Afterschool & STEM

System-Building Evaluation 2016

Patricia J. Allen, Ph.D.

Gil G. Noam, Ed.D., Ph.D. (Habil.)

The PEAR Institute: Partnerships in Education and Resilience

Harvard Medical School and McLean Hospital

Todd D. Little, Ph.D.

Eriko Fukuda, Ph.D.

Rong Chang, Ph.D.

Britt K. Gorrall, M.Ed.

Luke Waggenspack, B.A.

IMMAP: Institute for Measurement, Methodology, Analysis & Policy

Texas Tech University

AFTERSCHOOL & STEM SYSTEM-BUILDING EVALUATION 2016

Abstract

WHY WE CONDUCTED THIS EVALUATION

As the nation seeks ways to increase interest in science, technology, engineering, and math (STEM) education and careers, high-quality afterschool STEM programs will fill a growing need. With support from the Charles Stewart Mott Foundation and the Noyce Foundation (now STEM Next), states across the country are developing systems of support for more quality afterschool programs focused on STEM. System building elements include partnership and leadership development; evaluation and data collection activities; quality building and professional development opportunities; communication and policy; and financing and sustainability.

This evaluation is among the first at a large scale to measure the impact of afterschool programs on students' STEM-related attitudes and social-emotional/21st-century skills. The primary goals of this work were (1) to examine levels of change in youth outcomes among programs receiving resources and training support from system-building states; (2) to inform on national trends related to STEM learning, such as gender or grade differences in science interest; and (3) to link STEM program quality with student outcomes and facilitator beliefs.

WHAT WE FOUND

Participation in STEM-focused afterschool programs led to major, positive changes in students' attitudes toward science. More than 70% of students reported positive gains in areas such as *STEM interest*, *STEM identity*, *STEM career interest* and career *knowledge*, and *21st-century skills*, including *perseverance* and *critical thinking*. Female students were more likely to report gains in *relationships with adults and peers* in numbers significantly higher than their male counterparts. Larger positive effects were also noted in students who participated in their programs for a minimum of four weeks.

There was a quality-related effect on student outcomes, such that students participating in higher quality STEM programs reported more positive gains than students participating in lower quality STEM programs. There were clear variations in outcomes across states.

WHAT WE RECOMMEND

The Afterschool & STEM System-Building Evaluation serves as a proof point that it is possible to gather evidence of STEM learning in afterschool using common data-creating tools on a national scale. Recommendations include:

1. **Leverage leaders' strengths:** Support the growing community of system-builders in their efforts to address key system components: partnership and leadership development, quality building and professional development opportunities, communication and policy, and evaluation and data collection.
2. **Target professional development:** Provide professional development for facilitators and quality support in the areas of programming ideas, program management, and how to connect afterschool programming with the school day. Additional support would be helpful to improve STEM content learning, inquiry, reflection, relevance, and youth voice in the implementation of STEM activities.
3. **Focus on the linkage between STEM learning and 21st-century skills:** Integrate youth development and informal science in programs to simultaneously address the 21st-century needs of students while also sparking their curiosity and skills in science.
4. **Encourage use of data to inform practice:** Gather survey and observation data from programs to continuously improve. Encourage programs to work together to collectively pool data that will identify strengths and challenges on a city and state level to inform on the best ways to leverage training, resources, and support.
5. **Innovate out-of-school time evaluation and assessment strategies:** Consider innovative methods like the retrospective pretest-posttest format to gain a better understanding of outcomes than the traditional methods make possible.
6. **Prioritize evaluation in the system-building process:** Dedicate resources and build infrastructure in states around evaluation and assessment to track successes and challenges in afterschool STEM programming.

ABOUT THE METHODOLOGY

Nearly 1,600 students (Grades 4–12) enrolled in 160 afterschool STEM programs across 11 states completed a retrospective self-report survey called the Common Instrument Suite (CIS), which measures STEM-related attitudes and 21st-century skills. STEM facilitators completed a survey about their experiences leading afterschool STEM, and the programs' STEM activities were observed by professionals certified to use the Dimensions of Success (DoS) tool to establish levels of quality.

About the Research Team

This evaluation was developed by a large-scale collaboration between researchers, practitioners, funders and 11 statewide afterschool networks (FL, IA, IN, KS, MA, MD, MI, NE, OR, PA, SC). The evaluation was a collaboration between The PEAR Institute: Partnerships in Education and Resilience at Harvard University and McLean Hospital led by Gil Noam, Ed.D., Ph.D. (Habil.) and Patricia J. Allen, Ph.D. and IMMAP: Institute for Measurement, Methodology, Analysis & Policy at Texas Tech University led by Todd D. Little, Ph.D., with Eriko Fukuda, Ph.D., Rong Chang, Ph.D., Britt K. Gorrall, M.Ed., and Luke Waggenspack, B.A.

Acknowledgments

We would like to thank Ron Ottinger of STEM Next and Victoria Wegener of Mainspring Consulting for their continuous support throughout this project. We thank Ashima Shah, Ph.D., and Rebecca Browne, B.S., at The PEAR Institute for leading trainings and providing assistance related to the Dimensions of Success (DoS) observation tool. In addition, we thank the network leads of the 11 states, their staff, and all of the 160 programs, their facilitators and youth. We could not have done this work without everyone's active participation.

Special thanks to the Charles Stewart Mott Foundation for leadership in the afterschool field and support for this research.

TABLE OF CONTENTS

INTRODUCTION

The Charles Stewart Mott Foundation and the Noyce Foundation, through the Mott-Noyce STEM Initiative, have embarked on a nationwide capacity-building project that aspires to improve the quality, quantity, and accessibility of science, technology, engineering, and math (STEM) offerings to young people in afterschool across the United States. As of 2016, all 50 states have either statewide afterschool network or partnership grants, and over half have received either STEM system-building or planning grants. Significant effort and resources have been invested by the foundations and state afterschool networks to support informal STEM learning by building capacity, providing tools and trainings, creating communities of practice and sharing system-building strategies and advice.

States receiving system-building support engage key partners around a vision of quality STEM in afterschool; map the existing landscape of afterschool and STEM efforts; prioritize strategies and act to expand awareness of, supply, and quality of STEM in afterschool through communication, policy, and professional development; and measure the effectiveness of efforts. This major investment, involving large numbers of state afterschool networks and organizations that reach many committed staff and students, deserves an evaluation aimed to answer the question of whether afterschool STEM providers are learning how to advance the cause of STEM for children and youth in significant areas like STEM interest, engagement, skills, and motivation. To this end, The PEAR Institute: Partnerships in Education and Resilience at Harvard University and McLean Hospital, in partnership with IMMAP: Institute for Measurement, Methodology, Analysis, & Policy at Texas Tech University, devised an innovative plan to evaluate youth outcomes and quality of STEM activities in afterschool programs receiving resources and training support from Mott-Noyce system-building states. As detailed in this report, data collected using multiple methods substantiate the increase of quality afterschool STEM related to improved STEM learning.

The Evolution of Afterschool STEM

The state of the afterschool STEM field is rapidly evolving (e.g., Noam & Shah, 2013). Afterschool programs were originally conceived as safe, engaging, and enriching places for youth to participate in a variety of hands-on activities and avoid the dangers of unsupervised time while parents are still at work. The expectations had been that students would receive mentoring, homework help, and access to sports, games, or arts and crafts. Now, however, afterschool is increasingly being conceptualized as a place to complement and supplement learning from the school day. Since 2009, when STEM education was identified as a national priority for the coming decade, significant emphasis has been placed on teaching STEM inside and outside of school. There have been multiple influential collaborations in both the public and private sectors to ensure young people are motivated and inspired to excel in science and math. As a result, the role of afterschool is now shifting rapidly to incorporate access to science learning opportunities. Afterschool settings are considered ideally situated to foster student interest and engagement in STEM, in part because they can offer more hands-on and exciting activities than those typically provided in regular school settings. However, the demand for STEM-focused afterschool programming has outpaced professional development and the confidence of afterschool educators to teach STEM. This dramatic shift in educational priorities has placed added pressures on afterschool programs to provide quality STEM experiences, whether they are prepared to or not (Noam & Shah, 2013).

Quality, Quantity, and Outcomes

The Mott-Noyce STEM Initiative has focused on improving the quality and quantity of STEM offerings with significant training, resources, and support to improve the skills of a large community of practitioners leading STEM activities. This proactive approach to improving quality is a game changer for the afterschool STEM field. Many have rushed to measure outcomes (to prove that afterschool STEM is effective) among programs that are not yet properly equipped to teach informal STEM well. The Mott-Noyce STEM Initiative has worked to develop statewide systems to support STEM in afterschool by providing a process framework, concrete strategies, examples, and tools to inform the work of state afterschool networks and partners. We hypothesize that good outcomes can be achieved with adequate training, resources, technical assistance, infrastructure, and commitment.

Moreover, availability and expansion of afterschool STEM offerings are key issues as well. It is critical to know whether all communities are being reached (for instance, children living in rural, urban, and suburban settings) and even more importantly, whether groups traditionally underrepresented in STEM (including minority groups and women) are being served by afterschool programs. We hypothesize that the more success programs have, the more financial support they will receive from stakeholders (e.g., funders, businesses) to expand and increase opportunities for children and youth.

Evaluating STEM Learning

The PEAR Institute has been involved with STEM activities and assessment for many years. Our interest in promoting social-emotional well-being in children in-school and out-of-school, as well as our experience in developing assessment tools, led naturally to involvement with the movement to ensure that children have positive, high-quality experiences when they participate in afterschool STEM activities. To better understand the Mott-Noyce Initiative related to students, staff, and organizations, PEAR and IMMAP designed, coordinated, and executed the 2016 Afterschool & STEM System-Building Evaluation to test the relationship between STEM program quality and student outcomes. This national effort is at the cutting edge of the research on STEM learning in the afterschool field. With the help of funders, state network leaders, program directors, and STEM educators, our cross-state research team gathered three pieces of evidence of STEM learning from 160 afterschool STEM programs across the United States. Namely, data were collected using tools developed by The PEAR Institute that measure program quality, facilitator experience, and youth outcomes in STEM and 21st-century skills.

The specific questions guiding this evaluation were as follows:

Funder and state afterschool network support

- How has the support provided by funders and state networks impacted STEM practices and 21st-century skills among youth across the United States?

Student similarities and differences

- How are student characteristics, such as gender, grade level, and academic performance, related to student outcomes?

Program similarities and differences

- How are program characteristics, such as facilitator beliefs, program duration, and quality of STEM activities, related to student outcomes?

Converging evidence of STEM learning

- How is STEM program quality related to student outcomes and facilitator beliefs?

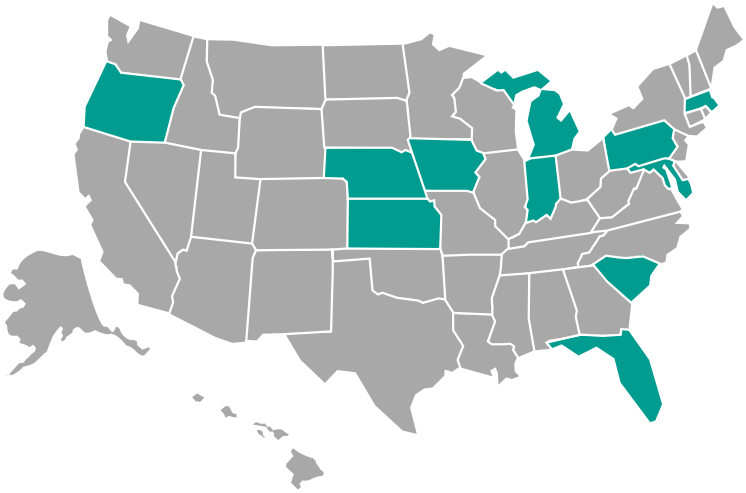
In summary, the primary goals of this evaluation were (1) to examine levels of change in STEM-related outcomes and 21st-century skills among youth in programs receiving resources and training support from system-building states; (2) to inform on national trends related to STEM learning, such as gender or grade differences in STEM interest; and (3) to link program quality with student outcomes and facilitator beliefs.

