021 third grade Stanford-10 Reading Comprehension and Mathematics Problem Solving subtests as the pretests. The combined treatment and control groups met WWC (2022) baseline equivalence standards for both Reading Comprehension (g = 0.10) and Mathematics Problem Solving (g = 0.09) (i.e., 0.05 < |Effect size at baseline| ≤ 0.25). The state assessment outcome data were analyzed using the same model as the Stanford-10 outcomes.

As shown in **Table 13**, The HLM coefficients for both outcomes were **positive**, meaning students in the treatment group performed **better** than the control group by approximately 0.10 standard deviations in Reading and 0.16 standard deviations in Math ($g_{reading}$ = 0.10; g_{math} = 0.16), but the effects were **not statistically significant** ($p_{reading}$ = 0.40; p_{math} = 0.25). The percentile ranks in **Table 13** indicate that a student who scored at the median (50th percentile) in the treatment group would rank at the 54th percentile (Reading) and 56th percentile (Math) in the comparison group respectively. To aid interpretation of p-values given the different sample sizes, CREP calculated the minimum detectable effect size (MDES) for each analysis, using the lower bound of the 95% confidence interval for the treatment estimate. The MDES shows the smallest effect size that would be required for a statistically significant result (i.e., the effect size for which p = 0.05). As shown in **Table 13**, the MDES values are *higher* than the Hedges' g effect sizes, indicating the sample sizes were likely not large enough to detect a statistically significant difference.

Table 13. Spring 2023 Reading and Math Combined State Assessment Outcomes

	HLM Coefficient	p value	Hedges' g	Percentile Rank	Imp. Index	MDES[1]
Reading	0.08	0.40	0.10	54	4	0.26
Math	0.13	0.25	0.16	56	6	0.32

[1] The MDES shows the smallest effect size that would be required for a statistically significant result (i.e., the effect size for which p = 0.05)

Table 14 shows the estimated marginal means (EMMs) of the standardized Reading and Math scores for the treatment and control groups. EMMs are average standardized scores on the state assessment that have been adjusted to control for differences between the treatment and control groups on other variables (e.g., pretest score, Economically Disadvantaged status). The Reading and

Math EMMs are both **larger** in treatment group than those in the control group, with a difference of 0.07 standard deviations in Reading and 0.12 standard deviations in Math.

Table 14. Spring 2023 Reading and Math Combined State Assessment Outcome SDs and Estimated Marginal Means

Group	*N*	EMM (Reading)	*SD* (Reading)	EMM (Math)	*SD* (Math)
Treatment	937	0.06	0.83	0.10	0.81
Control	763	-0.01	0.80	-0.02	0.86

Spring 2023 Combined State Assessments Technical Note

As North Carolina and South Carolina have their own state assessment systems, students in both states did not take the same test. This means that students with the same test score in different states may not represent the same level of learning performance. Thus, CREP transformed the state assessment scores into standard scores (*z*-scores) for the analyses. The standard scores were calculated for each state separately, and involved calculating separate *z*-scores for each students' achievement in Reading and Math. The formula used for standardizing the test scores within each state was as follows:

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Where x_i is the student's actual score (Reading or Math), $\bar{x}$ is the mean (i.e., average) score of the respective state's sample, and σ is the standard deviation of the scores within the respective state. The scores were standardized with a mean of 0 and standard deviation of 1. In other words, each student's performance was evaluated against the average of the state achievement in Reading and Math within their respective state. Thus, a mean score of zero indicates the mean for that student was exactly equal to the mean score for all students in the respective state (North Carolina or South Carolina). Negative scores indicate that students' performance was below the average, while positive scores indicate performance was above the average. Therefore, any differences observed between treatment and control groups in Reading and Math reflect relative differences in performance compared to the average scores of students within North or South Carolina.

Once the test scores were standardized within each state, the standardized scores were combined for students from both North Carolina and South Carolina. The combined standardized scores in Reading and Math for each student were then utilized for subsequent analyses to ensure a unified dataset.

In summary, standardizing the assessment scores within each state ensures that any observed differences in student achievement are relative to the average performance of students in the respective state, and combining the standardized scores from North Carolina and South Carolina can ensure the consistency of the analyses. The final standardized and combined dataset provides a fair basis for evaluating the effectiveness of the SSEC treatment effect across North and South Carolina.

Longitudinal Professional Development Analysis

The items on the four professional development teacher feedback surveys varied by both module type (Engineering or Physical Science), training type (Introductory or Intermediate) and delivery type (virtual, hybrid, or fully in-person). As a result, only six items were common across all four PD surveys and could be compared across all module, training, and delivery types. Another factor to consider is the grade level of the module on which the participant was trained. The three modules for **Engineering** included (a) How Can We Protect Animals When Their Habitat Changes? (Habitat), (b) How Can We Provide Energy to People's Homes? (Energy), and (c) How Can We Provide Freshwater to Those in Need? (Freshwater). The three **Physical Science** models included (a) How Can We Predict Patterns of Motion? (Motion), (b) How Does Motion Energy Change in a Collision? (Collision), and (c) How Can We Identify Materials Based on Their Properties? (Materials)

Results for the first five of these items involved participant responses to the performance of the SSEC-trained facilitators, and are shown in **Figure 2,** below. The sixth item involved participant understanding of the science content, and is shown in **Figure 3**. Readers should note that, in the interest of readability, the Y-axis of these charts starts at 50% rather than at 0%.

When interpreting these outcomes, readers should note that the PD for each combination of module and type (for example, introductory Engineering) was only delivered once. Additionally, PD delivery progressed from fully virtual at the beginning of the project, to hybrid in the middle, to fully in-person at the end. In addition, the facilitators also varied across years. These results, therefore, cannot be interpreted as being caused by any one specific factor. CREP addressed the four following questions related to longitudinal outcomes for the teacher PD surveys.

Question 1: What patterns, if any, can be seen in the response to introductory relative to intermediate PD?

Across the five questions about **facilitator performance (Figure 2)**, the intermediate (i.e., content-focused) PD responses were slightly *less favorable* than the introductory (curriculum-focused) PD responses. However, this effect was not consistent across all modules. For example, on the item about **facilitator encouragement of teaching strategies**, positive responses for the Collision module *decreased* from 100% to 81%, while another Physical Science module (Motion) module *increased* from 83% to 100% positive responses.

For the question about the **science content of the unit (Figure 3)**, the percentage of positive responses was, on average, stable between introductory and intermediate PD. Meanwhile, positive responses *increased* for the Collision module, and *decreased* for the Habitat and Energy modules.

Figure 2: The SSEC-trained facilitators...

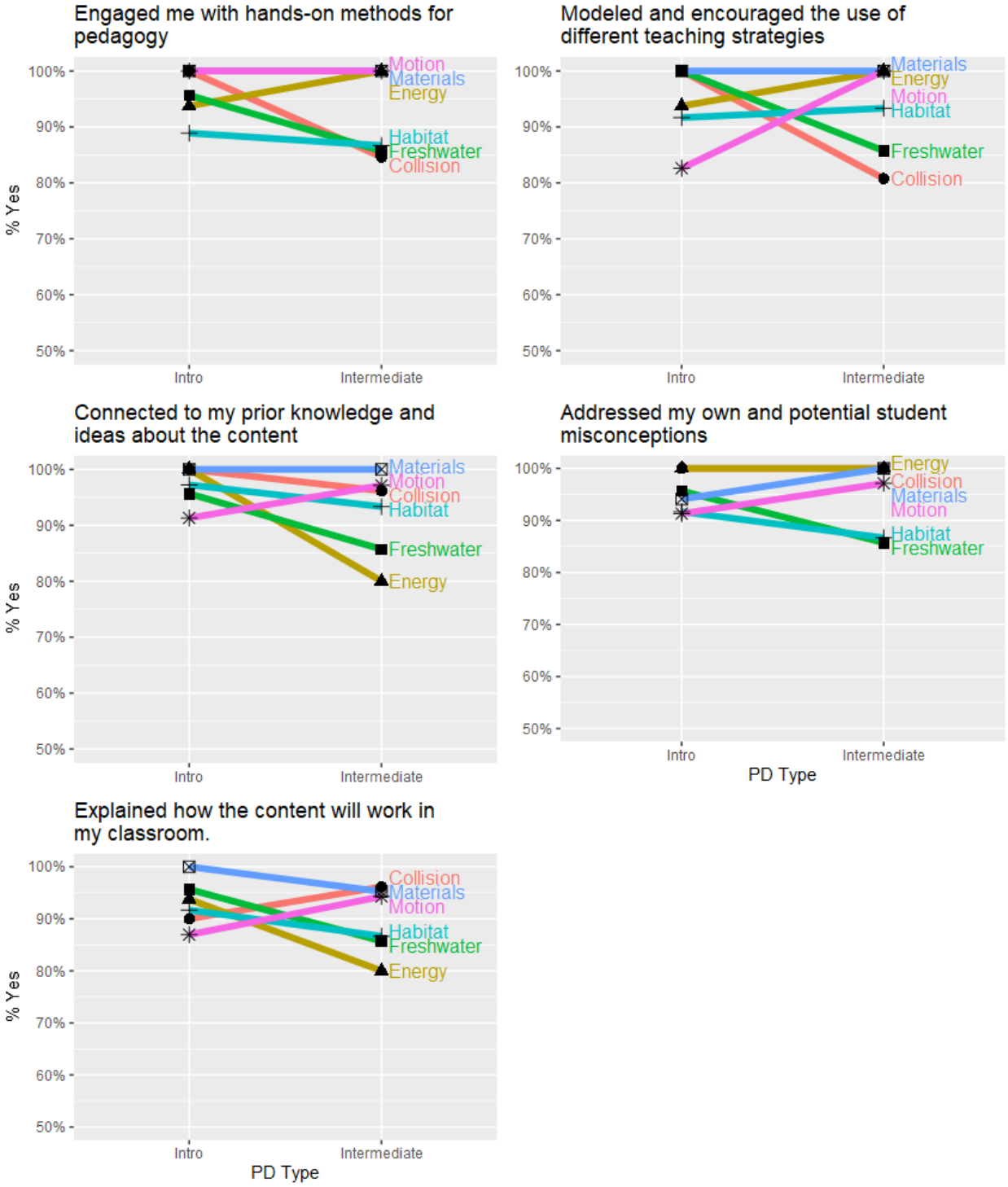
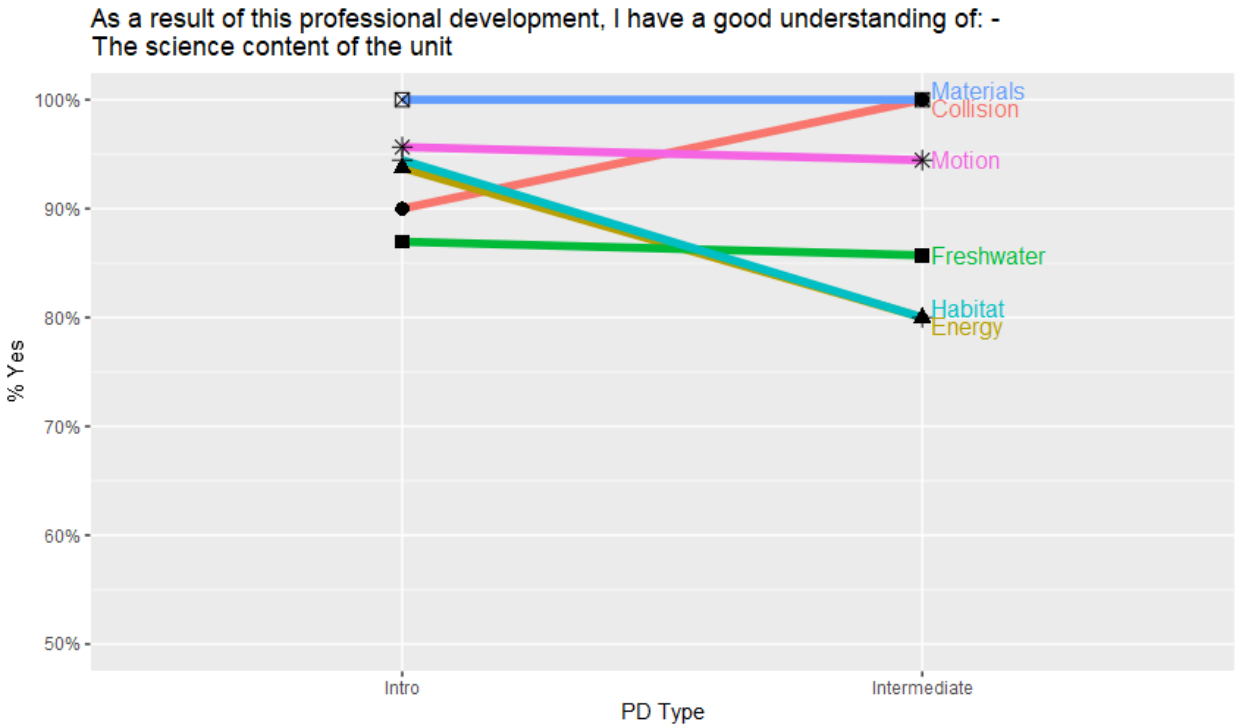


Figure 3: PD Participant Understanding of Science Content



Question 2: What patterns, if any, can be seen in the response to Engineering relative to Physical Science PD?