METHODS

Participants

A total of 11 state afterschool networks were chosen to participate in this evaluation based on *a priori* criteria agreed upon by funders and PEAR/IMMAP: (1) the collection of participating states together reflect the demographic diversity of the U.S., including rural, suburban, and urban composition; (2) the state afterschool networks receive system-building support from the two funders, and (3) the state afterschool networks demonstrate prior experience and capacity to implement a large-scale and complex evaluation within the designated evaluation time frame (March–June 2016). An expert demographer at Texas Tech University served as consultant to inform on the choice of states to ensure the representativeness of the sample.

Figure 1. Map of Participating State Afterschool Networks



Leaders from each of the 11 state networks (see Figure 1) consulted with PEAR/IMMAP to choose up to 15 informal STEM education programs that best represent the afterschool universe in their state, ensuring a variety of curricular offerings that are taught in different settings (school-based, community-based, or other), that range in level of formality, and that represent different demographics including age and race/ethnicity. PEAR/IMMAP provided program selection guidelines to state networks to assist in the recruitment process (see Table 1). The target demographic for this evaluation was youth in Grades 4–9, but younger and older students were surveyed at the discretion of programs. The programs included in the evaluation also met the target dosage and duration (greater than three weeks of STEM programming, for a minimum of one to two sessions per week), although there were students who self-reported participating in the program for less than one week. The majority of programs reported four or more STEM sessions per month, and most programs were ongoing throughout the academic year (Aug./Sept.– May/Jun.).

At the end of the evaluation in June 2016, a total of 160 programs that provide informal STEM instruction to students in grades 1 through 12 participated in the evaluation. Survey data were analyzed for students in Grades 4–12. The sites were reflective of the larger STEM afterschool universes across United States. Programs represented a variety of settings, including school-based (69.8%), community-based (28.2%) or other-based (2.0%), and the program size ranged widely between 3 and 54 students based on students observed participating in STEM activities. Programs reported using various types of STEM curriculum—with answers ranging from very specific to fairly broad in nature. Some examples of specific programs reported include Mindworks, Lego

Table 1. Program Selection Goals for State Afterschool Networks

Location	Choose programs that best represent the composition of the state in terms of rural, urban, suburban settings
Offerings	Select programs that best represent the curricular offerings of the state (i.e., no one curriculum dominates data pool for state)
Dosage/Duration	Aim to recruit programs offering STEM programming for three or more weeks, for a minimum of 1–2x/week
Grade Range	Aim to recruit programs with students in Grades 4–9 (younger and older students included at program's discretion)
Capacity	Choose programs that are able to complete all three components of the evaluation (student survey, facilitator survey, program quality observation)
State Support	Choose programs that have received varying degrees of state support (e.g., trainings, resources)

Robotics, Boston Museum of Science curriculum, Mindcraft/Coding, Zero Robotics, SciGirls, S.INQ Up, and NASA curriculum. Some less specific answers regarding program subjects include climate/animals, STEM, aviation, hands-on activities, web/Pinterest, and other activities researched online. Programs received varying levels of support through their state afterschool network's system-building work. Some examples of support include providing programs with STEM program quality observation training and certification, training of staff, coordination of STEM resources to support programming, curriculum-specific training (e.g., Wisdom Tools), resource materials, information/communications, grants and sustainability resources, coaching, technical assistance, evaluation, and professional development (generally).

Assessment Tools

In an effort to triangulate evidence of STEM learning, three assessment tools developed by Noam and team at The PEAR Institute were used. All of PEAR's tools were developed using a translational approach that combines academic research with feedback from practitioners in afterschool settings.

COMMON INSTRUMENT SUITE (CIS)

The CIS includes a battery of items that measure STEM-related attitudes and 21st-century skills (see Table 2). The core of this suite of tools is the Common Instrument (CI), a brief measure of student STEM interest in afterschool settings (Noam, Allen, et al., in preparation; Martinez, Linkow, & Velez, 2014). PEAR has recently expanded the CI to integrate other important STEM learning-related dimensions that can aid in the development of more effective afterschool science programming, including STEM career orientation and intrinsic motivation (adapted from OECD, 2010), STEM self-identity (adapted from Aschbacher, Ing, & Tsai, 2014; Cribbs, Hazari, Sonnert, & Sadler, 2015) and 21st-century skills such as critical thinking, perseverance, and relationships with peers and adults (Noam, Malti, & Guhn, 2012). The survey also included items to ascertain student characteristics, including gender, grade, race/ethnicity, primary language, and length of program participation.

Table 2. Outcome Measures for the Common Instrument Suite (CIS)

STEM-Related Attitudes	STEM Interest	How interested and enthusiastic a student is about science and science-related activities
	STEM Identity	How much a student self-identifies as a science person
	STEM Career Interest	How motivated a student is to pursue a career in science
	STEM Career Knowledge	How knowledgeable a student is about obtaining a career in science
	STEM Activity Participation	How often a student seeks out science activities
21st-Century Skills	Relationships With Adults	Positive connections and attitudes toward interactions with adults
	Relationships With Peers	Positive and supportive social connections with friends and classmates
	Perseverance	Persistence in work and problem solving despite obstacles
	Critical Thinking	Examination of information, exploration of ideas, and independent thought

Survey design. The CIS survey was created using the Qualtrics survey system and administered electronically using Wi-Fi-enabled tablet devices once at the end of STEM programming using a retrospective design. Students were provided with instructions and practice items at the start of the survey to ensure that students understood how to reflect on how much they felt they had changed as a result of participating in their program. Students were randomly assigned to complete one of two kinds of retrospective surveys: a retrospective pretest-posttest survey (75% of sample) or a retrospective change survey (25% of sample). The latter retrospective change design is a novel approach to the retrospective pretest-posttest format (see Appendix A for design description and instruction block). Pilot data for the retrospective change format will be utilized in future studies to evaluate the merits and validity of this innovative format with student populations. This report details the results from the well-established retrospective pretest- posttest design only.

The retrospective pretest-posttest method instructs students to rate each survey item twice from two different frames of reference: first to consider what they thought "before the program" and then to consider what they think "at this time." This design is similar to the traditional pretest-posttest method in that change is calculated by subtracting ratings for "Before the program" from "At this time." For the retrospective pretest-posttest survey, students' responses were recorded using a visual analog scale (VAS), a continuous scale of measurement (Gorrall, Curtis, Little, & Panko, 2016). The scale ranged from 0 (Strongly Disagree) to 99 (Strongly Agree), with a score of 49 representing the midpoint (Neutral).

At the end of programming, students were asked to think about themselves at a prior point in time (December 2015), rate themselves retrospectively, and then make a rating of themselves in the current state. Instructions for the retrospective pretest-posttest can be viewed in Figure 2. To help prime the retrospective thinking, a calendar image of December 2015 was presented in the instruction block. Students were additionally asked practice questions (see Figure 3), which were designed to help students understand the format of the retrospective pretest-posttest self-report design and VAS response format. To minimize survey length for the students, and to maximize the quality of data, a 10-form planned missing data (PMD) design was used (Rhemtulla, Savalei, & Little, 2016). A PMD design accounts for the reason that the data are missing and allows for the incomplete data to be easily recovered through multiple imputation (Enders, 2010; Little, Jorgensen, Lang, & Moore, 2014; van Buuren, 2012).

Figure 2. Retrospective Pretest-Posttest Survey Instruction Page

We would like to show you a few practice questions to help you understand how to answer some of the questions on this survey.

During this practice time you will be shown a sentence. Below the sentence you will see two places to pick an answer. For your answer, we want you to tell us how you felt before your afterschool program "Before the program." For your second answer, we want you to tell us how you feel right now "At this time."

Let's get started!

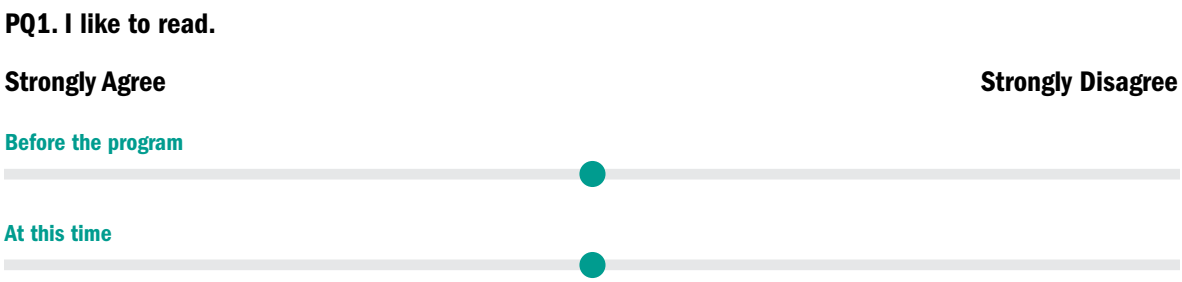
December 2015						
Sunday	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Before the program
Think back to this past December 2015 before you joined this afterschool program. Think about what was happening in December. Did you celebrate any holidays? Were you on winter break? What was the weather like? Did you see any movies?

When you go to pick your answers, remember to think back to how you felt in December. Then rate how much you agreed or disagreed with the sentence.

At this time
When you are asked how you feel "at this time," think about yourself right now because of your program. Then rate how much you agree or disagree with the sentence at this time.

Figure 3. Practice Block Question in Retrospective Pretest-Posttest Survey With Visual Analog Scale (VAS)



Design Rationale. The retrospective pretest-posttest design is an alternative method to the traditional pretest-posttest design that is commonly used to measure change in perceptions over time. In the traditional pretest-posttest design, a student responds to the same survey twice, such as before and after a given intervention. This is in contrast to the present retrospective pretest- posttest design, in which a student responds to the survey once, following a given intervention, but answers the questions from two frames of reference ("Before the program" and "At this time"). In our forthcoming paper on the retrospective pretest-posttest design, we detail many of the concerns associated with traditional pretest-posttest designs, which serves to support the choice to use the retrospective survey design (Little et al., in preparation). Briefly, the main outcomes measured using the CIS in the present evaluation are interest and self-beliefs about science-related activities. The traditional pretest-posttest design is likely to have biased responses for such outcomes at the pretest because the frame of reference of the respondent is unclear (Nieuwkerk & Sprangers, 2009). Ambiguous frames of reference lead to what is termed the "response-shift bias" (Howard, 1980; Schwartz, Sprangers, Carey, & Reed, 2004). In addition, a traditional pretest-posttest for self-related beliefs suffers from a lack of awareness on the part of the respondent. Other limitations of the designs include social desirability because responses cannot be anonymous (due to the need to track change over time) as well as retest effects and test-reactivity resulting from the repeat assessment of the exact same protocol (Bray, Maxwell, & Howard, 1984; Moore & Tananis, 2009).

The retrospective pretest-posttest design, on the other hand, does not suffer from these limitations (Howard & Dailey, 1979). The design forces the respondent to focus on the self at a particular point in time. Thus, the frame of reference for the respondent is assured (Drennan & Hyde, 2008). In addition, with the exposure to STEM activities, the respondent is capable of gauging prior levels of beliefs, interests, and attitudes compared to current levels of beliefs, interests, and attitudes. Reactivity and retest effects are also eliminated because the respondent must make two distinct judgments for each item (e.g., at the beginning of the program and at the time of assessment). These features of the retrospective pretest-posttest design are ideally suited to detect change when change occurs. Importantly, when change does not occur, the design is able to show the lack of change.

FACILITATOR SURVEY (CIS-FS)

The CIS-FS is a questionnaire for facilitators that was designed to complement the student CIS and was also developed to capture the unique qualities of STEM programs and the practitioners who lead STEM activities in afterschool programs. The CIS-FS contains questions about the training and professional development afterschool STEM facilitators received to lead informal STEM, the training and professional development that they would like to receive in the future, their ability and confidence levels for leading STEM, and their feelings about how they have impacted their students' proficiency and confidence in math and science as well as in social skills. The survey also included items to ascertain facilitator characteristics, including gender, race/ethnicity, highest level of education completed, and years of experience leading STEM. The survey was created using the Qualtrics survey system and administered electronically to facilitators at the end of STEM programming. For items related to perceptions of change, such as student's *math/science proficiency* and *math/science confidence*, facilitators responded to a sliding VAS that ranged from 0 (Decrease) to 99 (Increase), with a score of 49 representing the midpoint (No Change).

DIMENSIONS OF SUCCESS (DoS)

DoS is an observation tool for assessing program quality for STEM learning in afterschool programs (Shah, Wylie, Gitomer, & Noam, 2016; Noam & Shah, 2013; Papazian, Noam, Shah, & Rufo-McCormick, 2013). The tool is evidence-based and captures 12 dimensions of STEM program quality in afterschool along four organizing domains (see Table 3).

Table 3. Dimensions of Success (DoS) Program Quality Tool—Domains

Features of the Learning Environment	Activity Engagement	STEM Knowledge & Practices	Youth Development in STEM
Organization	Participation	STEM Content Learning	Relationships
Materials	Purposeful Activities	Inquiry	Relevance
Space Utilization	Engagement w/ STEM	Reflection	Youth Voice

Rigorous training and certification is required to perform DoS observations. State networks worked with DoS-certified individuals within their states to coordinate one or more program quality observations at each participating program to establish a score for program quality. Previous psychometric work by The PEAR Institute (Shah, Wylie, Gitomer, & Noam, 2014) has found DoS to have similar, and sometimes stronger, levels of agreement between raters than the agreement levels reported for observation tools used in studies in formal settings (Bell, Qi, Croft, Leusner, Gitomer, McCaffrey, & Pianta, 2014).

Observers recorded evidence of STEM learning during STEM activities for a minimum of 30 minutes (a maximum of 120 minutes, depending on activity length). Qualitative data from field notes were quantified using a standard rubric on a 4-point Likert scale from low (evidence absent) to high (compelling evidence). Field notes and ratings for each program were submitted electronically to The PEAR Institute, and feedback on program strengths and challenges was given to programs directly by DoS observers.

Procedure

This work can be conceptualized in three phases: methods preparation/program recruitment (Phase I), data collection (Phase II), and data analysis/reporting (Phase III). In Phase I, PEAR/IMMAP teams worked collaboratively to refine the student and facilitator surveys. For instance, we made modifications to items developed for older students (e.g., PISA-adapted) to fit the target age group in this evaluation (Grades 4–9). We also developed new items to measure other important student outcomes, including academic self-report and science identity.

In tandem with method refinement, the PEAR/IMMAP teams worked with Mainspring Consulting to select states and coordinate calls to answer the questions of state network leaders interested in participating in this work. In Phase II, PEAR/IMMAP finalized the electronic survey platforms, ordered and disseminated Kindle tablets across all networks, and hosted a training webinar (held in March 2016) to present an overview of the evaluation plan, goals, and tools to 160 individual programs across all 11 states. PEAR also assisted states requesting more individuals be trained and certified to observe STEM program quality using the DoS tool. In addition, PEAR/IMMAP developed and distributed to networks and participating programs a detailed student survey administration guide that described the goals of the evaluation, defined student outcomes being measured, and informed on best practices for administering the survey to students. Complementary technology and troubleshooting guides were also provided to assist programs with setting up and using the tablet devices to administer the student and facilitator surveys.

Lastly, a telephone hotline was provided to programs in the case of technology issues, and an automated email reminder system was established to notify programs (with permission from state networks) when it was time to administer the student and facilitator surveys. In Phase III, PEAR processed all ratings of program quality and IMMAP processed all student and facilitator data to prepare for data analysis. IMMAP imputed data for the PMD design and coordinated analysis/reporting with PEAR.

RESULTS

Participants

STUDENTS

A total of 1,599 students (733 female, 866 male) in Grades 4–12 completed the CIS student survey. Given that most programs were serving elementary and middle school children and youth, students in Grades 9-12 were combined to form a "high school" group.

Figure 4. Gender Distribution Across Grades 4–12

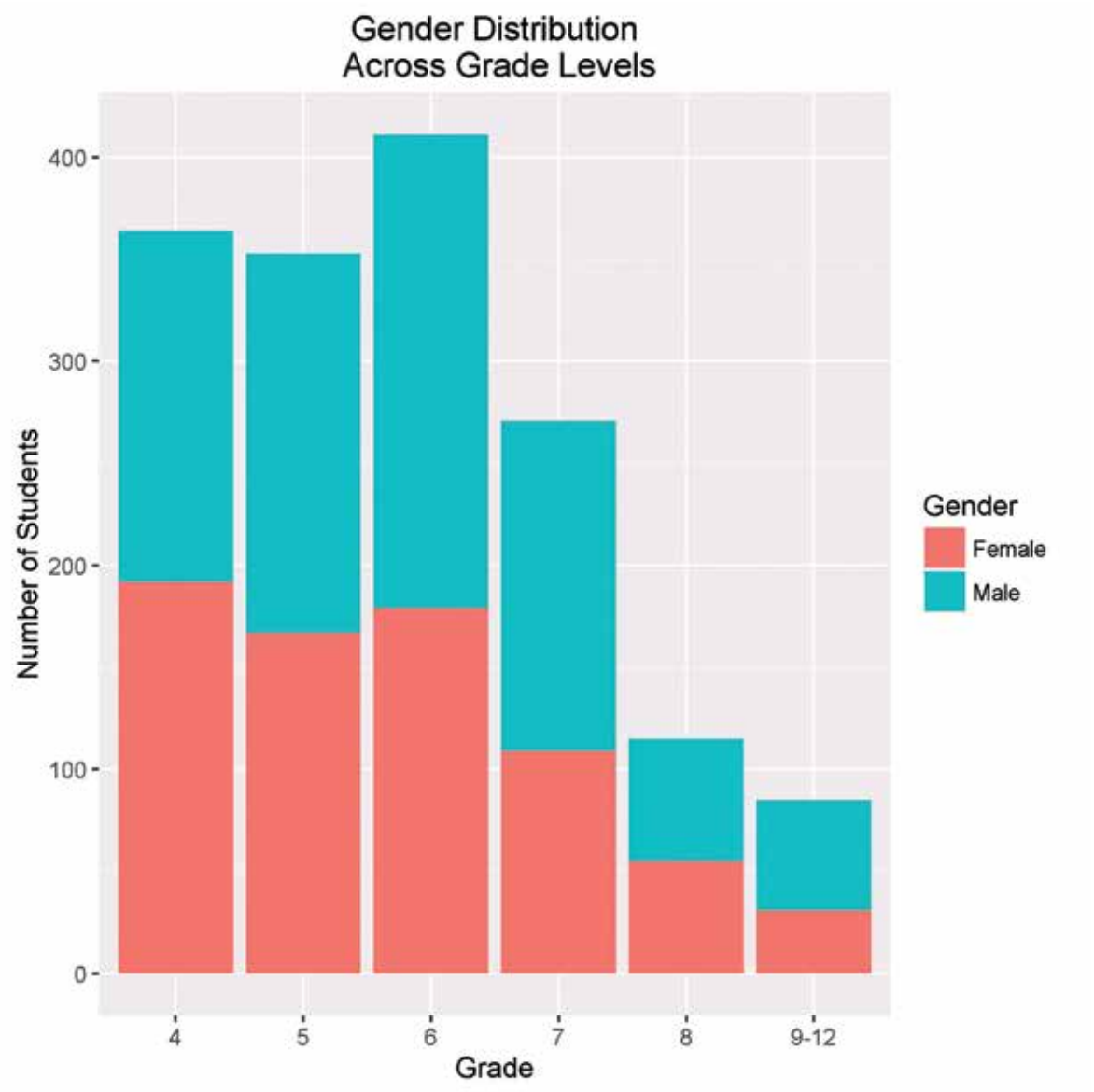


Figure 5. Gender Distribution Across Grade Levels by State

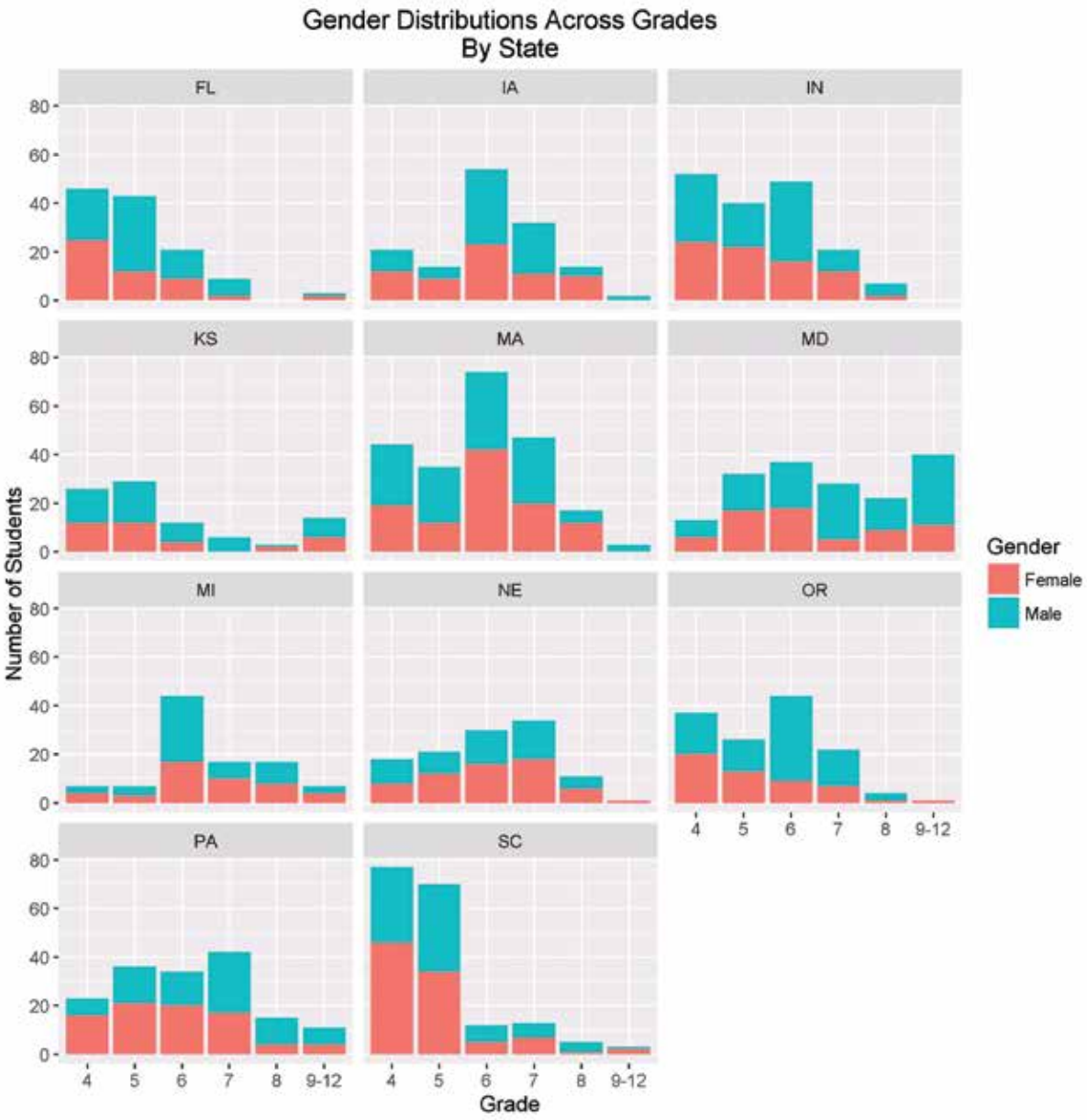


Figure 4 displays the gender distribution across the grade levels. Grade 6 had the largest number of participants, and Grades 8–12 had the smallest numbers of participants. The gender distribution across grade levels by state was also examined (see Figure 5). Indiana was the only state that did not have afterschool students in Grades 9–12, while a majority of students enrolled in programs in South Carolina were in Grades 4–5.