As shown in **Table 15**, when averaged across PD type (introductory and intermediate), responses to the **Physical Science** PD were slightly *more positive* than responses to the **Engineering** PD. However, these differences are relatively small, and their causal factor is unclear. For example, since the Engineering PD preceded the Physical Science PD, based on the teacher feedback on the PD surveys, SSEC may have become better at designing the PD and training the trainers for the Physical Science PD. In addition, it is possible that teachers had greater initial comfort with the Physical Science vs. Engineering concepts, and further benefited from the hybrid and in-person modality of the Physical Science trainings relative to the Engineering workshops.

Table 15: Average PD Responses for Engineering and Physical Science

The SSEC trained facilitators…	% Yes	
	Engineering	Physical Sci
Engaged me with hands-on methods for pedagogy	91.8%	97.4%
Modeled and encouraged the use of different teaching strategies	94.1%	93.9%
Connected to my prior knowledge and ideas about the content	92.0%	97.4%
Addressed my own and potential student misconceptions	93.3%	97.1%
Explained how the content will work in my classroom.	88.9%	93.8%
As a result of this PD, I have a good understanding of	**Engineering**	**Physical Sci**
The science content of the unit	91.7%	97.1%

Question 3: Across the four workshops, are there any noteworthy results pertaining to a specific training or grade-level breakout?

Across both PDs and grade levels, responses were primarily positive for all question categories. Among items with at least 10 responses, the **lowest-scoring category** was the *thought homework* in the **Introductory Engineering PD**, with an average of 79.7% positive responses across all modules and questions. However, participants in the fourth grade breakout still gave the thought homework an average of 91.7% positive responses across all questions.

The **second lowest-scoring category** was the *review of past student work* in the **Intermediate Engineering PD**, with an average of 82.6% positive responses across all modules and questions. However, this PD had a considerably lower number of total responses than average: *N* = 27, with only *n* = 5 for fourth grade and *n* = 7 for fifth grade.

Question 4: Across the four workshops, which portions of the training were perceived to be most helpful or viewed most positively (e.g., reflection time, content breakouts, review of student work)?

Many questions had close to 100% positive responses, but with a very small sample size. Among questions with at least 10 responses, 12 had 95%+ positive feedback. All 12 questions were in the **Introductory Engineering PD**.

Table 16: Highest Positive PD Responses for Engineering and Physical Science

Module	Question	N	% Yes
Habitat	Connected to my prior knowledge and ideas about the content	36	97.2%
Energy	Connected to my prior knowledge and ideas about the content	16	100.0%
Energy	Addressed my own and potential student misconceptions	16	100.0%
Energy	The overview of supports for Smithsonian Science for the Classroom through Carolina Science Online presented on day 1: - Showed me new ways to access information and resources	16	100.0%
Energy	As a result of this professional development, I have a good understanding of: - Why the unit is organized the way it is	16	100.0%
Energy	As a result of this professional development, I have a good understanding of: - The use and management of the materials	16	100.0%
Freshwater	Engaged me with hands-on methods for pedagogy	23	95.7%
Freshwater	Modeled and encouraged the use of different teaching strategies	23	100.0%
Freshwater	Connected to my prior knowledge and ideas about the content	23	95.7%
Freshwater	Addressed my own and potential student misconceptions	23	95.7%
Freshwater	Explained how the content will work in my classroom.	23	95.7%
Freshwater	During the sections where I led lesson(s) for my peers, I felt: - That I received sufficient guidance from the trainer	23	95.7%

Note: For questions with at least 10 responses

Longitudinal Teacher Module Logs

As part of their participation in *Smithsonian Science for North and South Carolina Classrooms,* teachers were asked to fill out Module Logs each year where they reported their experiences and opinions regarding classroom use of the modules. Of the six *Smithsonian Science* modules used in this intervention, only the three Engineering modules had more than one year of module log data available at the time of this report: *How Can We Protect Animals When Their Habitat Changes?*, *How Can We Provide Energy to People's Homes?*, and *How Can We Provide Freshwater to Those in Need?*

This section of the report uses these data to examine whether teacher opinions of the modules changed over time, with an emphasis on module use in the classroom. While CREP also collected data on virtual learning during the 2020-2021 and 2021-2022 academic years, the number of responses to those questions was too low to support meaningful analysis.

Use of Modules in Classrooms

Module Log items 1-16 asked teachers about the ways they used modules in the classroom. The longitudinal results for modules with three years of data are shown in **Figure 4** and **Figure 5**. Results are presented for each module individually, and across all modules combined. Response counts by module and year are given in **Table 17**.

Across all three years, teachers **consistently** reported they:

- had all necessary materials
- felt comfortable with the science content of the modules
- had sufficient training to teach the modules

Teachers became **less** likely to report that:

- they taught lessons in the suggested sequence
- modules were easy to use
- modules could comfortably fit in a class period
- they taught the module during instructional time not intended for science

Teachers became **more** likely to report that:

- they supplemented the lessons with materials from other sources

Table 17: Response Counts to Engineering Module Use Questions

Year	Engineering Module			Combined
	Habitat	Energy	Freshwater	
2020-21	8	6	5	19
2021-22	20	12	12	44
2022-23	14	4	6	24
Total	**42**	**22**	**23**	**87**