Figure 6. Student Demographic Information

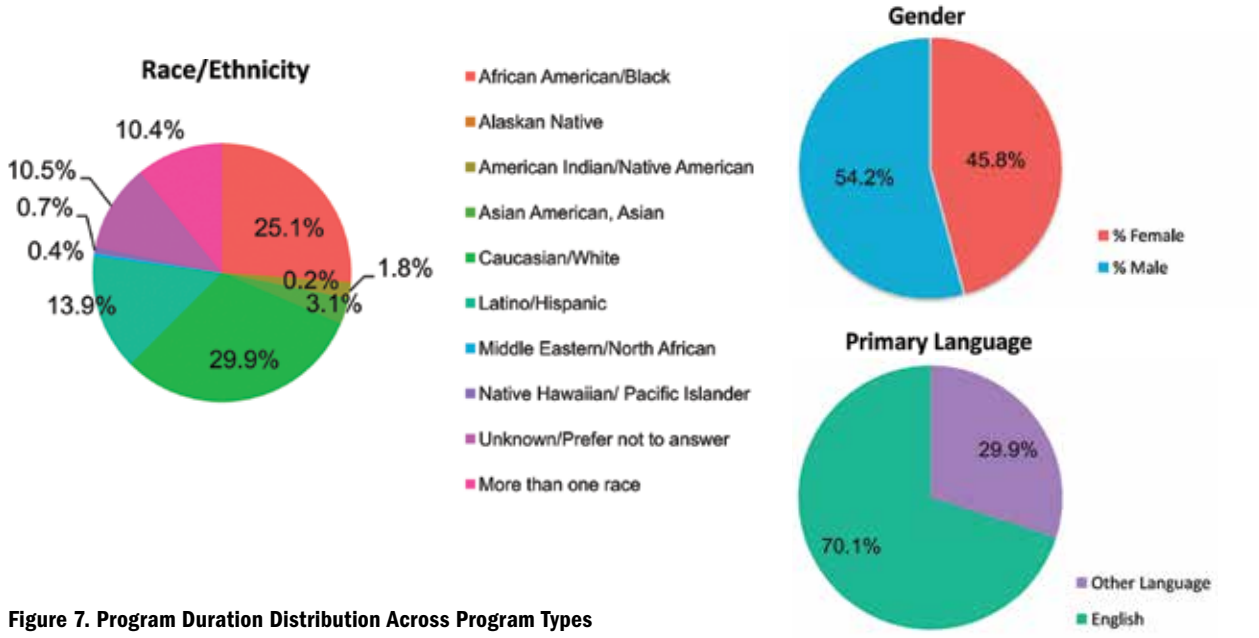
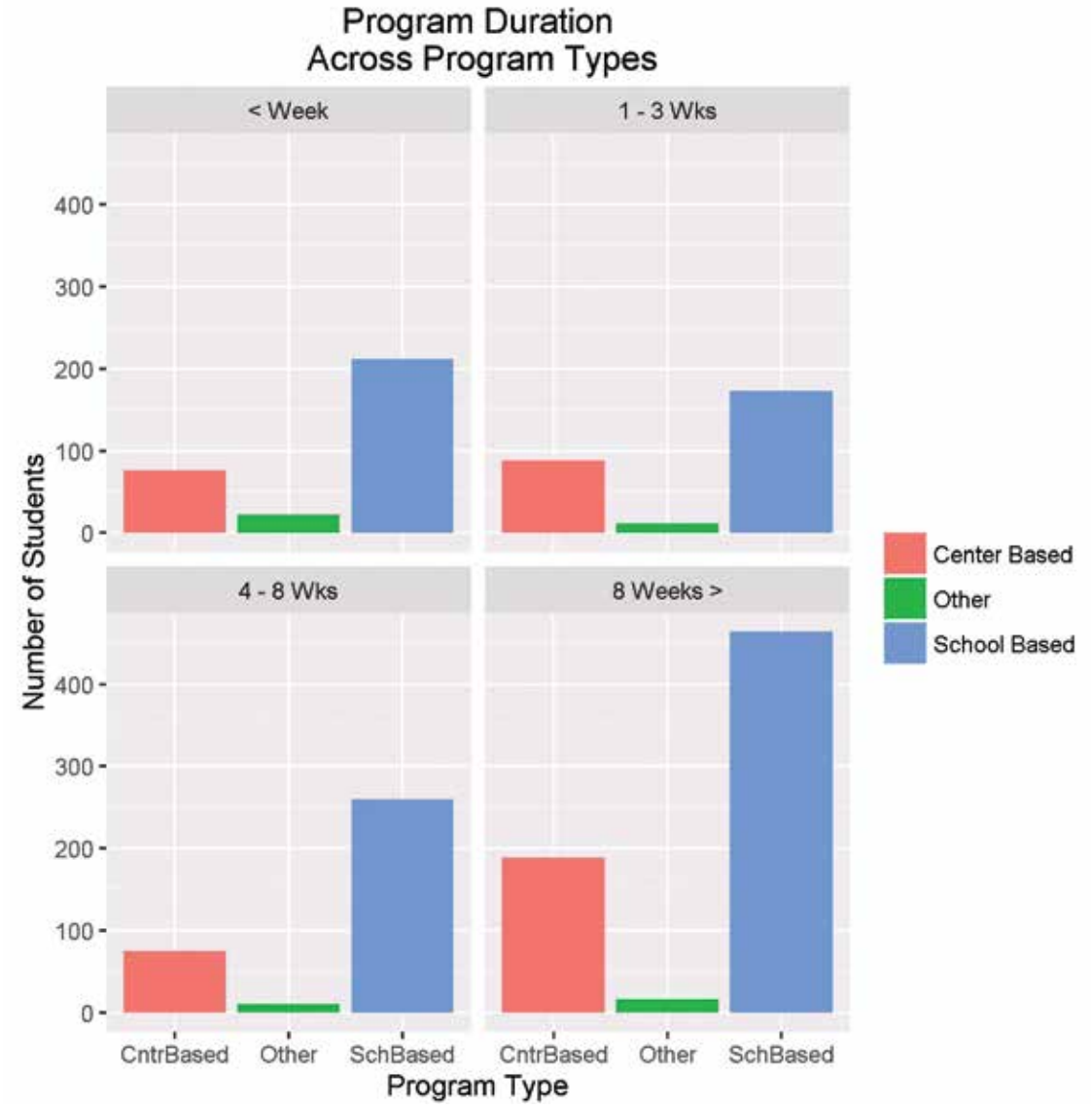


Figure 7. Program Duration Distribution Across Program Types



The sample was diverse (see Figure 6) and included groups that are traditionally underrepresented in STEM. Specifically, students identified as American Indian (1.83%) or Alaska Native (0.21%), African American/Black (25.05%), Asian/Asian American (3.11%), Latino or Hispanic (13.90%), Middle Eastern/North African (0.42%), Native Hawaiian or other Pacific Islander (0.71%), and more than one group (10.44%). About 10.5% of students preferred not to answer.

In addition, approximately one third of students speak a language other than English at home (29.9%). More than 60% of student survey respondents reported participating in STEM programming for 4 weeks or longer, and programming most often took place in a school-based setting (see Figure 7).

FACILITATORS

Across the 11 states, a total of 148 facilitators completed the facilitator survey (CIS-FS). Table 4 displays the sample characteristics of the facilitators. The average age of program facilitators was 38.21, and the 148 facilitators had an average of 15.48 years of experience working with children. A total of 65 (43.92%) of the facilitators were female and 95 (64.18%) of the facilitators had completed a higher education degree (associate, bachelor's, or master's).

Table 4. Demographics of Facilitators From Program-Level Survey

N = 148 Facilitators	Value
Average age of facilitator	38.21 years
Average years spent working with children	15.48 years
Female facilitators	43.92%
Race/Ethnicity	
Caucasian/White	45.95%
African American/Black	13.51%
More than 1 race/ethnicity	40.54%
Hispanic/Latino/Spanish Origin self-identity	38.51%
Education Level	
High school diploma/GED	29.05%
Technical degree	6.76%
Associate's degree	14.86%
Bachelor's degree	29.05%
Master's degree	20.27%

Student Survey Ratings

For the CIS retrospective pretest-posttest results, mean difference testing was conducted using paired *t*-tests to compare the change in ratings from "Before the program" (retrospective pretest) to "At this time" (posttest). Positive difference scores between the retrospective pretest and posttest provide evidence that students benefited from their enrollment in a STEM-focused afterschool program. Conversely, negative difference scores indicate that students did not benefit from their STEM- focused afterschool program. A neutral difference score (zero) reflected the student did not experience a change in attitude in either the positive nor negative direction.

OVERALL CHANGE IN RATINGS

In the following section, we highlight the findings across the nine core CIS constructs and academic perceptions. Data from students in Grades 4–12 were utilized to examine overall changes in STEM-related attitudes and 21st-century skills. To quantify the effects that program enrollment made on students' attitudes, we conducted simple *t*-tests and calculated effect sizes using Cohen's *d*. Cohen's *d* measures the difference between the retrospective pretest score and posttest score in standard deviation units (Hattie, 2009). In addition, proportions of difference scores were calculated to show the percentage of students in each state that experienced negative change, no change, or positive change after participation in their STEM-focused afterschool program. This descriptive statistic provides information about the magnitude (e.g., large or small) of the effects. Table 5 summarizes the results for the nine core CIS constructs across the 11 states. The detailed results for each state can be found in Appendix B, Table B1.

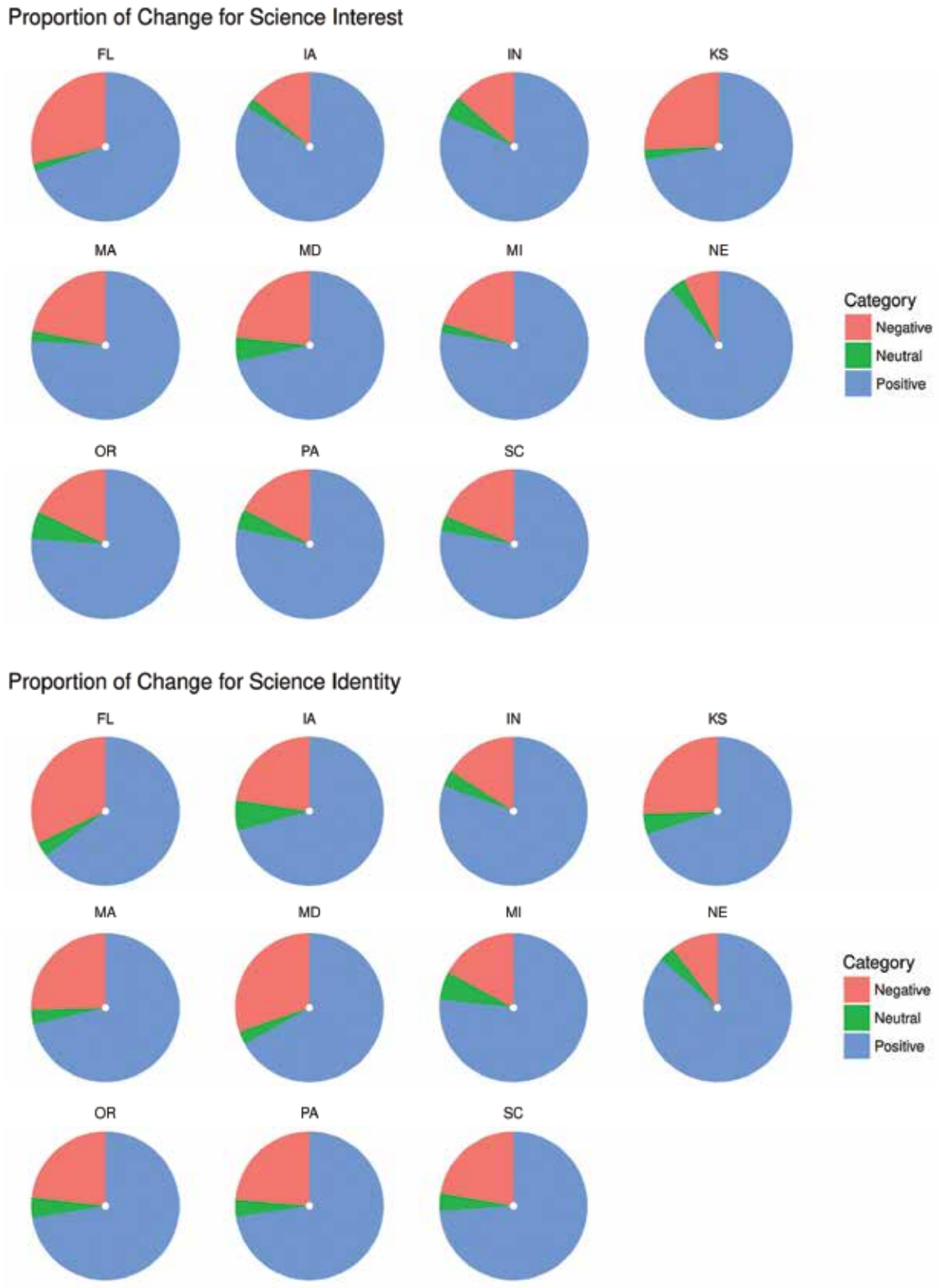
Table 5. Descriptive Statistics, Comparison Test Results, and Proportion of Changes in 11 states

Nationwide (*n* = 1,599)

	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
Variable	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
STEM Interest	60.174	19.864	70.023	19.254	26.559	p < 0.001	0.503	77.5%	3.50%	18.9%
STEM Identity	50.162	22.767	58.073	22.282	23.892	p < 0.001	0.351	73.1%	4.00%	22.9%
STEM Career Knowledge	44.989	22.318	55.684	21.502	28.557	p < 0.001	0.488	79.7%	3.10%	17.1%
STEM Career Interest	51.006	22.014	59.880	21.866	25.318	p < 0.001	0.404	75.7%	3.70%	20.6%
STEM Activity Participation	36.523	21.160	45.306	22.426	26.804	p < 0.001	0.403	76.7%	3.50%	19.8%
Relationships With Adults	64.240	19.744	71.736	18.440	20.653	p < 0.001	0.392	71.0%	4.90%	24.1%
Relationships With Peers	72.923	18.189	78.427	16.191	16.767	p < 0.001	0.320	64.5%	5.30%	30.2%
Perseverance	67.298	20.109	76.161	17.266	22.406	p < 0.001	0.473	72.4%	4.80%	22.8%
Critical Thinking	68.337	19.658	77.138	16.999	23.575	p < 0.001	0.479	72.9%	4.20%	23.0%

Science interest and identity. The overall results indicated that participation in a STEM-focused afterschool program made a positive impact on students' reported attitudes toward *STEM interest* and *STEM identity* (see Figure 8). For *STEM interest*, the aggregated proportion of positive change (positive difference score) across the 11 states was 78%, meaning that more than three-quarters of students across all states experienced a positive change in their self-reported attitudes towards *STEM interest* following participation in their afterschool program. For *STEM identity*, the aggregated proportion of positive change was 73%. Close to 75% of all students reported their *STEM identity* positively increased following afterschool program participation.

Figure 8. Proportion of Students Who Experienced Negative Change, No Change, or Positive Change in Their Science Interest and Science Identity Across the 11 States



Career orientation and intrinsic motivation. The overall results indicated there were significant increases in *STEM career interest*, *STEM career knowledge*, and *STEM activity participation* (see Figure 9). Effect size testing was conducted to quantify the impact program participation had on *STEM career interest*, *STEM career knowledge*, and *STEM activity participation*. All three constructs showed aggregated proportions of positive change in difference scores over 75%, with *STEM career knowledge* having a proportion of positive change close to 80%. More than 75% of students reported positive gains in *STEM career interest* and *STEM activity participation*, and almost 80% of the students across the 11 states reported a positive gain in their *science career knowledge* following program participation.

21st-century skills. Results showed participation in a STEM-focused afterschool program made a positive impact across all 11 states on students' 21st-century skills, including *perseverance*, *critical thinking*, and *quality of relationships with adults and peers* (see Figure 10). Effect size testing was conducted to quantify the impact program participation had on all four constructs. Over 72% of students across all states reported their *perseverance* and *critical thinking* skills positively increased following afterschool program participation. For *relationships with adults and peers*, 71% and 65% of students experienced a positive change in their self-reported social relation skills following program participation.

Figure 9. Proportion of Students Who Experienced Negative Change, No Change, or Positive Change in Their Science Career Interest, Science Career Knowledge and Science Activity Participation Across the 11 States

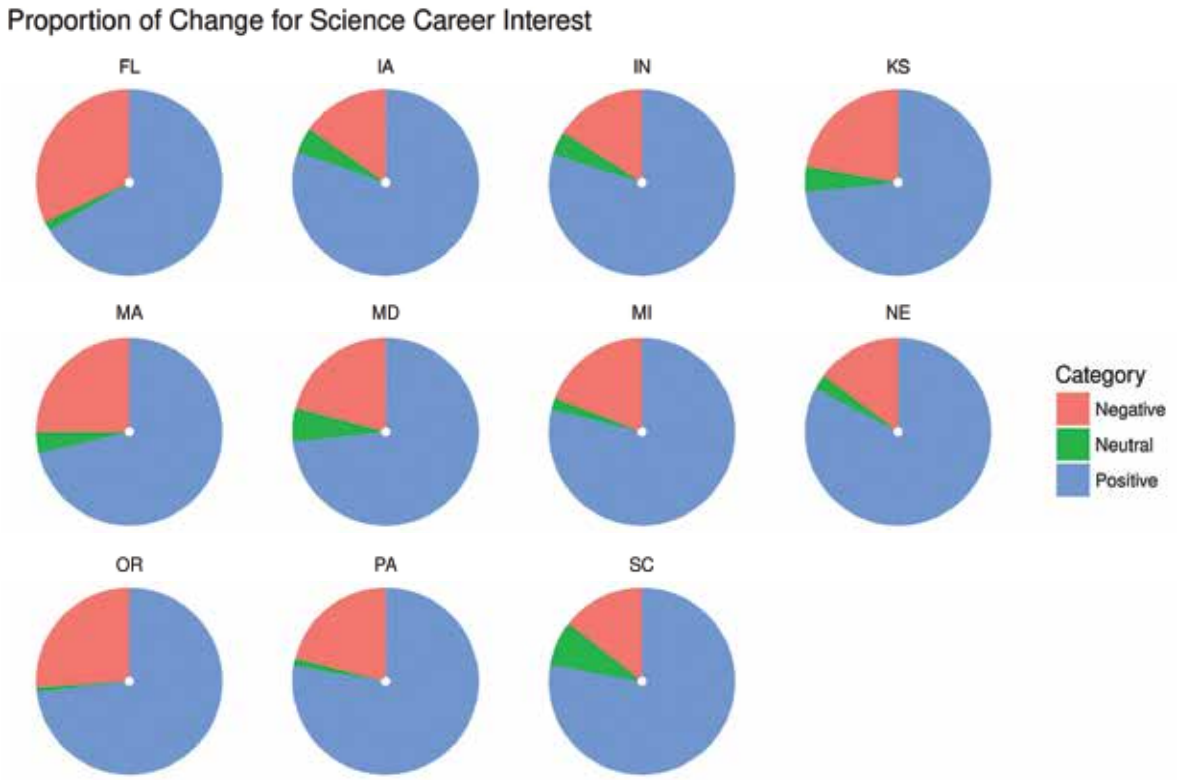
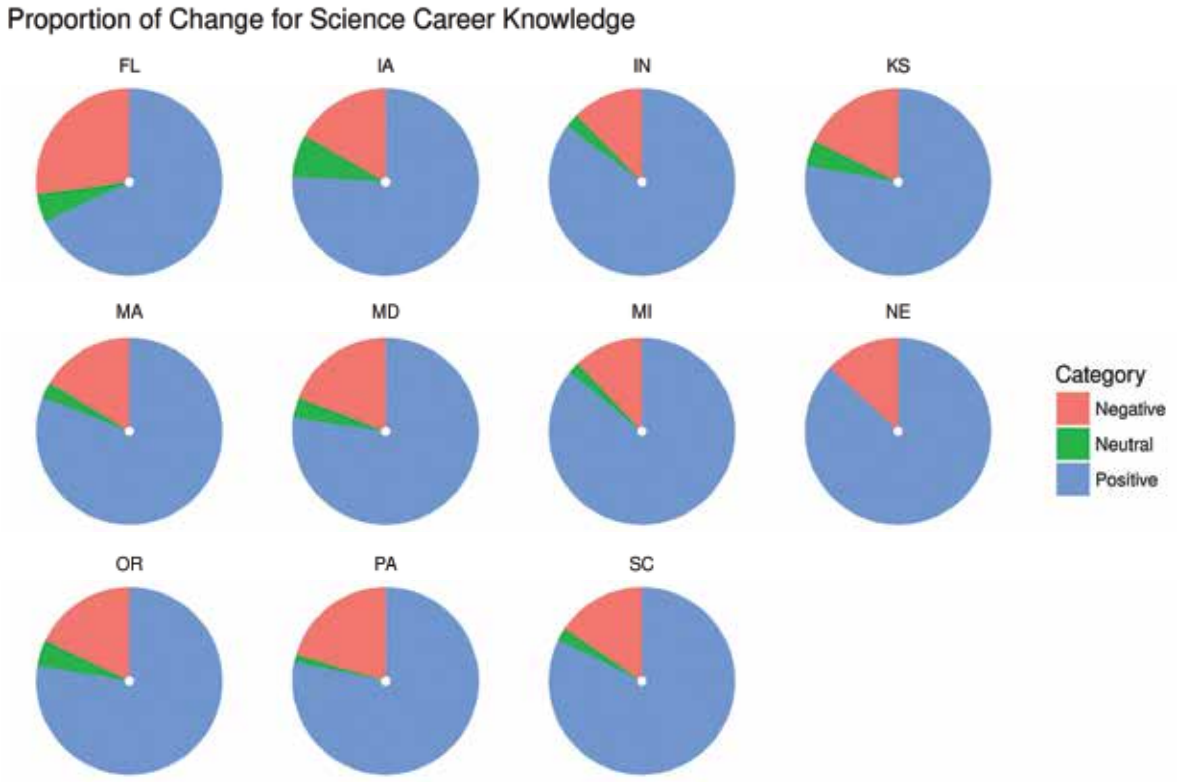