Figure 4: Responses by Module, Q1-Q8

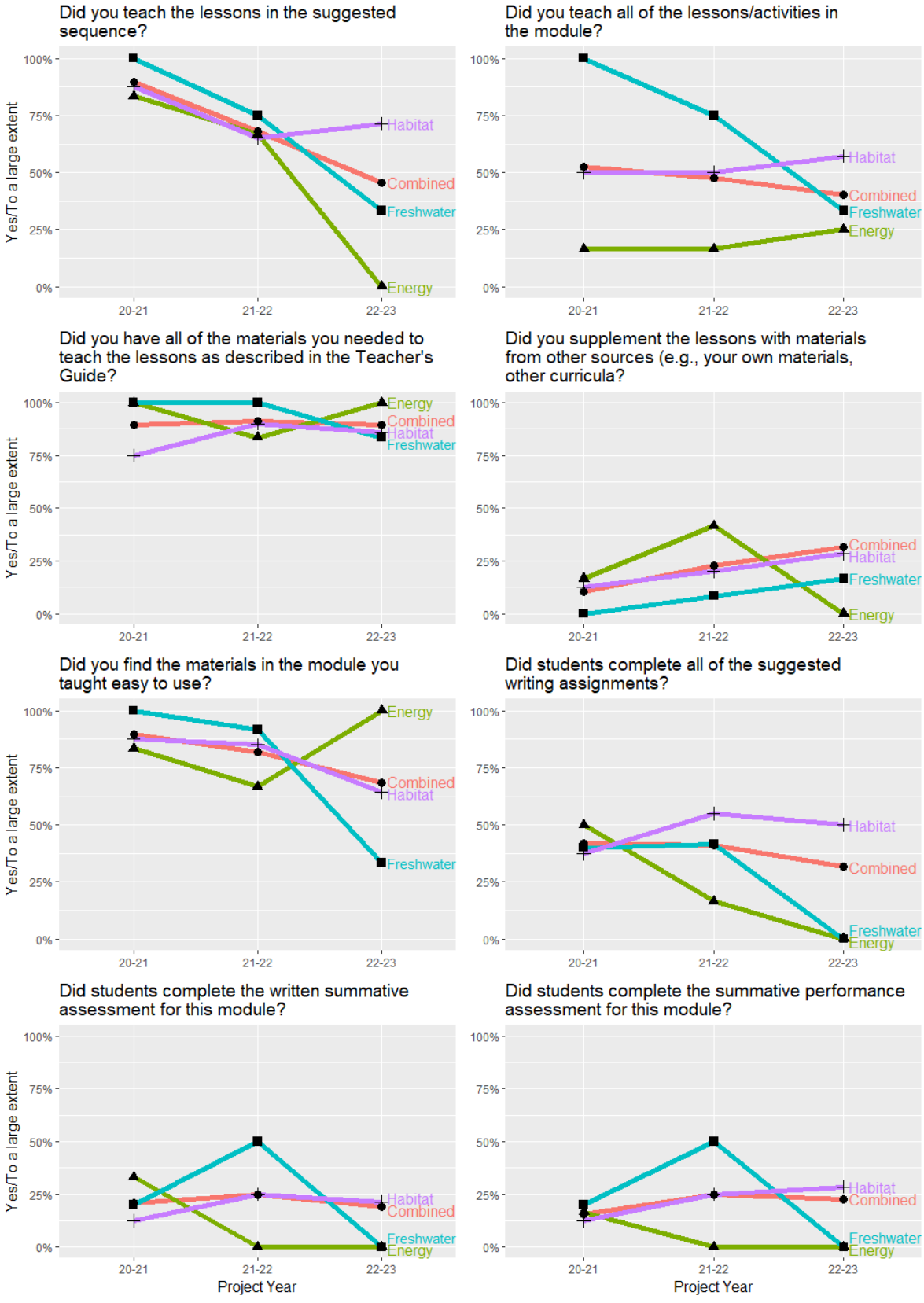


Figure 5: Responses by Module, Q9-Q16

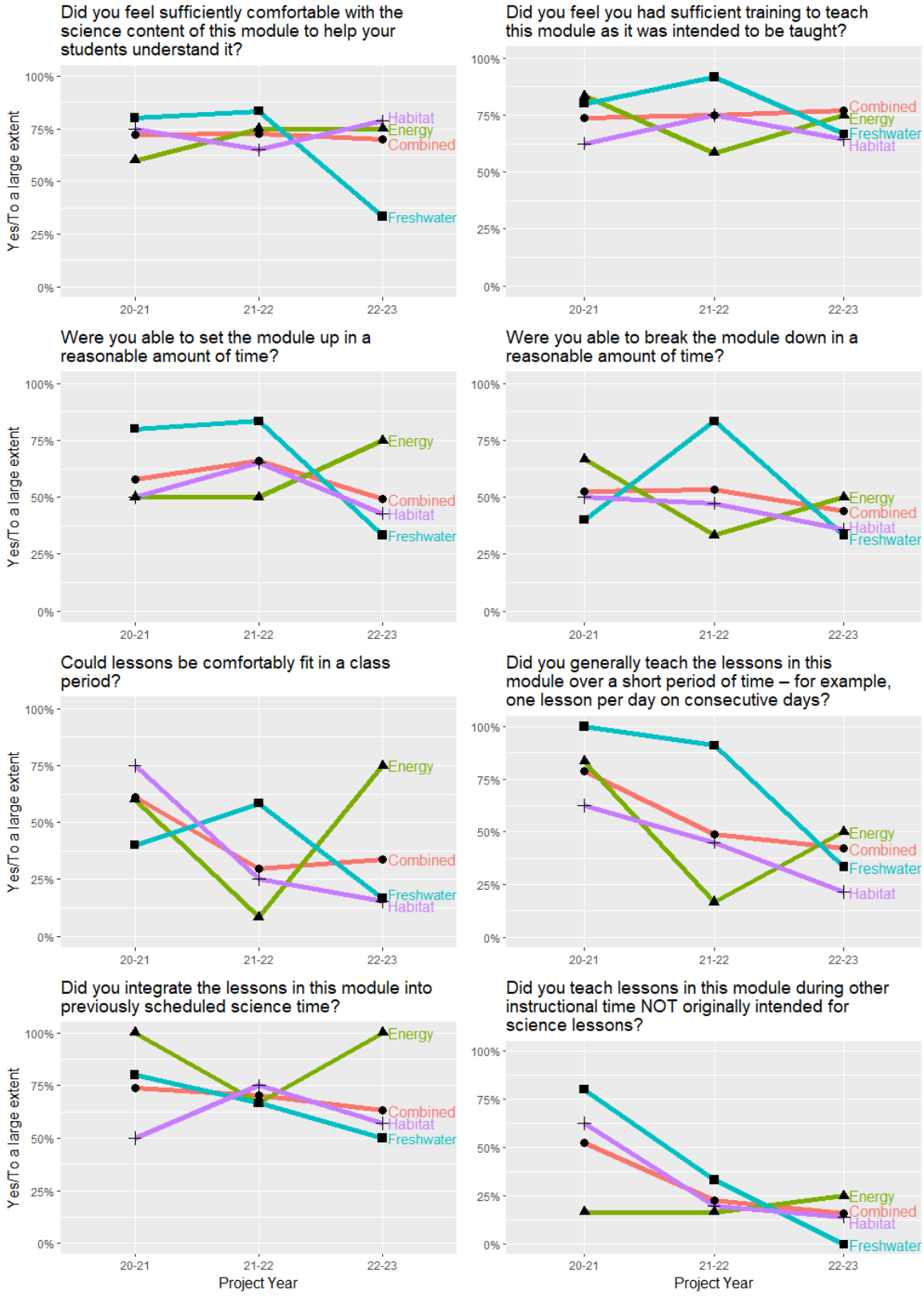


Table 18: Emphasis on Student Notebook Quality

Module	Year	Count				Percentage		
		Not much emphasis	Moderate emphasis	Strong emphasis	Total	Not much emphasis	Moderate emphasis	Strong emphasis
How Can We Protect Animals When Their Habitat Changes?	2020-21	4	2	2	*8*	50.0%	25.0%	25.0%
	2021-22	4	13	3	*20*	20.0%	65.0%	15.0%
	2022-23	5	9	0	*14*	35.7%	64.3%	0.0%
How Can We Provide Energy to People's Homes?	2020-21	1	4	1	*6*	16.7%	66.7%	16.7%
	2021-22	4	8	0	*12*	33.3%	66.7%	0.0%
	2022-23	0	4	0	*4*	0.0%	100.0%	0.0%
How Can We Provide Freshwater to Those in Need?	2020-21	1	3	1	*5*	20.0%	60.0%	20.0%
	2021-22	2	10	0	*12*	16.7%	83.3%	0.0%
	2022-23	2	3	0	*5*	40.0%	60.0%	0.0%

Table 19: Assessment of Student Learning

Module	Year	Count				Percentage[1]		
		Strategic Questions	Provided Asmnt	Notebook Entries	Total	Strategic Questions	Provided Asmnt	Notebook Entries
How Can We Protect Animals When Their Habitat Changes?	2020-21	5	3	4	*7*	71.4%	42.9%	57.1%
	2021-22	13	7	12	*19*	68.4%	36.8%	63.2%
	2022-23	13	3	9	*14*	92.9%	21.4%	64.3%
How Can We Provide Energy to People's Homes?	2020-21	5	3	5	*6*	83.3%	50.0%	83.3%
	2021-22	10	4	4	*11*	90.9%	36.4%	36.4%
	2022-23	4	1	3	*4*	100.0%	25.0%	75.0%
How Can We Provide Freshwater to Those in Need?	2020-21	2	2	2	*4*	50.0%	50.0%	50.0%
	2021-22	6	9	6	*10*	60.0%	90.0%	60.0%
	2022-23	3	0	1	*4*	75.0%	0.0%	25.0%

[1] Sums to more than 100% because teachers could select multiple options

For modules that had three years of data, teacher responses to the question, *How much emphasis did you place on the quality of student notebook entries?* are shown above in **Table 18**, with the highest percentage response every year for each module highlighted in green. Teachers maintained approximately the same level of emphasis (Moderate) on notebook entries for all three years for the *Habitat* and *Energy* modules. Teachers using the *Freshwater* module became slightly less likely to place emphasis on notebook entries over time. However, the low number of teachers responding in the first and third years means any differences should be interpreted cautiously.

Teacher responses to the question, *How did you determine what students learned?* are shown above in **Table 19**. For the *Habitat* and *Energy* modules, teachers were consistently more likely to assess learning with strategic questions than with student notebook entries or the provided assessments. However, no clear pattern emerged for teachers using the *Freshwater* module.

Teacher Overall Opinions of Modules

Teacher opinion of module levels of alignment with their state science education standards are presented in **Table 20**. Because these questions may have differential responses by state, their results are reported separately by state. Due to resulting small counts in each cell, the response options "A little" and "not at all" have been collapsed into "No", and the options "A great deal" and "completely or almost completely" have been collapsed into "Yes".

Teachers in **North Carolina** consistently felt the *Habitat* and *Freshwater* modules were not strongly aligned with their state standards. While their opinions of the *Energy* module appear to have shifted over time, the low response count for that module makes it difficult to draw trustworthy conclusions.

Teachers in **South Carolina** were more positive about module alignment with state standards, with most teachers feeling the *Habitat* module was strongly aligned for all three years. While responses were generally negative for the *Energy* module and moderate for the *Freshwater* module, the low response count per cell for those modules is again a challenge to drawing conclusions.

Table 20: Teacher Opinion of Module Fit to Science Standards

		Count				Percentage		
Module	**Year**	**No[1]**	**Somewhat**	**Yes[2]**	**Total**	**No**	**Somewhat**	**Yes**
North Carolina								
How Can We Protect Animals When Their Habitat Changes?	20-21	3	0	0	*3*	100.0%	0.0%	0.0%
	21-22	7	1	0	*8*	87.5%	12.5%	0.0%
	22-23	6	0	0	*6*	100.0%	0.0%	0.0%
How Can We Provide Energy to People's Homes?	20-21	0	1	2	*3*	0.0%	33.3%	66.7%
	21-22	1	5	2	*8*	12.5%	62.5%	25.0%
	22-23	1	1	0	*2*	50.0%	50.0%	0.0%
How Can We Provide Freshwater to Those in Need?	20-21	2	0	1	*3*	66.7%	0.0%	33.3%
	21-22	4	1	3	*8*	50.0%	12.5%	37.5%
	22-23	3	0	0	*3*	100.0%	0.0%	0.0%
North Carolina Total		*36*	*16*	*13*	*65*	*55.4%*	*24.6%*	*20.0%*

South Carolina								
How Can We Protect Animals When Their Habitat Changes?	20-21	0	2	3	*5*	0.0%	40.0%	60.0%
	21-22	2	3	6	*11*	18.2%	27.3%	54.5%
	22-23	3	1	4	*8*	37.5%	12.5%	50.0%
How Can We Provide Energy to People's Homes?	20-21	0	2	1	*3*	0.0%	66.7%	33.3%
	21-22	3	0	1	*4*	75.0%	0.0%	25.0%
	22-23	2	0	0	*2*	100.0%	0.0%	0.0%
How Can We Provide Freshwater to Those in Need?	20-21	0	2	0	*2*	0.0%	100.0%	0.0%
	21-22	3	1	0	*4*	75.0%	25.0%	0.0%
	22-23	0	1	0	*1*	0.0%	100.0%	0.0%
South Carolina Total		*16*	*17*	*17*	*50*	*32.0%*	*34.0%*	*34.0%*

[1] Sum of responses "a little" and "not at all"

[2] Sum of responses "A great deal" and "completely or almost completely"

As shown in **Table 21**, teachers in both states most often responded 'Maybe' when asked about their willingness to continue using modules. Teachers in **South Carolina** were slightly more likely to respond 'Yes' than teachers in **North Carolina**. Differences between states were greatest for the *Freshwater* module, with 100% of **North Carolina** teachers during 2022-23 reporting that they *would not* continue using the module, and 100% of **South Carolina** teachers reporting that they *would* continue using it. However, the small response counts in each cell limit the ability to draw conclusions for this module.

Table 21: Teacher Willingness to Continue Using Modules

		Count				Percentage		
Module	**Year**	**No**	**Maybe**	**Yes**	**Total**	**No**	**Maybe**	**Yes**
North Carolina								
How Can We Protect Animals When Their Habitat Changes?	20-21	0	2	1	*3*	0.0%	66.7%	33.3%
	21-22	3	5	0	*8*	37.5%	62.5%	0.0%
	22-23	2	2	2	*6*	33.3%	33.3%	33.3%
How Can We Provide Energy to People's Homes?	20-21	0	2	1	*3*	0.0%	66.7%	33.3%
	21-22	0	3	5	*8*	0.0%	37.5%	62.5%
	22-23	0	2	0	*2*	0.0%	100.0%	0.0%
How Can We Provide Freshwater to Those in Need?	20-21	1	2	0	*3*	33.3%	66.7%	0.0%
	21-22	2	3	3	*8*	25.0%	37.5%	37.5%
	22-23	3	0	0	*3*	100.0%	0.0%	0.0%
North Carolina Total		*11*	*21*	*12*	*44*	*25.0%*	*47.7%*	*27.3%*
South Carolina								
How Can We Protect Animals When Their Habitat Changes?	20-21	0	2	3	*5*	0.0%	40.0%	60.0%
	21-22	2	6	3	*11*	18.2%	54.6%	27.3%
	22-23	2	5	1	*8*	25.0%	62.5%	12.5%
	20-21	0	1	2	*3*	0.0%	33.3%	66.7%
	21-22	1	3	0	*4*	25.0%	75.0%	0.0%

How Can We Provide Energy to People's Homes?	22-23	0	2	0	*2*	0.0%	100.0%	0.0%
How Can We Provide Freshwater to Those in Need?	20-21	0	0	2	*2*	0.0%	0.0%	100.0%
	21-22	0	1	3	*4*	0.0%	25.0%	75.0%
	22-23	0	0	1	*1*	0.0%	0.0%	100.0%
South Carolina Total		*5*	*20*	*15*	*40*	*12.5%*	*50.0%*	*37.5%*

Classroom Observations

School Observation Measure (SOM)

Targeted observations were conducted in available treatment and control school science classrooms during the 2021-22 and 2022-23 academic years using the School Observation Measure (SOM). To make results comparable across years, only schools that participated in observations both years were included for analysis. For **treatment schools**, these observations incorporate science taught using the *Smithsonian Science* curricular modules as well as science taught using other methods. However, as **control schools** have not begun *Smithsonian Science* implementation, their observations incorporate science instructional methods other than curricular modules. Each targeted observation consisted of a single science classroom visit that lasted at least 30 minutes, although many continued for the duration of the class period. The results of the most and least frequently observed items will be presented by treatment and control groups.

Treatment Schools.

During the 2021-22 and 2022-23 academic years, a total of 95 targeted SOM observations were conducted in treatment science classrooms across both North and South Carolina. The percentage of responses for each category and strategy, along with the two summary items, is shown in **Table 22** by group (treatment or control), with the most prevalent activities ***observed*** highlighted in green. For most strategies, the highest percentage was either "Not Observed" or "Rarely Observed" (17/24 = 71% of strategies). Of those observed, the most prevalent strategies, based on being rated "Extensively" or "Frequently" observed (**Table 22**) both years were: Direct Instruction (2021-22: 33.3% and 2022-23: 49.2%) and Teacher acting as coach facilitator (2021-22: 36.7% and 2022-23: 32.3%). In terms of the *overall classroom environment* (i.e., the summary items), high academically focused class time was most often "Extensively" or "Frequently" observed both years (2021-22: 56.7% and 2022-23: 70.8%).

Control Schools.