Figure 9. Continued



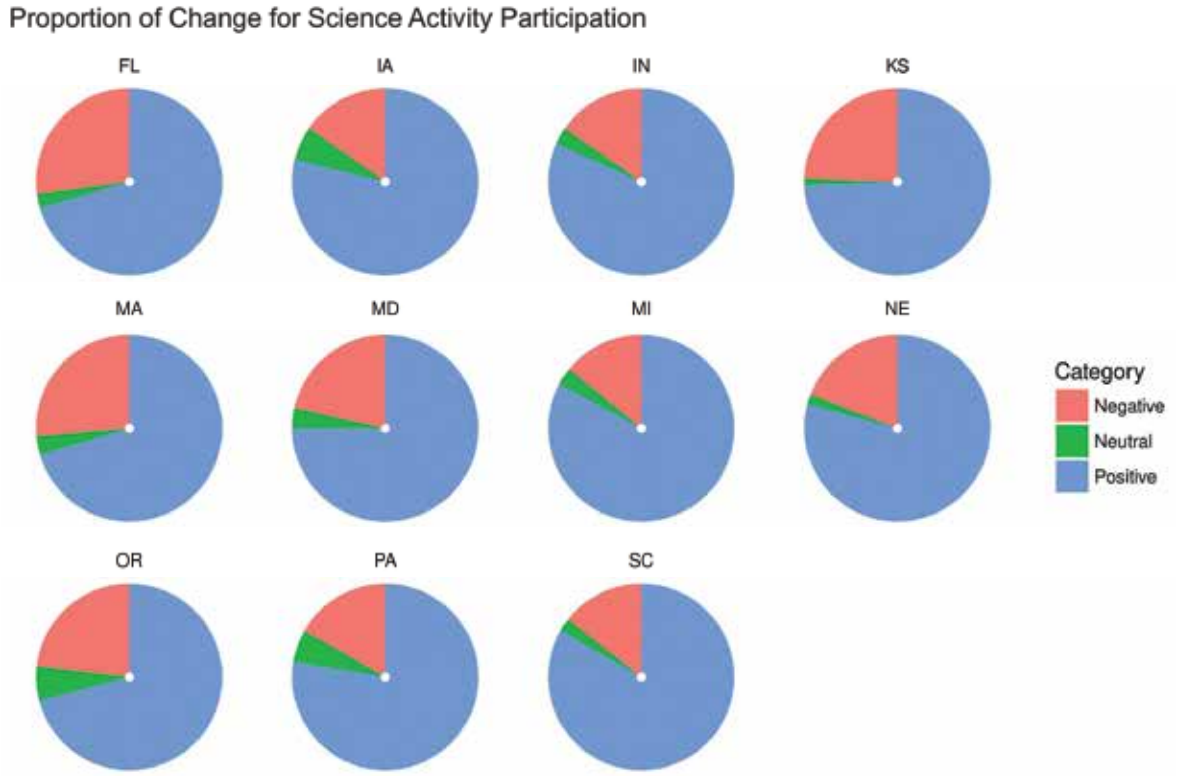
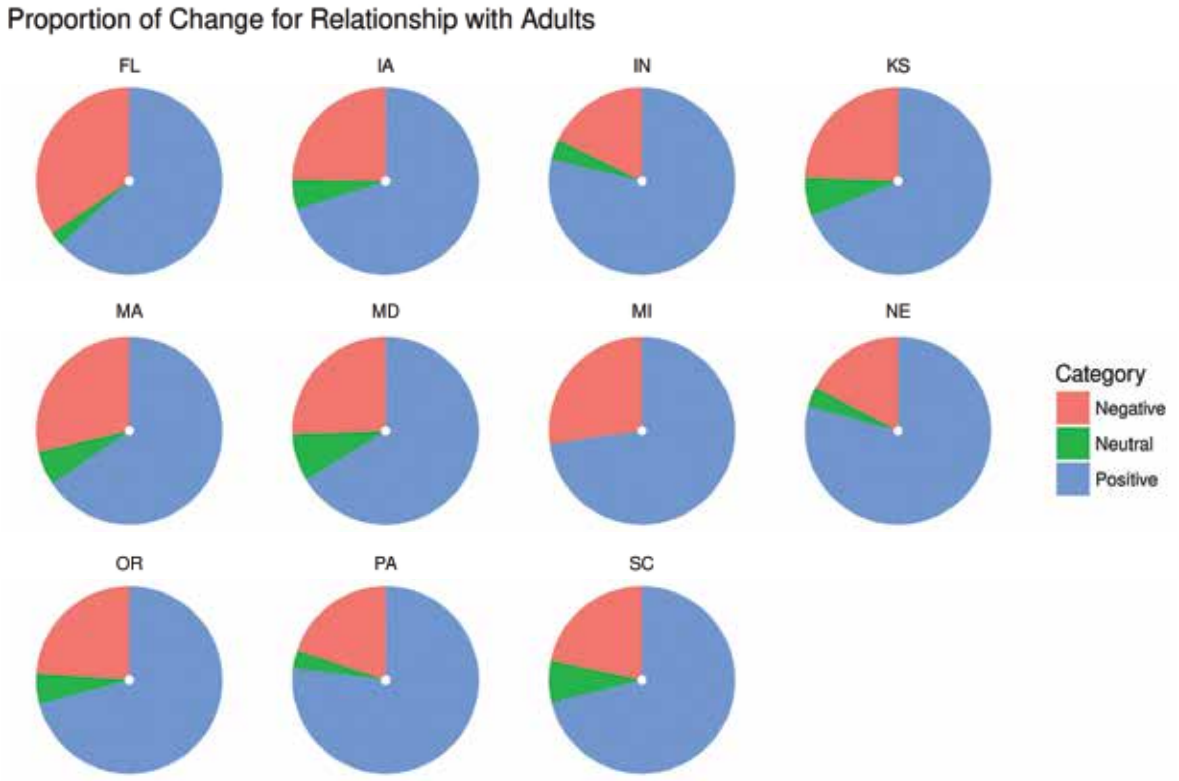
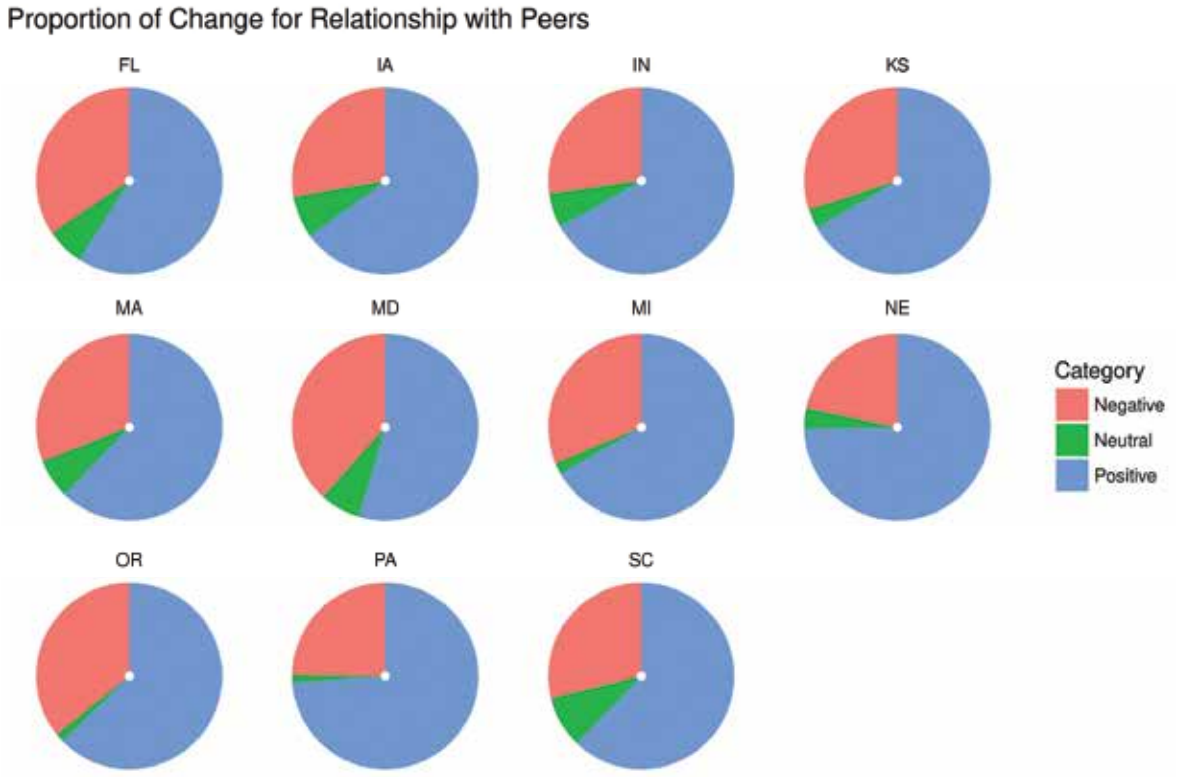


Figure 10. Proportion of Students Who Experienced Negative Change, No Change, or Positive Change in Their Perseverance, Critical Thinking, and Quality of Relationships With Adults and Peers Across the 11 States

Figure 10. Continued





Academic perceptions. Students were asked to report their attitudes (beliefs), perceptions of behavior (opinion of), and perceived cognitive state (how they feel) regarding their performance in math and science classes during the school period. Overall, posttest scores showed that students felt they had experienced academic gains after program enrollment. Effect size testing revealed that 61% of students had positive changes in their attitudes, 66% of students had positive changes in their perceptions of behaviors, and 65% of students had positive changes in their perceived cognitive state related to their math and science performance during the regular school day following program completion.

GROUP COMPARISONS

Gender. Difference testing was conducted to determine if students self-identifying as male or female responded differently to the retrospective pretest-posttest self-report design method. An analysis of variance (ANOVA) was conducted using the retrospective pretest scores and posttest scores by gender. Results revealed a statistically significant effect of gender on *relationships with adults and peers*, *with female students reporting higher retrospective pretest and posttest scores compared to male students* (see Appendix B, Table B2 for complete analyses results). To quantify the effects gender differences could have made on students' posttest scores for *relationships with adults and peers*, partial eta squared (η_p^2) was calculated to evaluate the magnitude of the effect. This effect size partials out the effects of the retrospective pretest score and will account for the proportion of total variance that is associated with group difference effects (Richardson, 2010). The effects of gender were minimal (η_p^2=0.004) for *relationships with adults and peers*, and males and females reported similar levels of positive change based on the difference between retrospective pretest scores to retrospective posttest scores. There were no other statistically significant effects on survey outcomes by gender.

Grade. Difference testing was conducted to determine if students at various grade levels responded differently to the retrospective pretest-posttest self-report design method. An ANOVA was conducted using the retrospective pretest scores and posttest scores (see Appendix B, Table B3 for complete analyses results). No statistically significant differences were found between students in Grades 4–12 on STEM-related attitudes and 21st-century skills. These results indicate that grade level did not significantly influence posttest score gains.

State. An ANOVA was conducted to determine if significant differences across the 11 states existed for students in Grades 4–12. Across all nine constructs, statistically significant differences were found among the 11 states. To quantify the effects group differences could have made on students' posttest scores, partial eta squared (η_p^2) was calculated to evaluate the magnitude of the effect. In Table 6, the effect sizes (η_p^2) and Tukey's HSD results are presented to show where significant differences exist between states on student-reported STEM-related and 21st- century skill attitudes. Across seven of the nine constructs (*science interest*, *science identity*, *science career knowledge*, *science career interest*, *science activity participation*, *relationships with adults*, *perseverance*), statistically significant differences were found between Florida and South Carolina. In South Carolina, the retrospective pretest scores were lower compared to Florida retrospective pretest scores. However, South Carolina students reported posttest scores higher in magnitude compared to students in Florida. On *science career knowledge*, nine statistically significant differences between states were found (see Table 6, refer to Appendix B, Tables B1 for state retrospective pretest and posttest scores). Retrospective pretest scores in Indiana, Michigan, and South Carolina were

lower compared to Florida. Yet students reported higher posttest scores in Indiana, Michigan, and South Carolina compared to students in Florida. Students in Maryland reported higher retrospective pretest and posttest scores compared to students in Florida. On *science career knowledge*, students in Indiana, Maryland, Michigan, and South Carolina reported higher retrospective pretest scores and posttest scores compared to students in Oregon. An aggregated effect size was calculated across the constructs to gain understanding of the average effect state characteristics had on STEM-related attitudes and 21st-century skills. On the statistically significant constructs, the effect of state characteristics contributed to 2% of the variance in the posttest scores following program enrollment.

Table 6. Effect Sizes and Tukey's HSD to Locate Significant Differences Between States

Variable	Effect Size (η^2)	Significant State Differences (Tukey's HSD)
STEM Interest	.017	FL:SC*
STEM Identity	.019	FL:NE*; FL:SC*; NE*:OR; SC*:OR
STEM Career Knowledge	.028	FL:IN*, FL:MD*; FL:MI*; FL:SC*; IA:MI*; IN*:OR; MD*:OR; MI*:OR; SC*:OR
STEM Career Interest	.017	FL:SC*
STEM Activity Participation	.018	FL:PA*; FL:SC*; MA:PA*
Relationships With Adults	.025	FL:PA*; FL:SC*; MA:PA*; MA:SC*
Relationships With Peers	.017	NE*:OR; SC*:OR
Perseverance	.018	FL:SC*; MA:PA*
Critical Thinking	.020	MD:PA*; MD:SC*

* denotes state with the greater difference between posttest and retrospective pretest

Program type. Difference testing between varying program types (center-based, school-based, or other) for students in Grades 4–12 revealed significant effects on all constructs except *STEM activity participation* and *relationships with peers* (see Appendix B, Table B4). On *STEM interest*, students in "other" program types reported higher retrospective pretest scores compared to students in school-based and center-based programs. However, students in school-based programs had posttest scores that were larger in magnitude compared to students in center-

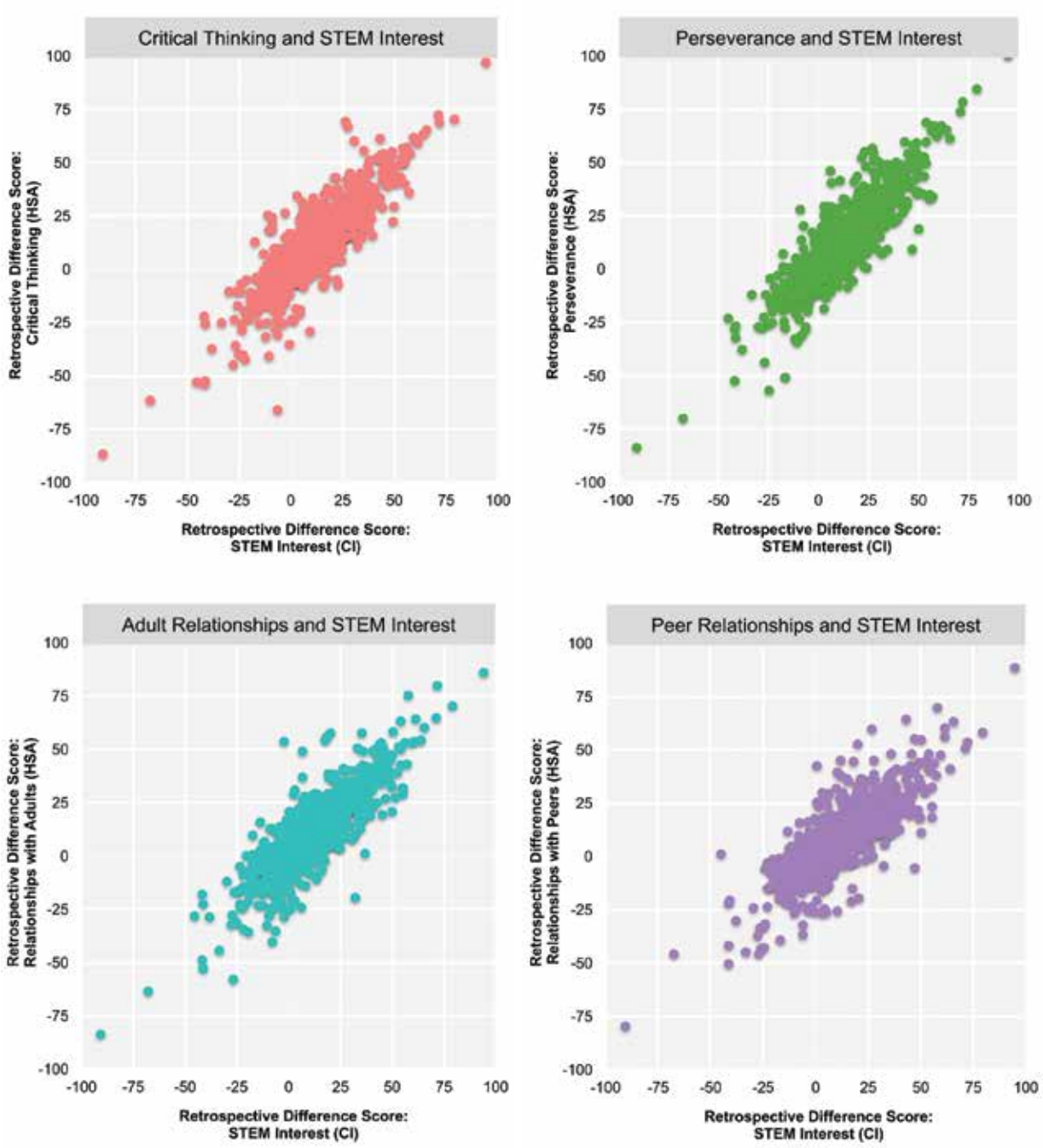
based and other-based programs. On *STEM career knowledge* and *critical thinking*, self-reported retrospective pretest scores were lower for students in school-based programs compared to students in center-based and other program types. Following program enrollment, posttest scores reported by students in school-based programs were greater in value compared to students in center-based and other program types. For *STEM identity* and *relationships with adults*, significant differences were found between school-based and other program types. On both of these constructs, students in school-based programs reported lower retrospective pretest scores compared to students in other program types. Following program enrollment, posttest scores reported by students in school-based programs on *STEM identity* and *relationships with adults* were higher in magnitude compared to posttest scores reported by students in other program types. Interestingly, on *STEM career interest* and *perseverance*, the effects of program types on posttest scores was found to be statistically significant; however, Tukey's HSD revealed no significant differences existed between program types. To quantify the effects group differences could have made on students' posttest scores, partial eta squared (η_P^2) was calculated to evaluate the magnitude of the effect. Of the seven constructs that showed statistically significant differences between students in varying program types, an aggregated effect size ($\eta_P{}^2$) was measured. Less than 1% of the known variance on the posttest scores is related to program type characteristics.

Program duration. The results of difference testing on students' self-reported program duration (less than 1 week, 1–3 weeks, 4–8 weeks, or greater than 8 weeks of programming) revealed that significant differences existed between all nine constructs on the retrospective pretest-posttest (see Appendix B, Table B5). To understand the effect time spent in the program had on STEM-related attitudes and 21st-century skills, effect size ($\eta_P{}^2$) testing was conducted. Close to 8% of the known variance in posttest scores across the nine CIS outcomes is related to the effects of program duration. Tukey's HSD indicated larger effects on posttest scores were found among students who participated in afterschool STEM for four to eight weeks and greater than eight weeks. Yet, there were no significant differences in the level of change reported by students who indicated they participated in afterschool STEM for a period of four to eight weeks and eight or more weeks.

Correlations: Science Attitudes and 21st-Century Skills

There were significant, strong positive correlations between students' self-reported change in all five science-related outcomes (*science interest*, *science career interest*, *science career knowledge*, *science activity participation*, and *science identity*) and self-reported change in the 21st-century skills *critical thinking*, *perseverance*, and *relationships with peers and adults* (see Figure 11). All *r*-values between the science-related outcomes and 21st-century skills ranged between 0.698 and 0.891. The 21st-century skills showed the highest correlations with *science interest*, with all *r*-values above 0.803. In other words, students who endorsed positive change in skills related to flexible thinking, problem solving, grit, and positive relationships with others also tended to endorse positive change in *science interest* (as well as other science-related outcomes).

Figure 11. Correlations Between Science Interest and 21st-Century Skills



Facilitator Survey Ratings

FACILITATOR PERCEPTIONS

Facilitators were asked to rate the level of change, if any, that they and their students had experienced throughout the duration of their program. Specifically, facilitators were asked to rate their perception of change in their students' *math proficiency*, *math confidence*, *science proficiency*, *science confidence*, and *social skills*. Regarding proficiency, 91% of facilitators perceived their students made significant improvements in *math proficiency*, and 91% perceived significant improvements in *science proficiency*. Regarding confidence, 92% of facilitators perceived their students made significant improvements in *math confidence*, and 91% perceived significant improvements in *science confidence*. For social skills, 91% of facilitators perceived their students made significant improvement. Additionally, facilitators were asked to rate their own perceived change in levels of *confidence*, *ability*, *interest*, and *professional development in STEM facilitation*, *frequency of attendance at professional development opportunities*, and the *priority placed on professional development*. Again, all of these metrics were shown to have significant positive changes. Across these constructs, 88% of facilitators felt that their confidence in STEM facilitation improved, 90% of facilitators perceived that their ability in STEM facilitation improved, and 92% of facilitators perceived that their interest in their STEM facilitation improved. Additionally, 92% of facilitators felt that professional development experiences had a positive impact on their STEM facilitation abilities, 86% of facilitators reported attending professional development opportunities more frequently, and 92% of facilitators felt that professional development in STEM facilitation was a higher priority for them.

There were also significant, positive correlations between facilitators' levels of *interest*, *confidence*, and *ability in STEM facilitation* and their perceptions of their students' *proficiency* and *confidence* in math and science. Specifically, facilitators reporting greater *interest* and *ability in STEM facilitation* perceived greater gains in their students' science and math *confidence* and *proficiency*. In addition, facilitators reporting gains in their *confidence in STEM facilitation* perceived greater gains in their students' *science confidence*; however, there was no relationship between facilitators' *confidence* levels and perceived gains in students' levels of *math confidence*. Similarly, facilitators reporting greater *interest* and *ability in STEM facilitation* perceived greater gains in the *social skills* of their students, though there was no relationship found between facilitator *confidence* and perceived change in *social skills* in students.

PROGRAM CHARACTERISTICS

Facilitators answered a series of questions about their afterschool program, including whether they used a specific STEM curriculum and what type of community their program served. Analyses using repeated measures ANOVA were performed to examine whether specific program characteristics had a significant effect on facilitators' perceptions of their students' *proficiency* and *confidence* in science and math, as well as whether program characteristics had a significant effect on students' self-reported science-related attitudes, 21st-century skills, and academic perceptions.

Regarding STEM curriculum, results indicated that facilitators who reported using a specific STEM curriculum perceived more positive gains in their students' math and science *proficiency* as well as in math and science *confidence* than facilitators who reported not using a specific STEM curriculum to guide their daily STEM activities. Overall, these data indicate that a specific STEM curriculum can be beneficial for facilitator perceptions of student gains. There were no differences, however, in pre, post, nor change over time for student self-reported science-related attitudes, 21st-century skills, and academic perceptions based on their program using a specific STEM curriculum.

Results also indicated that facilitators of programs that serve urban and suburban communities perceived more positive change in their students' math *proficiency* and *confidence* than peers in rural communities. Additionally, facilitators in programs serving suburban communities perceived more positive gains in their students' science *proficiency* and *confidence*. However, students enrolled in programs that serve urban communities (as determined by facilitator response) self-reported lower scores in *science interest*, *relationships with adults*, *relationships with peers*, *perseverance*, and *critical thinking* for both pre- and post- measurements. Additionally, students enrolled in programs that serve suburban communities self-reported lower pre- and post- scores in *science activity participation* and *relationships with peers*. Furthermore, students enrolled in programs that serve suburban communities self-reported higher pre- and post- scores in *relationships with adults*. Change over time for each of these student outcomes, however, were equivalent across all communities served.

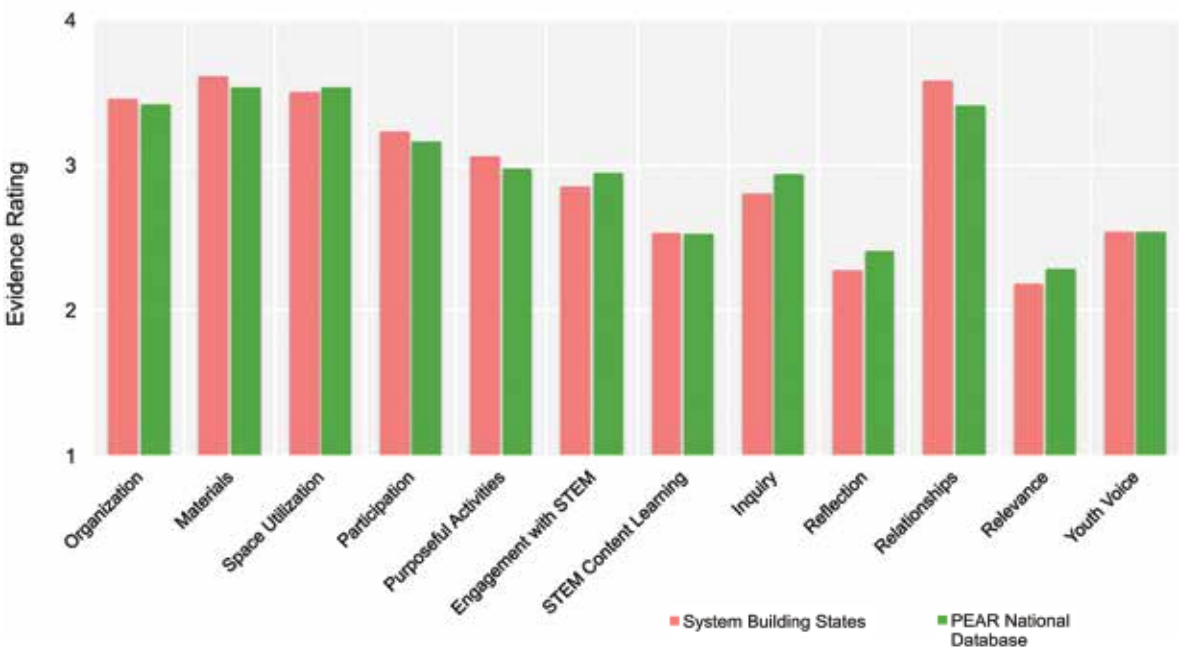
Lastly, facilitators of programs that focus specifically on teaching math outside of school (as opposed to science, engineering, or technology) perceived more gains in students' science *proficiency* and *confidence* (though interestingly there was no relationship with math *proficiency* and *confidence*). Students enrolled in math-focused programs self-reported higher pre- and post- scores for *science career knowledge* and *critical thinking*. Students enrolled in technology-focused programs (as determined by facilitator response) self-reported higher pre- and post- scores for all science-related attitudes and 21st-century skills. While both the pre- and post- scores were higher for students in technology-focused programs, the change over time was not different from nontechnology-focused programs for any of these outcomes.