A total of 90 targeted observations were conducted in available control science classrooms across both North and South Carolina during the 2021-22 and 2022-23 academic years. However, these schools have not begun implementing *Smithsonian Science* curricular modules in their classrooms. Like treatment schools, for most strategies, the highest percentage was either not observed or rarely observed (21/24 = 88% of strategies), with three of those strategies (13%) rated at 100% "not observed or rarely" across both years of the study: Individual tutoring, Parent/community involvement in learning activities, and Student self-assessment. The most prevalent strategies observed and rated "Extensively" or "Frequently" (**Table 22**) both years were: Direct Instruction (2021-22: 52.0% and 2022-23: 63.1%) and

Teacher acting as coach facilitator (2021-22: 32.0% and 2022-23: 30.8%). As with treatment classrooms, in terms of the *overall classroom environment* (i.e., the summary items), high academically focused class time was most often "Extensively" or "Frequently" observed both years (2021-22: 56.0% and 2022-23: 72.3%).

Three notable areas where the treatment and control groups diverged included (based on the highest percentage of responses):

- **Cooperative/Collaborative Learning**: Extensively/Frequently observed more often across both years of the study in **treatment** (20% and 45% of observations, respectively) than in control school (8% and 23% respectively) classrooms.
- **Use of higher-level questioning strategies**: Extensively/Frequently observed more often across both years of the study in **treatment** (10% and 26% of observations respectively) than in control school (4% and 11% respectively) classrooms.
- **Experiential hands-on learning:** Extensively/Frequently observed more often in the last year of the study in **treatment** (37% of observations) than in control school (23%) classrooms.

Three noteworthy areas that moved from being observed "Extensively" or "Frequently" more often in control schools in 2021-22 to treatment schools in 2022-23 (based on percentage point differences) were the strategies of **Student discussion** (-10.0 in treatment 2021-22 to +6.2 in treatment 2022-23) and **Performance assessment strategies** (-12.0 in treatment 2021-22 to +1.5 in treatment 2022-23), and the summary item **High level of student attention/interest/engagement** (-1.3 in treatment 2021-22 to +13.9 in treatment 2022-23).

Table 22: SOM for North and South Carolina 2021-2022 and 2022-23

Strategy	Treatment (N=95)						Control (N=90)					
	% Not Observed + Rarely		% Occasionally		% Extensively + Frequently		% Not Observed + Rarely		% Occasionally		% Extensively + Frequently	
	2021 22	2022 23	2021 22	2022 23	2021 22	2022 23	2021 22	2022 23	2021 22	2022 23	2021 22	2022 23
Instructional Orientation												
Direct Instruction	23.3%	24.6%	43.3%	26.2%	**33.3%***	**49.2%***	40.0%	21.5%	8.0%	15.4%	**52.0%***	**63.1%***
Team teaching	93.3%	86.2%	0.0%	3.1%	6.7%	10.8%	100.0%	84.6%	0.0%	0.0%	0.0%	15.4%
Cooperative/collaborative learning	60.0%	43.1%	20.0%	12.3%	20.0%	44.6%	80.0%	64.6%	12.0%	12.3%	8.0%	23.1%
Individual tutoring (teacher, peer, aide, adult volunteer)	100.0%	96.9%	0.0%	1.5%	0.0%	1.5%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%
Classroom Organization												
Ability groups	93.3%	96.9%	0.0%	1.5%	6.7%	1.5%	100.0%	98.5%	0.0%	1.5%	0.0%	0.0%
Multi-age grouping	100.0%	98.5%	0.0%	0.0%	0.0%	1.5%	100.0%	98.5%	0.0%	0.0%	0.0%	1.5%
Work centers (for individuals or groups)	100.0%	98.5%	0.0%	0.0%	0.0%	1.5%	100.0%	96.9%	0.0%	0.0%	0.0%	3.1%
Instructional Strategies												
Higher level instructional feedback (written or verbal) to enhance student learning	83.3%	78.5%	13.3%	13.8%	3.3%	7.7%	96.0%	86.2%	0.0%	9.2%	4.0%	4.6%
Integration of subject areas (interdisciplinary/thematic units)	93.3%	93.8%	6.7%	4.6%	0.0%	1.5%	96.0%	89.2%	4.0%	1.5%	0.0%	9.2%
Project-based learning	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	92.0%	96.9%	0.0%	1.5%	8.0%	1.5%

Use of higher-level questioning strategies	73.3%	47.7%	16.7%	26.2%	10.0%	26.2%	76.0%	66.2%	20.0%	23.1%	4.0%	10.8%
Teacher acting as a coach/facilitator	33.3%	44.6%	30.0%	23.1%	**36.7%***	**32.3%***	56.0%	56.9%	12.0%	12.3%	**32.0%***	**30.8%***
Parent/community involvement in learning activities	100.0%	96.9%	0.0%	1.5%	0.0%	1.5%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%
Student Activities												
Independent seatwork (self-paced worksheets, individual assignments)	73.3%	95.4%	16.7%	4.6%	10.0%	0.0%	64.0%	93.8%	24.0%	4.6%	12.0%	1.5%
Experiential, hands-on learning	70.0%	52.3%	16.7%	10.8%	13.3%	36.9%	64.0%	66.2%	16.0%	10.8%	20.0%	23.1%
Systematic individual instruction (differential assignments geared to individual needs)	100.0%	96.9%	0.0%	1.5%	0.0%	1.5%	100.0%	98.5%	0.0%	1.5%	0.0%	0.0%
Sustained writing/composition (self-selected or teacher-generated topics)	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	96.9%	0.0%	1.5%	0.0%	1.5%
Sustained reading	100.0%	98.5%	0.0%	1.5%	0.0%	0.0%	100.0%	98.5%	0.0%	0.0%	0.0%	1.5%
Independent inquiry/research on the part of students	96.7%	84.6%	3.3%	4.6%	0.0%	10.8%	96.0%	86.2%	4.0%	7.7%	0.0%	6.2%
Student discussion	76.7%	86.2%	20.0%	6.2%	3.3%	7.7%	96.0%	90.8%	4.0%	7.7%	13.3%	1.5%
Technology Use												

Computer for instructional delivery (e.g., CAI, drill & practice)	36.7%	44.6%	40.0%	27.7%	23.3%	27.7%	48.0%	47.7%	32.0%	20.0%	20.0%	32.3%
Technology as a learning tool or resource (e.g., Internet research, spreadsheet or database creation, multi-media, CD Rom, Laser disk)	73.3%	92.3%	20.0%	4.6%	6.7%	3.1%	84.0%	80.0%	4.0%	7.7%	12.0%	12.3%
Assessment												
Performance assessment strategies	93.3%	98.5%	6.7%	0.0%	0.0%	1.5%	88.0%	100.0%	0.0%	0.0%	12.0%	0.0%
Student self-assessment (portfolios, individual record books)	100.0%	96.9%	0.0%	1.5%	0.0%	1.5%	100.0%	100.0%	0.0%	0.0%	0.0%	0.0%
Summary Items												
High academically focused class time	10.0%	0.0%	33.3%	29.2%	**56.7%***	**70.8%***	12.0%	0.0%	32.0%	27.7%	**56.0%***	**72.3%***
High level of student attention/interest/engagement	3.3%	3.1%	50.0%	26.2%	46.7%	70.8%	0.0%	1.5%	52.0%	41.5%	48.0%	56.9%

Note: Percentages in bold with an asterisk (*) represent the top three most prevalent strategies observed for each group.

Rubric for Inquiry-Based Assessment (RIBA)

Observers used the RIBA concurrently with the SOM to rate (a) the frequency of inquiry-based learning strategies employed in the classroom and (b) the overall level of class time dedicated to inquiry-based science. As with the SOM, RIBA observations were conducted in treatment and control school science classrooms in the 2021-22 and 2022-23 academic years. Results have been made comparable across years by including only schools that participated in observations both years. The results of the most and least frequently observed items are presented by treatment and control groups below.

Treatment Schools.

During the 2022-23 academic year, 95 RIBA observations were conducted in available treatment science classrooms across both North and South Carolina. The percentage of responses for each strategy is shown in **Table 23** by group (treatment or control). For most student-centered activities (7/10 = 70%), the higher percentage both years was "Not Observed." The three most prevalent activities that were observed both years in treatment classrooms (highlighted in green) were: Prepared science kits or modules in use (2021-22: 32.3% and 2022-23: 67.7%), Students engaged in experimentation (2021-22: 29.0%, and 2022-23: 52.3%), and Students gathering or recording evidence (2021-22: 29.0% and 2022-23: 56.9%).

Control Schools.

A total of 90 RIBA observations were conducted in available control science classrooms across both North and South Carolina during the 2022-23 academic year. Compared to treatment schools, *"Not Observed"* was the *higher percentage for all student-centered activities across both years of the study (10/10 = 100%).* The three most prevalent activities observed both years (**Table 23**) were: Students hypothesizing or making predictions (2021-22: 24.0% and 2022-23: 32.8%), Students engaged in experimentation (2021-22: 32.0% and 2022-23: 37.3%), and Students gathering or recording evidence (2021-22: 28.0% and 2022-23: 34.3%).

Of note, **7 out of 10 student-centered activities were observed at a higher rate in treatment classrooms** than control classrooms in the final year (2022-23). Five notable areas where the treatment and control groups **diverged** across one or both years were the following (based on the percentage observed):

- **Prepared science kits or modules in use**: Was observed across *both years* of the study at least twice **as often** in **treatment** vs. control classroom observations.
- **Students organizing data or preparing to organize data**: Was observed almost 50% more often in **treatment** vs. control classroom observations in the final (2022-23) year.
- **Students engaged in experimentation**: Was observed over 40% more frequently in **treatment** vs. control classroom observations in the final year.
- **Students gathering or recording evidence**: Was observed nearly two-thirds more often in treatment schools in the final year, and was observed more often than control schools across *both years* of the study.

- **Students evaluating evidence**: Was observed in the final year over twice as often in **treatment** vs. control classroom observations.

Table 23: RIBA for North and South Carolina 2021-22 and 2022-23

Student Centered Activities	Treatment (N=95)				Control (N=90)			
	Not Observed		Observed		Not Observed		Observed	
	2021 22	2022 23	2021 22	2022 23	2021-22	2022-23	2021-22	2022-23
Prepared science kits or modules in use	67.7%	32.3%	**32.3%***	**67.7%***	84.0%	94.0%	16.0%	6.0%
Students organizing data or preparing to organize data	96.8%	58.5%	3.2%	41.5%	84.0%	71.6%	16.0%	28.4%
Students making predictions or hypothesizing	83.9%	64.6%	16.1%	35.4%	76.0%	67.2%	**24.0%***	**32.8%***
Students designing their own procedures	96.8%	89.2%	3.2%	10.8%	96.0%	80.6%	4.0%	19.4%
Teacher demonstrating	90.3%	84.6%	9.7%	15.4%	84.0%	82.1%	16.0%	17.9%
Students engaged in experimentation	71.0%	47.7%	**29.0%***	**52.3%***	68.0%	62.7%	**32.0%***	**37.3%***
Students initiating questions about the experiment	90.3%	86.2%	9.7%	13.8%	88.0%	83.6%	12.0%	16.4%
Students gathering or recording evidence	71.0%	43.1%	**29.0%***	**56.9%***	72.0%	65.7%	**28.0%***	**34.3%***
Students evaluating evidence	74.2%	72.3%	25.8%	27.7%	80.0%	86.6%	20.0%	13.4%
Students reporting findings to others	93.5%	87.7%	6.5%	12.3%	92.0%	98.5%	8.0%	1.5%

Note: Percentages in bold with an asterisk (*) represent the top three most prevalent activities observed for each group.

The RIBA summary item "**Level of class time dedicated to inquiry-based science**" was rated as "high" four times as often in control schools (12% vs. 3%) in the first year of the study (**Figure 6**), but over 50% higher in treatment (34%) vs. control (22%) classroom observations in the last year of the study (**Figure 7**).

Figure 6: *RIBA Level of Class time Dedicated to Inquiry-Based Science 2021-22*

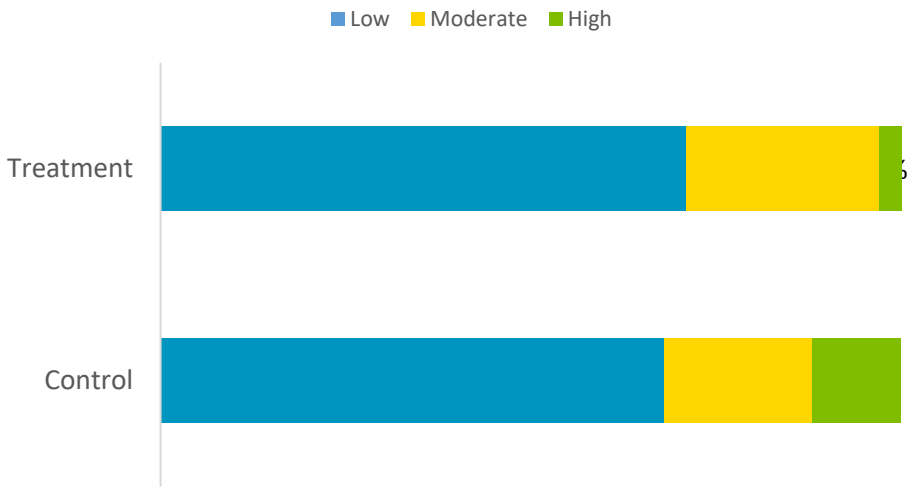
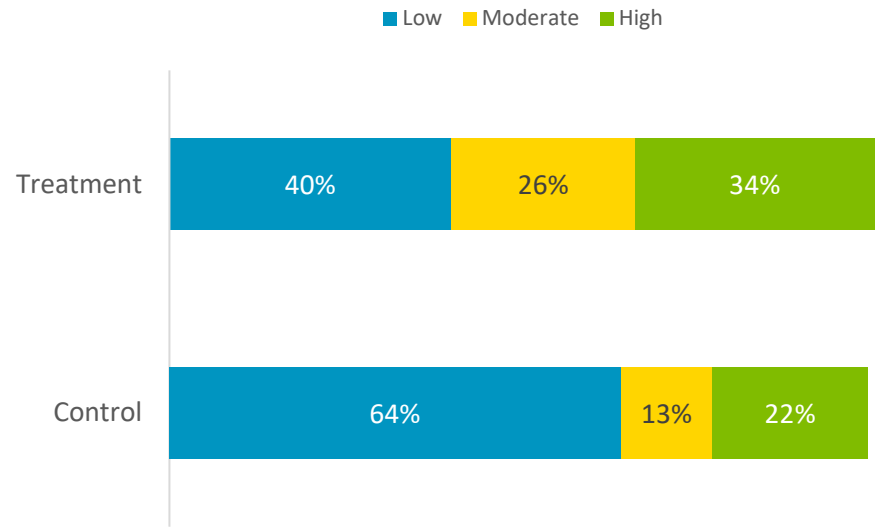


Figure 7: RIBA Level of Class time Dedicated to Inquiry-Based Science 2022-23



Inter-Rater Reliability

For both the SOM and the RIBA, inter-rater reliability (IRR) results were acceptable in both years during which school observations were conducted. Ratings for the **RIBA** were considerably higher in Year 2 (2022-23) than in Year 1 (2021-22). However, IRR for the RIBA was assessed using a different

procedure in Year 2 (Gwet's AC2) than in Year 1 (Cohen's weighted Kappa, Kw). Between the two, AC2 is considered more accurate (Gwet, 2014), and thus is likely closer to the true values.

2021-2022.

Inter-rater reliability between each rater and expert pair on the **SOM and RIBA** was assessed using Cohen's weighted Kappa (linear weights), which determines the extent of agreement between two observers that is greater than expected by chance (chance corrected agreement). For these analyses, the weighted Kappa statistic (k_w) is particularly appropriate when ratings are provided in ordered-categorical form, as in this case, and where raters scored on a continuum with five levels ranging from 0-4 (i.e., Not Observed=0, Rarely Observed=1, Occasionally Observed=2, Frequently Observed=3, Extensively Observed=4). For items like these, k_w would assign less "weight" to ratings that were farther apart (more disagreement).

For the interpretation of k_w, whether weighted or not, values between 0.21 and 0.40 are conventionally interpreted as an indication of "fair" agreement between two raters, 0.41 to 0.60 are an indication of "moderate" agreement, 0.61 to 0.80 are an indication of "substantial" agreement, while values of 0.81 or higher are conventionally interpreted as signs of "almost perfect" agreement (Landis & Koch, 1977) (**Table 24**). Averaged across the nine videos, the weighted Kappa values across all raters were in the "Moderate" or "Substantial" range on the SOM (**Table 27**), and in the "Fair", "Moderate", or "Substantial" range on the RIBA (**Table 28**).

As a note, one observer combined videos seven through nine into a single sheet. Since that data could not be separated out by video, it was not included in this analysis.

Table 24. Interpretation of Kappa Values (Landis & Koch, 1977)

Values Between	Level of Agreement
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 or higher	Almost Perfect

2022-2023.

Inter-rater reliability between each rater and expert pair on the **SOM** was assessed with an ICC(3,1) absolute agreement model as described in Gwet (2014). This is a mixed ANOVA with rater as a fixed effect, item as a random effect, and rating as the dependent variable. Rater was a fixed rather than random effect because CREP was interested in assessing these specific raters, not generalizing to a population of raters. The ICC values in **Table 29** and **Table 30** below are calculated from the ratio of rater agreement to total variance.

On the **SOM**, ratings are provided in ordered-categorical form, with five levels ranging from 0-4 (i.e., Not Observed = 0, Rarely Observed = 1, Occasionally Observed = 2, Frequently Observed = 3, Extensively Observed = 4). The ICC is appropriate for these data, and any data where ANOVA would be appropriate, since it is essentially an effect size for a specified type of ANOVA (Gwet, 2014).

For the overall average, an ICC (3,1) absolute agreement model was calculated on all data for a rater-expert pair, with a rater-video interaction effect included to account for dependent observations. However, the model failed to complete for some raters due to multicollinearity. As a result, the 'Overall' column for the SOM is the average of all ICC scores for each observer, with average *p*-values calculated using Stouffer's method (Stouffer, 1949).

The following interpretation of ICC is given by Koo (2016): ICC values less than 0.5 are interpreted as "poor" agreement; values $\geq$ 0.50 and < 0.75 are considered "moderate" agreement; values $\geq$ .075 and < 0.9 are considered "good" agreement, and values of 0.9 or above are "excellent" agreement (**Table 25**). Across the seven observers, the overall ICC values were either Good (*N* = 4) or Moderate (*N* = 3) (**Table 29**).

Table 25: Interpretation of ICC Values (Koo, 2016)

Values Between	Level of Agreement
< 0.5	Poor
≥ 0.50 and < 0.75	Moderate
≥ 0.75 and < 0.90	Good
≥ 0.90	Excellent

Inter-rater reliability between each rater and expert pair on the **RIBA** (**Table 30**) was calculated through Gwet's AC_2 (linear weights), which determines the extent of agreement between two observers that is greater than expected by chance (chance-corrected agreement, CAC), and which is not subject to the biases of other CAC procedures (Gwet, 2014). As a weighted analysis, AC_2 assigns less weight to ratings that are farther apart. It is appropriate for dichotomous and categorical data. This makes it a good fit for the RIBA, on which every item is dichotomous (Not Observed = 0, Observed = 1) except for the last summary item. The Overall column was calculated by analyzing data across all videos for a given rater-expert pair because AC_2 does not assume independence of observations.

AC_2 values are interpreted by using the value's confidence interval to determine which rating band it has a $\geq$ 95% chance of falling into or above. Scores between 0.0 and 0.2 are "poor", scores from 0.2 to 0.4 are "fair", scores from 0.4 to 0.6 are "moderate", scores from 0.6 to 0.8 are "good", and scores greater than 0.8 are "very good" (Gwet, 2014) (**Table 26**). Across the seven observers, the overall ICC values were either Good (*N* = 4) or Very Good (*N* = 3) (**Table 30**).

Table 26: Interpretation of AC2 Values (Gwet, 2014)

Values Between	Level of Agreement
0.0 to 0.2	Poor
0.2 to 0.4	Fair
0.4 to 0.6	Moderate
0.6 to 0.8	Good
> 0.8	Very Good

Table 27: Inter-Rater Reliability Statistics for SOM 2021-22

Rater	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9	Video 10	Overall	Category
Rater 1	0.821***	0.845***	0.666***	0.572***	0.597***	0.552***	n/a	n/a	n/a	0.600***	0.665***	Substantial
Rater 2	0.758***	0.649***	0.465***	0.507***	0.541***	0.505***	0.611***	0.511**	0.374*	0.355**	0.528***	Moderate
Rater 3	0.621***	0.708***	0.698***	0.487**	0.549***	0.581***	0.418**	0.568***	0.665***	0.708***	0.600***	Moderate
Rater 4	0.767***	0.765***	0.411**	0.742***	0.563***	0.62***	0.477**	0.673***	0.652***	0.664***	0.633***	Substantial

*$p < .05$, ** $p < .01$, *** $p < .001$

Table 28: Inter-Rater Reliability Statistics for RIBA 2021-22

Rater	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9	Video 10	Overall	Category
Rater 1	0.522*	0.686**	-0.122	0.522*	0.522*	0.522*	n/a	n/a	n/a	0.621*	0.467***	Fair
Rater 2	0.645**	0.593**	0.676**	0.732**	0.732**	0.738***	0.313	0.756***	0.645**	0.621*	0.653***	Substantial
Rater 3	0.029	0.738***	0.441	0.732**	0.694**	0.582**	0.686**	0.409*	0.313	0.845***	0.541***	Moderate
Rater 4	0.389	0.511*	0.327	0.522*	0.732**	0.203	0.463*	0.542**	0.542**	0.441	0.467***	Fair

*$p < .05$, ** $p < .01$, *** $p < .001$

Table 29: Inter-Rater Reliability Statistics for SOM 2022-23

Rater	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9	Video 10	Overall	Category
Rater 1	0.978 ***	0.896 ***	0.665 ***	0.878 ***	0.722 ***	0.827 ***	0.870 ***	0.705 ***	0.654 ***	0.819 ***	0.801 ***	Good
Rater 2	0.805 ***	0.589 ***	0.465 **	0.659 ***	0.693 ***	0.564 **	0.564 **	0.676 ***	0.695 ***	0.858 ***	0.657 ***	Moderate
Rater 3	0.767 ***	0.758 ***	0.627 ***	0.559 **	0.803 ***	0.744 ***	0.877 ***	0.825 ***	0.778 ***	0.625 ***	0.736 ***	Moderate
Rater 4	0.761 ***	0.917 ***	0.633 ***	0.806 ***	0.733 ***	0.697 ***	0.806 ***	0.797 ***	0.834 ***	0.712 ***	0.770 ***	Good
Rater 5	0.758 ***	0.762 ***	0.602 ***	0.795 ***	0.576 ***	0.491 **	0.713 ***	0.697 ***	0.476 **	0.918 ***	0.679 ***	Moderate
Rater 6	0.679 ***	0.713 ***	0.543 **	0.752 ***	0.825 ***	0.756 ***	0.737 ***	0.843 ***	0.660 ***	0.824 ***	0.733 ***	Good
Rater 7	0.825 ***	0.839 ***	0.763 ***	0.742 ***	0.854 ***	0.798 ***	0.786 ***	0.721 ***	0.590 ***	0.685 ***	0.760 ***	Good

*$p < .05$, ** $p < .01$, *** $p < .001$

Table 30: Inter-Rater Reliability Statistics for RIBA 2022-23

Rater	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9	Video 10	Overall	Category
Rater 1	0.895 ***	0.800 ***	0.772 ***	0.769 ***	0.918 ***	0.787 ***	0.791 ***	0.880 ***	0.840 ***	0.639 *	0.887 ***	Very Good
Rater 2	0.780 ***	0.706 ***	0.869 ***	1.000 ***	0.829 ***	0.723 ***	0.840 ***	0.835 ***	0.838 ***	0.518 ***	0.877 ***	Very Good
Rater 3	0.734 ***	0.753 ***	0.816 ***	0.652 **	0.840 ***	0.614 **	0.706 **	0.800 ***	0.848 ***	0.397 ***	0.839 ***	Good
Rater 4	0.890 ***	0.769 ***	0.816 ***	0.615 *	0.681 **	0.681 ***	0.840 ***	0.747 ***	0.710 ***	0.299 ***	0.829 ***	Good
Rater 5	0.724 ***	0.790 ***	0.816 ***	0.917 ***	0.913 ***	0.729 ***	0.769 ***	0.790 ***	0.780 ***	0.323 ***	0.861 ***	Very Good
Rater 6	0.723 ***	0.792 ***	0.774 ***	0.738 ***	0.738 ***	0.771 ***	0.829 ***	0.706 ***	0.780 ***	0.455 ***	0.836 ***	Good
Rater 7	0.724 ***	0.800 ***	0.428 *	0.755 ***	0.521 *	0.681 ***	0.769 ***	0.790 ***	0.840 ***	0.323 ***	0.820 ***	Good

*$p < .05$, ** $p < .01$, *** $p < .001$

Overall Summary of Teacher Interviews and Focus Groups for Years 3 (2021-22) and 4 (2022-23)

In Spring 2022 (2021-22 school year), teacher interviews were conducted with only treatment teachers. Five out of 25 teachers initially contacted (20%) were interviewed. In May 2023 (2022-23 school year), teacher focus groups were conducted with control schools (over Zoom) and in July and August 2023 (in person) for treatment schools. Fourteen teachers were interviewed. This summary represents findings across both years from 19 teachers representing six of the seven districts and 20 out of 36 schools. Teacher responses are summarized by common topics covered in both years below. Questions were unique to the treatment group unless indicated otherwise.

Impacts of COVID 19 on Instruction

Multiple teachers in **both treatment and control groups** spoke about the pandemic's effect on the general loss of students' reading and mathematics comprehension and skills, and lack of prior science knowledge. For example, some **control group** teachers mentioned adjustments made after the pandemic, including science time being used for interventions or small pull-out groups to catch students up in reading and math.

In the **treatment group**, one teacher mentioned having a science lab that could be utilized, but the frequency of use was greatly reduced due to the pandemic. Also, during the pandemic, some teachers would have to load all science materials onto a rolling cart and move from classroom-to-classroom instead of the students coming to a single space.

Since the pandemic, multiple teachers in the **treatment group** also agreed there was an impact on increasing teacher's confidence in teaching science. Multiple teachers agreed that engaging students in inquiry-based learning or hands-on activities helped students fully comprehend the content and increased inquiries, helping retain the knowledge of the concepts. Teachers also mentioned that students' science knowledge is up-to-date, and they are grasping concepts better. All teachers interviewed found the modules were interactive, and several mentioned their students enjoyed the lessons.

The Landscape of Science Instruction

Across both years, both states, and **both treatment and control groups**, multiple teachers spoke about the state-tested nature of science in their school being a determining factor for the structure of their science teaching. They mentioned following state standards for which topics to discuss and in which order. State standards had to be addressed, and yet teachers also mentioned a wide range of time for science instruction from *not taught at all* to *15 minutes each day*.

Numerous teachers from the **treatment** groups in both states reported a heavy emphasis on reading and math, along with many teachers incorporating teacher-developed standards into their classrooms, and following state standards for which topics to discuss. Some also mentioned how science was integrated into other subject areas. For example, one teacher in **South Carolina** reported due to the lack of teachers at their school, she was teaching all subject areas, and therefore incorporated science content into other subjects. In addition, multiple teachers discussed the lack of a science curriculum at their schools.

Furthermore, multiple participants in **both treatment and control groups** reported using online assessments and materials to assess and track their students' progress. Several mentioned using online informal assessments (e.g., Study Island, BrainPOP, Generation Genius) with students to gauge comprehension of materials as well as using games and videos. One teacher in **North Carolina** reported using videos and informal assessments to fill in gaps in science knowledge. Teachers in the **control group** mentioned using online materials (e.g., Study Island) as well to teach and assess their students' progress and comprehension of concepts.

Fidelity of Implementation (Treatment group only)

In the first year following the pandemic (2021-22 school year), most teachers interviewed reported using the Smithsonian modules in the classroom. When asked how they presented the modules, teachers mentioned (a) following the instructions with fidelity, (b) utilizing the booklets that are provided with the modules, and (c) reading through the material with students while discussing it as a group. In addition, all teachers interviewed found the modules were interactive. Teachers mentioned their students enjoyed the modules, as they were able to work together as a group, brainstorm solutions, and be hands-on with science. Teachers did not suggest any improvements for the modules, and those that used the modules found them to be beneficial and engaging for their students.

Teachers also mentioned that the modules did not fit their allotted time of 30 to 45 minutes, and mentioned not implementing the modules fully due to modules not fitting into their state standards for science. Also, some teachers did not use the assessments included with the modules, although one teacher who did utilize the assessments said their students found them difficult due to the length and difficulty level of the questions.

In the second year following the pandemic (2022-23 school year), teachers in both states reported using both the Engineering and Physical Science modules to some extent. Multiple participants reported pulling specific lessons or using parts of the module that aligned with state-tested standards. However, only a few teachers reported using an entire module since they found it contained everything they needed. Teachers mentioned feeling guilty because they did not use the entire module or only used one instead of both.

Teachers also reported that implementation of the program was dependent on the teacher, grade level standards, and time available for science instruction. Many reported the science block was too short, so had difficulty fitting in all the lessons, especially as some did not align with state-tested standards. In addition, several reported using supplemental materials or other lessons, in tandem with *Smithsonian Science* modules, to teach an entire unit. Several teachers mentioned the modules did not fit their allotted time, especially with the amount of time needed to set up and prepare them for students. Moreover, teachers mentioned that fifth grade does not have as much time to implement anything that does not align to science standards, as a great deal of science instruction is used to catch students up on tested standards in preparation for testing that does not occur in other grades.

Future Smithsonian Science Implementation

Many teachers from the **control group** stated they were given little to no information about the implementation of *Smithsonian Science* at their schools. Many had mixed reactions to implementation of *Smithsonian Science* at their school, while a few were excited to welcome the program. For instance,

one teacher reported excitement for the opportunity to obtain materials that were complementary with their current curriculum, as they felt gathering materials was not their strongest skill. On the other hand, multiple respondents were apprehensive about adding new materials to their teacher-developed standards and curricula. Teachers also expressed concerns about continuing to have freedom with the way lessons are taught or continuing to have the option to "remove some things that might not work," particularly in helping students pass the Science End-of-Grade (North Carolina state assessment) for fifth graders. Teachers also mentioned not being fully informed about when the trainings would be held and when the materials would be provided. A few teachers also reported being advised that modules would be provided, but were not mandatory to use.

Although respondents in the **control group** were unsure how the implementation of *Smithsonian Science* would work at their schools, across the board, respondents stated they would be compliant in implementing *Smithsonian Science* at their schools, though teachers also mentioned they would implement *Smithsonian Science* as long as they are able to cover all their standards.

Summary

Project implementation largely occurred as planned during the 2022-23 school year (the final year of implementation), and in-person activities resumed after several years of disruptions from the COVID-19 pandemic. Third through fifth grade teachers from treatment schools received in-person content-focused professional development (PD) for the first time in summer 2023, after two sessions of virtual PD (spring and summer 2021) and one session of hybrid PD (summer 2022). Over the three years of implementation (2020-21 through 2022-23 school years), program evaluation activities and outcomes were as follows:

- **Student Achievement**:

  For the three main (i.e., confirmatory) findings on the combined sample (i.e., both states combined), the effect of *Smithsonian Science* was **positive** (g = 0.18) and **statistically significant** (p = .032) on the **Stanford-10 science assessment**. For the **state assessment** combined samples, standardized scores in **Reading** (g = 10) and **Math** (g = 16) were both **positive**, meaning students in the treatment group performed better than the control group, but the effects were **not statistically significant**.

- **Teacher Professional Development**: Across the five questions about **facilitator performance**, the intermediate (i.e., content-focused) PD responses were slightly less favorable than the introductory (curriculum-focused) PD responses. However, this effect was not consistent across all modules. For the question about the **science content** of the unit, the percentage of positive responses was, on average, stable between introductory and intermediate PD. When averaged across PD type (introductory and intermediate), responses to the **Physical Science** PD were slightly more positive than responses to the **Engineering PD**. However, these differences are relatively small, and their causal factor is unclear. Across both PDs and grade levels, responses were primarily positive for all question categories. In addition, many questions had close to 100% positive responses, but with a very small sample size.

- **Teacher Module Logs:** Across all three years, teachers consistently reported they (a) had all necessary materials, (b) felt comfortable with the science content of the modules, and (c) had

sufficient training to teach the modules. Over time, teachers became **less likely** to report that (a) they taught lessons in the suggested sequence, (b) modules were easy to use, (c) modules could comfortably fit in a class period, and (d) they taught the module during instructional time not intended for science. Meanwhile, teachers became **more likely** to report that they supplemented the lessons with materials from other sources.

- **Classroom Observations**: On the **SOM**, for **treatment and control schools** who participated in classroom observations during the 2021-22 and 2022-23 academic years, the most prevalent strategy observed both years was "Direct Instruction." Additionally, the overall classroom environments were similar between treatment and control schools. However, **treatment and control schools had a large divergence on "Experiential hands-on learning "** in the last year of the study, which was "*Extensively/Frequently observed*" over 60% more often in treatment vs. control schools (37% vs. 23% of the time, respectively). On the **RIBA**, two of the three activities most frequently observed over both years were the same for treatment and control schools: "Students engaged in experimentation" and "Students gathering or recording evidence". Meanwhile, over both years, "Prepared science kits or modules in use" was observed more frequently in treatment schools while "Students hypothesizing or making predictions" was observed more frequently in control schools. The **level of class time dedicated to inquiry-based science** was rated as "high" four times as often in control schools in the first year of the study, but over 50% higher in treatment classroom observations in the last year of the study.

- **Teacher Focus groups:** Findings from the focus groups should be interpreted with caution as they only represent 19 teachers across both years from six of the seven districts and 20 out of 36 schools, limiting the representativeness of their responses. Participants mentioned lingering impacts from COVID on the general loss of students' reading and mathematics comprehension and skills, and lack of prior science knowledge. Since the pandemic, multiple teachers in the **treatment group** also agreed there was an impact on increasing teacher's confidence in teaching science. In addition, several **treatment teachers** agreed that engaging students in inquiry-based learning or hands-on activities helped students fully comprehend the content and increased inquiries, helping retain the knowledge of the concepts. Several teachers in both states and both years spoke about the state-tested nature of science in their school being a determining factor for the structure of their science teaching. Teachers in both years also mentioned issues with alignment between the models and their state standards and having difficulty fitting the module in the limited time allotted for science instruction.

Next Steps

In the current year (2023-24), CREP's evaluation activities include:

- Collection of teacher Module Logs
- Treatment school focus groups
- Surveying participants of curriculum professional development in summer 2024

In sum, despite some lingering impacts from COVID-19 on the originally planned implementation schedule and methods, *Smithsonian Science for North and South Carolina Classrooms* returned to full in-person implementation during the 2022-23 school year, received positive feedback

from teachers, and **demonstrated statistically significant positive impacts on student achievement in science** after three years of implementation.

Appendix A: Fidelity Matrix

Table 31: Fidelity Matrix for Implementation Year 1

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
Key Component 1= Professional Development (Implementation Year 1)											
Engineering Curriculum-Tied PD (Spring 2021)	12 hours	Teacher	Attendance records collected at training by workshop facilitator	By 03/31, attendance records delivered to evaluator by email	**0 (low)** = < 50% of total hours **1 (moderate)** = 50%-79% of total hours **2 (high)** = 80% of total hours	Adequate implementation at teacher level = score of "2"					
All indicators	NA	NA	NA	NA	0 - 2	Teacher-level: Adequate implementa-tion score = 2	School-level **0** = < 25% teachers with score of "2" **1** = 26-50% teachers with score of "2" **2** = 51-75% of teachers with score of "2" **3** = > 75% teachers with score of "2" **Threshold for fidelity = score of "3"**	NA	Sample-level **0** = < 25% schools with score = 3 **1** = 26-50% schools with score = 3 **2** = 51-75% schools with score = "3" **3** = > 75% schools with score = 3 **Threshold for fidelity = score of "3"** [*in more than 75% of schools, the majority of*	All schools in which intervention is being implemented (*n* = 18 schools)	2020-21 (1st year of implementa-tion)

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
									teachers have high implementation of PD]		

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
Key Component 2= Access to Materials (Implementation Year 1)											
Engineering Curricular Modules Shipped	Curricular modules shipped by target date by Carolina Biological (Sufficient number of modules for each school to serve all students grades 3-5)	Sample	Carolina Biological	Before the last session of virtual Professional Development	**0 (low)** = < 80% of ordered modules shipped on time **1 (moderate)** = 80%-89% of ordered modules shipped on time **2 (high)** = 90% of ordered modules shipped on time	Adequate implementation at sample level = score of "2"					
All indicators	NA	NA	NA	NA	0 - 2	Sample-level	NA	NA	Sample-level	All schools in which intervention is	2020-21 (1st year of

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
						Adequate implementation score = 2			Adequate implementation score = 2 **Threshold for fidelity = score of "2"** [*90% or more of ordered modules shipped by target date]*	being implemented (*n* = 18 schools)	implementa-tion)

Table 32: Fidelity Matrix for Implementation Year 2

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
Key Component 1= Professional Development (Implementation Year 2)											
Engineering Content-Tied PD (Summer 2021)	12 hours	Teacher	Attendance records collected at training by institute facilitator	By 09/30, attendance records delivered to evaluator by email	**0 (low)** =< 50% of total hours **1 (moderate)** = 50%-79% of total hours **2 (high)** = 80% of total hours	Adequate implementation at teacher level = score of "2"					
All indicators	NA	NA	NA	NA	0 - 2	Teacher-level: Adequate implementation score = 2	School-level **0** = < 25% teachers with score of "2" **1** = 26-50% teachers with score of "2" **2** = 51-75% of teachers with score of "2" **3** = > 75% teachers with score of "2" **Threshold for fidelity = score of "2"**	NA	Sample-level **0** = < 25% schools with score = 3 **1** = 26-50% schools with score = 3 **2** = 51-75% schools with score = "3" **3** = > 75% schools with score = 3 **Threshold for fidelity = score of "3"** [*in more than 75% of schools, the majority of teachers have*	All schools in which intervention is being implemented (*n* = 18 schools)	2021-22 (2nd year of implementa-tion)

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
									high implementation of PD]		

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
Key Component 2= Access to Materials (Implementation Year 2)											
Additional Engineering Curricular Modules Shipped	Any additional curricular modules needed shipped by target date by Carolina Biological (If increases in student enrollment occur)	Sample	Carolina Biological	By October 1 of the fall semester following summer PD	**0 (low)** = < 80% of additional ordered modules shipped on time **1 (moderate)** = 80%-89% of additional ordered modules shipped on time **2 (high)** = 90% of additional ordered modules shipped on time or no additional modules needed.	Adequate implementation at sample level = score of "2"					

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
All indicators	NA	NA	NA	NA	0 - 2	Sample-level Adequate implementation score = 2	NA	NA	Sample-level Adequate implementation score = 2 **Threshold for fidelity = score of "2"** [*90% or more of ordered modules shipped by target date]*	All schools in which intervention is being implemented (*n* = 18 schools)	2021-22 (2nd year of implementa-tion)

Table 33: Fidelity Matrix for Implementation Year 3

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
Key Component 1= Professional Development (Implementation Year 3)											
Science Curriculum-Tied PD (Summer 2022)	12 hours	Teacher	Attendance records collected at training by institute facilitator	By 09/30, attendance records delivered to evaluator by email	**0 (low)** =< 50% of total hours **1 (moderate)** = 50%-79% of total hours **2 (high)** = 80% of total hours	Adequate implementation at teacher level = score of "2"					
All indicators	NA	NA	NA	NA	0 - 2	Teacher-level: Adequate implementation score = 2	School-level **0** = < 25% teachers with score of "2" **1** = 26-50% teachers with score of "2" **2** = 51-75% of teachers with score of "2" **3** = > 75% teachers with score of "2" **Threshold for fidelity = score of "2"**	NA	Sample-level **0** = < 25% schools with score = 3 **1** = 26-50% schools with score = 3 **2** = 51-75% schools with score = "3" **3** = > 75% schools with score = 3 **Threshold for fidelity = score of "3"**	All schools in which intervention is being implemented (n = 18schools)	2022-23 (3rd year of implementation)

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
									[in more than 75% of schools, the majority of teachers have high implementation of PD]		

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
Key Component 2= Access to Materials (Implementation Year 3)											
Science Curricular Modules Shipped and Additional Engineering Curricular Modules Shipped	Curricular modules shipped by target date by Carolina Biological (Sufficient number of Science curriculum modules for each school to serve	Sample	Carolina Biological	By October 1 of the fall semester following summer PD	**0 (low)** = < 80% of ordered modules shipped on time **1 (moderate)** = 80%-89% of ordered modules shipped on time **2 (high)** = 90% of ordered	Adequate implementation at sample level = score of "2"					

Indicators	Definition	Unit of implem-entation	Data Source(s)	Data Collection (who, when)	Score for levels of implementation at unit level	Threshold for adequate implementation at unit level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to next higher level if needed (score and threshold): Indicate level	Roll-up to program level (score and threshold for adequate implementation at sample level)	Expected sample for fidelity measure	Expected years of fidelity measurement
	students grades 3-5, and additional Engineering modules if increases in student enrollment occur.)				modules shipped on time.						
All indicators	NA	NA	NA	NA	0 - 2	Sample-level Adequate implementation score = 2	NA	NA	Sample-level Adequate implementation score = 2 **Threshold for fidelity = score of "2"** [90% or more of ordered modules shipped by target date]	All schools in which intervention is being implemented (n = 18 schools)	2022-23 (3rd year of implementation)

Appendix B: Technical Details of the Stanford-10 Analysis

Covariates

Level-1 (student) covariates obtained from spring 2021 state assessment data included gender, BIPOC status (non-White), economic disadvantage, English language learner status, and special educational status. Other level-1 covariates were the Stanford-10 pretests for Reading Comprehension and Mathematics Problem Solving.

Level-2 (school) covariates were obtained from the 2022-2023 Common Core of Data (National Center for Educational Statistics, 2023). They included the percentage of White students in a school, the percentage of free lunch-eligible students in a school, schoolwide Title I status, the student-to-teacher ratio, and school locale (such as urban, suburban, and rural). Additionally, randomization block (i.e., blocks used to randomly assign schools to treatment or control) was used as a level-2 covariate.

Missing Data

Students with missing outcome data were dropped from analysis. Missing covariate data were imputed through dummy variable imputation. For missing continuous covariate data, the cluster mean was used as the imputation value. For missing categorical covariate data, a value was set for the variable to indicate missingness. Since all level-1 categorical data for each state came from a single source, all level-1 categorical variables had the same missingness pattern. In other words, if one of the variables was missing for a given student, all were missing. Therefore, only one variable to mark missingness was used for all categorical level-1 covariates (DemoMISS). No data were missing at level-2.

Model Specifications

Baseline Model

Level-1: Individual Level

$$BaseVar_{ij} = \beta_{0j} + \varepsilon_{ij}$$

Level-2: Cluster Level

$$B_{0j} = \gamma_{00} + \gamma_{01}(T_j) + \sum_{p=1}^{P-1} \gamma_{02.p} Block_{pj} + \mu_{0j}$$

Where,

$BaseVar_{ij}$ = the baseline measure for the ith student in the jth cluster. For the Science outcome, this would be the SAT-10 Reading and Math Problem Solving subtests.

β_{0j} = the intercept for cluster j.

γ_{00} = the unadjusted mean baseline value for clusters in the comparison group and in the reference block.

γ_{01} = the difference in means between clusters in the treatment and comparison groups.

T_j = 1 if cluster *j* is assigned to treatment, and = 0 if assigned to comparison.

$\gamma_{02.p}$ = the effect of block (i.e., the difference in the intercept between block p and the reference block).

$Block_{pj}$ =1 if cluster *j* is in block *p* (*p*=1, 2, …, *P*), otherwise = 0. Blocking would be done by district, and potentially, achievement strata (e.g., Low, Medium, High) within district.

μ_{0j} = random intercept term for cluster *j*.

ε_{ij} = a residual error term for individual *i* in cluster *j*.

Outcome Model

Level-1: Individual Level (Student):

$$Science_{ij} = \beta_{0j} + \beta_{1j}Gender_{ij} + \beta_{2j}BIPOC + \beta_{3j}ED_{ij} + \beta_{4j}ELL_{ij} + \beta_{5j}SPED_{ij} + \beta_{6j}DemoMISS_{ij} + \beta_{7j}ReadComp_{ij} + \beta_{8j}ReadCompMISS_{ij} + \beta_{9j}MathPS_{ij} + \beta_{10j}MathPSMISS_{ij} + r_{ij}$$

Level-2: Cluster Level (School):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}pctWhite_j + \gamma_{02}pctFL_j + \gamma_{03}schTitleI_j + \gamma_{04}schLocale_j + \gamma_{05}Block_{pj} + \gamma_{06}Treatment_j + u_{0j}$$

$\beta_{1j} = \gamma_{10}$

$\beta_{2j} = \gamma_{20}$

$\beta_{3j} = \gamma_{30}$

$\beta_{4j} = \gamma_{40}$

$\beta_{5j} = \gamma_{50}$

$\beta_{6j} = \gamma_{60}$

$\beta_{7j} = \gamma_{70}$

$\beta_{8j} = \gamma_{80}$

$\beta_{9j} = \gamma_{90}$

$\beta_{10j} = \gamma_{100}$

Mixed Model:

$$Science_{ij} = \gamma_{00} + \gamma_{01}pctWhite_j + \gamma_{02}pctFL_j + \gamma_{04}schTitleI_j + \gamma_{05}schLocale_j + \gamma_{06}Block_{pj} + \gamma_{07}Treatment_j + \gamma_{10}Gender_{ij} + \gamma_{20}Minority_{ij} + \gamma_{30}ED_{ij} + \gamma_{40}ELL_{ij} + \gamma_{50}SPED_{ij} + \gamma_{60}DemoMISS_{ij} + \gamma_{70}ReadComp_{ij} + \gamma_{80}ReadCompMISS_{ij} + \gamma_{90}MathPS_{ij} + \gamma_{100}MathPSMISS_{ij} + u_{0j} + r_{ij}$$

Where,

$Science_{ij}$ = the SAT-10 Science outcome for the *i*th student in the *j*th school.

β_{0j} = the intercept for school *j*.

$\beta_{1j} - \beta_{10j}$ = the effect of covariates in school *j*.

r_{ij} = a residual error term for student *i* in school *j*.

γ_{00} = the mean intercept.

$\gamma_{01} - \gamma_{06}$ = the effect of school-level covariates.

$Block_{pj}$ = 1 if the school *j* was assigned to the treatment or comparison condition within the (randomization or matching) block *p*, and = 0 otherwise.

γ_{07} = the effect of treatment.

$Treatment_j$ = 1 if school *j* is assigned to treatment, and = 0 if school *j* is assigned to comparison.

μ_{0j} = random intercept term – deviation of block *j*'s mean from the grand mean, conditional on covariates; assumed to be normally distributed with mean 0 and variance τ_{00}^2.

Analysis

The main Stanford-10 analysis results are reported in the Stanford-10 Outcomes section. Data were analyzed in an HLM using restricted maximum likelihood, with school as a random effect. The analysis was conducted in R 4.3.1 (R Core Team, 2023) using the packages lme4 v. 1.1.34 (Bates et al., 2015) and lmerTest v. 3.1.3 (Kuznetsova et al., 2017). Estimated marginal means were calculated through the ggeffects package, v. 1.2.3 (Lüdecke, 2018). The output from the analysis is provided below.

Output

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula: SciSS ~ fem + min + ell2 + ed2 + sped2 + demo_miss + ReadCompSS +

MathPSSS + ReadComp_MISSING + MathPS_MISSING + (1 | school) +

block + schLocale + pctfl23 + pctWhite23 + str23 + schTitleI + treatment

Data: sat1023

REML criterion at convergence: 16186.3

Scaled residuals:

Min	1Q	Median	3Q	Max
-5.7070	-0.6063	-0.0141	0.6163	3.9739

Random effects:

Groups	Name	Variance	Std.Dev.
school	(Intercept)	23.44	4.841
Residual		636.99	25.239

Number of obs: 1752, groups: school, 36

Fixed effects:

	Estimate	Std. Err	df	t value	Pr(>|t|)
(Intercept)	483.6147	26.61123	29.09623	18.173	2E-16 ***
fem	-3.25976	1.40193	1715.532	-2.325	0.02018 *
min	-4.38152	1.86266	1716.196	-2.352	0.01877 *
ell2	-5.84914	3.43912	1711.751	-1.701	0.08917 .
ed2	-3.23072	1.63239	1717.262	-1.979	0.04796 *
sped2	-8.47643	2.14633	1717.199	-3.949	0.0000816 ***
demo_miss	-7.64055	2.62756	1681.066	-2.908	0.00369 **
ReadCompSS	0.17663	0.02154	1690.18	8.201	4.86E-16 ***
MathPSSS	0.16712	0.0234	1706.603	7.142	1.35E-12 ***
ReadComp_MISSING	-3.84645	2.51717	1719.858	-1.528	0.12668
MathPS_MISSING	2.49245	2.56675	1711.287	0.971	0.33166
block121	4.27256	5.53577	18.2052	0.772	0.45013
block122	9.06666	5.52838	16.99894	1.64	0.11937

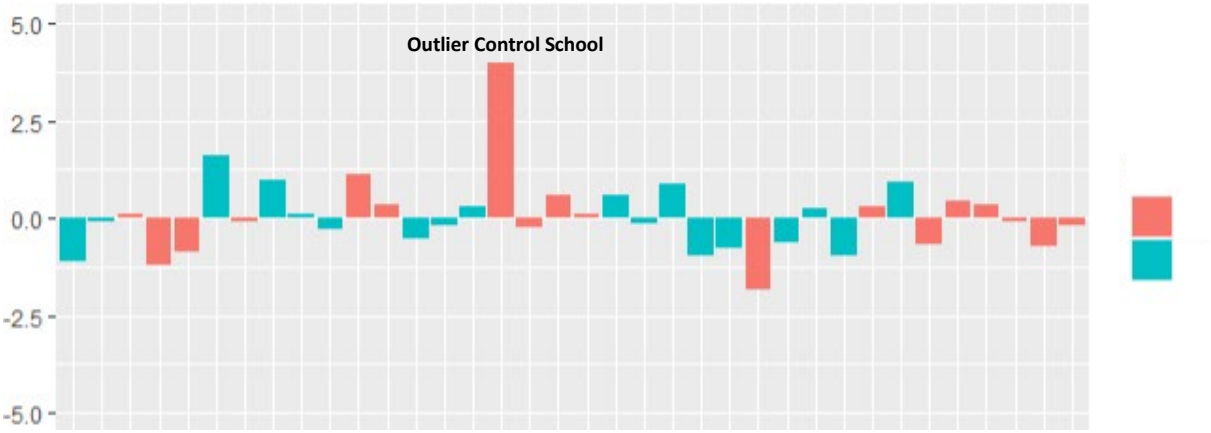
block131	-10.8304	5.43351	17.42133	-1.993	0.06213 .
block132	-13.1957	6.02718	20.47445	-2.189	0.04031 *
block141	10.34509	7.38139	18.2632	1.402	0.17783
block211	10.83747	7.79281	18.3167	1.391	0.18098
block221	13.84736	8.75596	16.77077	1.581	0.13244
block231	16.49938	9.45482	19.02375	1.745	0.0971 .
block232	11.14652	9.80827	18.90744	1.136	0.26997
schLocale22-Suburb: Mid-size	-7.29193	5.69314	15.84567	-1.281	0.21868
schLocale32-Town: Distant	-7.02088	7.21691	17.76442	-0.973	0.3437
schLocale41-Rural: Fringe	-7.54799	5.92019	17.48558	-1.275	0.21901
schLocale42-Rural: Distant	-18.6781	6.83613	18.83401	-2.732	0.0133 *
schLocale43-Rural: Remote	-3.86751	9.67946	22.61813	-0.4	0.69323
pctfl23	-35.0535	17.30834	17.38434	-2.025	0.05848 .
pctWhite23	15.13028	10.29516	19.99846	1.47	0.15721
str23	-1.29473	0.75689	16.08767	-1.711	0.10637
schTitleI	-2.90813	6.80575	15.21519	-0.427	0.67514
treatment	5.58619	2.40427	18.60311	2.323	0.03166 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Outlier School Analysis

One control school scored very high for its demographic profile. As shown in **Figure 8**, the school's level-2 standardized residual (e.g., average prediction error for the whole school) was 4.0, meaning its scores were extremely high when compared to demographically similar schools. This was one of the smallest schools in the study, with only 15 students. However, to a certain extent, groups are treated equally in HLM regardless of sample size (i.e., despite this school representing only 0.8% of the students, it represented 2.8% of schools).

Figure 8: Level-2 Standardized Residuals



Because of (a) the way groups are treated in HLM, and (b) this school's extremely high average score, it had a larger amount of influence over the estimate of treatment effect: Despite being the smallest school in the study, it had approximately four times the influence of an average school, or potentially a disproportional impact on the treatment estimate.

To investigate this possibility, CREP ran a secondary analysis with this school removed. The results are in **Table 34** below. As before, the results were positive and statistically significant, but at a much larger level: Statistical significance increased from $p = .043$ to $p = .010$, and the estimated effect size increased from $g = 0.17$ to $g = 0.22$. In other words, the main analysis may be an *underestimate* of the treatment effect. However, the entire school could not be removed for the main analysis, as that could call the evaluation results into question.

Table 34: Secondary Stanford-10 Analysis Results

HLM Coefficient	95% CI	*p* value	Hedges' *g*	Percentile Rank	Imp. Index
6.82	3.36, 9.55	.010	0.22	59	9

References

Alvarez-Rivero, A., Odgers, C., & Ansari, D. (2023). Elementary school teachers' perspectives about learning during the COVID-19 pandemic. *NGP Science of Learning, 40*. https://doi.org/10.038/s41539-023-00191-w

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational researcher*, 25(9), 6-14.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01.

Blank, R.K., & de las Alas, N. (2009). *Effects of teacher professional development on gains in student achievement: How meta analysis provides scientific evidence useful to education leaders*. Council of Chief State School Officers. http://files.eric.ed.gov/fulltext/ED544700.pdf+

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. Journal of Research on Educational Effectiveness, 1(4), 289-328. https://files.eric.ed.gov/fulltext/ED503202.pdf

Brown, S., & Browning, S. (2021). Elementary Students' STEM Investigations. *Elementary STEM Journal*, *26*(2), 15–20. Retrieved from https://eds.s.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=3bd4f6c0-d2f5-468b-ab6c-fbef227ba8c6%40redis

Cheung, A., Slavin, R. E., Kim, E., & Lake, C. (2017). Effective secondary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, 54(1), 58-81.

Crawford, B. A. (2014). From inquiry to scientific practices in the science classroom. In N. G. Lederman & S. K. Abell (Eds.), Handbook of research on science education, volume II (pp. 515– 541). Routledge.

Duran, M., & Dökme, I. (2016). The effect of the inquiry-based learning approach on student's critical-thinking skills. Eurasia Journal of Mathematics Science and Technology Education, 12(12).

Fletcher-Wood, H., & Zuccollo, J. (2020). The effects of high-quality professional development on teachers and students: A rapid review and meta-analysis. *Education Policy Institute*.

Gonzalez, Kathryn, Kathleen Lynch, and Heather C. Hill. (2022). A Meta-Analysis of the Experimental Evidence Linking STEM Classroom Interventions to Teacher Knowledge, Classroom Instruction, and Student Achievement. (EdWorkingPaper: 22-515). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/d9kc-4264

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Herb, N. E. (2022). *Elementary Teachers' Experiences With Professional Development in Inquiry-Based Science Education* (Doctoral dissertation, Shippensburg University).

Kennedy, M. M. (2016). How does professional development improve teaching?. Review of educational research, 86(4), 945-980.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Krajcik, J., Schneider, B., Miller, E. A., Chen, I. C., Bradford, L., Baker, Q., ... & Peek-Brown, D. (2023). Assessing the effect of project-based learning on science learning in elementary schools. *American Educational Research Journal*, *60*(1), 70-102.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software, 82(13), 1-26. doi:10.18637/jss.v082.i13.

Lee, C. A. & Houseal, A. (2003). Self-efficacy, standards, and benchmarks as factors in teaching elementary school science. Journal of Elementary Science Education, 15(1), 37-56. https://doi.org/10.1007/BF03174743

Lewis, E., Ross, S. M., & Alberg, M. (1999). *Reliability Analysis for School Observation Measure*. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Lo, C. K. (2021). Design Principles for Effective Teacher Professional Development in Integrated STEM Education: A Systematic Review. Educational Technology & Society, 24 (4), 136–152. Retrieved from https://eds.s.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=7b185f8e-9209-4d1e-8b7e-adb523d67a55%40redis

Lüdecke D (2018). "ggeffects: Tidy Data Frames of Marginal Effects from Regression Models." *Journal of Open Source Software*, *3*(26), 772. doi:10.21105/joss.00772.

Lynch, K., Hill, H.C., Gonzalez, K.E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis, 41*(3), 260–293. https://doi.org/10.3102/01623737198490443

Moreno, N. P., Garay, D. V., Harris, K. A., Newell, A. D., Perez-Sweeney, B., Camacho-Lopez, E., & Shargey, B. A. (2021). What the Pandemic Experience Taught Us about STEM Higher Education-School Partnerships. *Journal of STEM Outreach*, *4*(2). Retrieved from https://eric.ed.gov/?id=EJ1311065

National Center for Educational Statistics. (2023). *Common core of data* [Data set]. https://nces.ed.gov/ccd/

National Science Foundation. (2022). *The state of U.S. science & engineering: 2022 science & engineering indicators* (National Science Board Publication NSB-2022-1). https://ncses.nsf.gov/pubs/nsb20221#

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Ross, S. M., Smith, L. J., & Alberg, M. (1998). *School Observation Measure (SOM©) [*Manual and training videos]. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Ross, S. M., Smith, L. J., Alberg, M., & Lowther, D.L. (2004). Using Classroom Observation as a Research and Formative Evaluation Tool in Educational Reform: The School Observation Measure in H. Waxman (Ed.). *Observation Research in U.S. Classrooms: New Approaches for Understanding Cultural and Linguistic Diversity* (pp. 144-173). Cambridge, MA: Cambridge University Press.

Slavin, R.E., Lake, C., Hanley, P., & Thurston. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching, 51*(7), 870-901. https://doi.org/10.1002/tea.21139

Songer, N. B., Lee. H.-S., & Kam, R. (2001). Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? *Journal of Research in Science Training, 39(2)*, 128-150. https://doi.org/10.1002/tea.10013

Pearson Education. (2018). Stanford Achievement Test Series, Tenth Edition. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Academic-Learning/Comprehensive/Stanford-Achievement-Test-Series-%7C-Tenth-Edition/p/100000415.html?tab=overview

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.

Sterbinsky, A., & Ross, S. M., (2003). *School Observation Measure Reliability Stud*y. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

Strat, T. T. S., Henriksen, E. K., & Jegstad, K. M. (2023). Inquiry-based science education in science teacher education: a systematic review. Studies in Science Education, 1-59.

van Uum, M.S.J., Verhoeff, R.P, & Peeters, M. (2016). Inquiry-based science education: towards a pedagogical framework for primary school teachers. *International Journal of Science Education, 38*(16), 450-469. https://doi.org/10.1080/09500693.2016.1147660

What Works Clearinghouse. (2019). *Review protocol for primary science, version 4.0 (March 2019)*. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_ps_protocol_v4.0_508.pdf

What Works Clearinghouse. (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 (NCEE). https://ies.ed.gov/ncee/wwc/Handbooks

World Health Organization (2020). Retrieved from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020Zuur, A. F., Leno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, *1*(1), 3-14.

Yoon, H. G., Joung, Y. J., & Kim, M. (2012). The challenges of science inquiry teaching for pre-service teachers in elementary schools: Difficulties on and under the scene. *Research in Science Education*, 42, 589-608. https://doi.org/10.1007/s11165-011-9212-y

Zinger, D., Sandholtz, J. H., & Ringstaff, C. (2020). Teaching science in rural elementary schools: Affordances and constraints in the age of NGSS. *Rural Educator*, *41*(2), 14-30.

The University of Memphis
College of Education
Center for Research in Educational Policy
610 Goodman Street, 201 Newport Hall
Memphis, TN 38152