Program Quality Ratings

There were a total of 252 DoS observations performed across 152 programs from March to June 2016. Note that this number differs from the total number of participating programs, as not all programs were observed by DoS during the evaluation period. Overall, the highest rated dimensions were *organization*, *materials*, *space utilization*, and *relationships*. Dimensions that may benefit from focus and professional development include *STEM content learning*, *inquiry*, *reflection*, *relevance*, and *youth voice* dimensions. To provide context for these results, the DoS ratings from the 11 state afterschool networks (n = 252) participating in the present evaluation were compared to PEAR's national database (n = 354), which represents observations performed across 10 states from January 2013 to January 2016. Note that PEAR's national database includes DoS data from system-building states that were collected prior to the start of the current evaluation, but there were no differences found between data collected from system-building states and data collected from other sources during this three-year period of time. The program quality results collected from the 11 states during 2016 are consistent with trends observed previously with one exception: system-building states showed significantly higher ratings for *relationships* compared to observations in PEAR's national database (see Figure 12). This difference is modest but suggests that the youth development focus of afterschool programs is improving through STEM system-building support.

To examine the relationship between student outcomes and program quality measures, standard scores were calculated using the sum total of ratings across all DoS dimensions. Programs receiving DoS ratings that were one standard deviation above the national average were designated as "higher quality," whereas DoS ratings that were one standard deviation below the national average were designated as "lower quality," and the remaining scores within one standard deviation of the national average were labeled "average quality." Data were categorized as higher, average or lower quality to better understand how DoS ratings for each state fared relative to the national

Figure 12. Comparison of DoS Data Collected by System-Building States Compared to PEAR Database

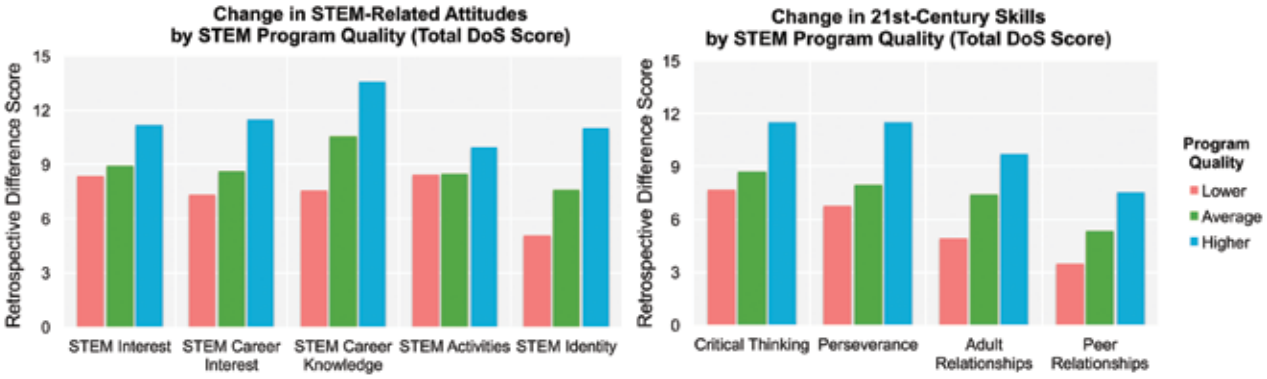


sample. This is important because it was determined that certain DoS dimensions often receive higher ratings than other dimensions, which indicates that some domains are harder to succeed in than others. For example, results indicated that most programs received high ratings on average (M = 3.51, SD = 0.64) for each of the dimensions within Features of the Learning Environment (*organization*, *materials*, *space utilization*), whereas most programs received low ratings on average (M = 2.59, SD = 0.92) for the dimensions in STEM Knowledge and Practices (*STEM content learning*, *inquiry*, and *reflection*). Analyses further demonstrated statistically significant differences in ratings between the four DoS domains (based on sum of mean scores for the three dimensions within each domain). The order of the domains ranging from most positive ratings to least positive ratings is as follows: *Features of the Learning Environment* (M = 10.56, SD = 1.50), *Activity Engagement* (M = 9.08, SD = 2.07), *Youth Development* (M = 8.40, SD = 1.84) and *STEM Knowledge and Practices* (M = 7.74, SD = 2.32). To better understand the average sum of scores that define lower-, average- and higher-quality ratings for each of the four domains, refer to the Appendix C.

Results showed that the magnitude of change in students' science-related attitudes and 21st-century skills, measured using the CIS, varied by program quality level (low, average, or high), as measured using DoS. Specifically, students attending programs that received higher-quality ratings (based on the sum of scores across all 12 DoS dimensions) reported greater gains in *science interest*, *science career interest*, *science career knowledge* and *science identity* compared to students attending programs that received lower-quality ratings. This was a quality-dependent effect, with outcomes for students attending average-quality programs in between those of students enrolled in lower- and higher-quality programs.

Likewise, students attending programs that received higher-quality ratings using DoS reported greater gains in *critical thinking*, *perseverance* and *relationships with adults and peers* compared to students attending programs that received lower-quality ratings. This was also a quality-dependent effect with outcomes for students attending average quality programs in between those of students enrolled in lower- and higher-quality programs (see Figure 13).

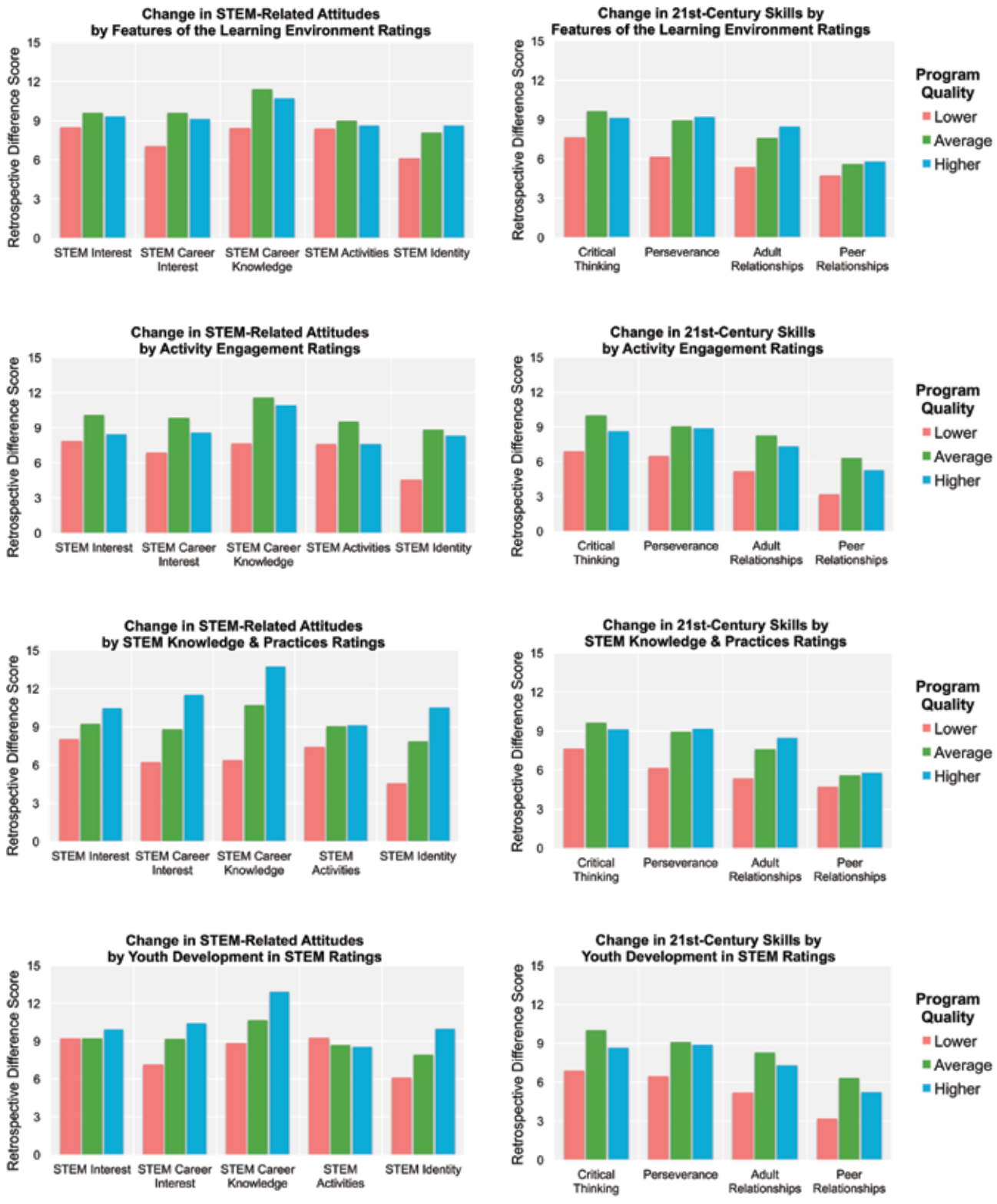
Figure 13. Comparison of Quality Development Across Programs on Science-Related Attitudes and 21st-Century Skills



The relationships between DoS ratings and CIS outcomes were further explored by examining the magnitude of change in CIS outcomes for each of the four DoS domains separately (*Features of the Learning Environment*, *Activity Engagement*, *STEM Knowledge & Practices and Youth Development in STEM*). This is important as it is clear that average ratings differ between each of the four DoS domains, and there may be differences in the contribution of different quality measures to change in student outcomes. Similar to what was found based on the total sum of DoS scores (collapsed across all four domains), when ratings from each DoS domain were analyzed separately, higher quality ratings were associated with more positive gains in student outcomes (see Figure 19). While the strength of the relationships between DoS and CIS varied by domain and by outcome, the most consistent finding was that an "average" quality rating or better was necessary to achieve positive change in student outcomes. All DoS domains were positively correlated with change in student outcomes, with the exception that *STEM Knowledge and Practices* was not correlated with students' self-reported *relationships with adults* and *Features of the Learning Environment* was not correlated with students' self-reported *relationships with peers*. Programs that received the highest ratings for the three dimensions within the *STEM Knowledge & Practices* domain had the most positive science-related outcomes, particularly for students' self-reported change in *science career interest*, *science career knowledge*, and *science identity*.

The relationship between DoS ratings and facilitators' perceptions of students' proficiency and confidence in math and science were also examined. Facilitators' ratings of change in their students' *science proficiency* and *math proficiency* levels were positively correlated with the dimensions within the *Features of the Learning Environment* and *Youth Development* domains. Additionally, facilitators' ratings of their students' change in *math proficiency* were also positively correlated with dimensions within the *Youth Development* domain. Facilitators' ratings of change in their students' confidence levels in science and math as well as social skills were positively correlated with the dimensions within *Features of the Learning Environment*, *Activity Engagement*, and *Youth Development*. On the contrary, the ratings for the *STEM content knowledge* dimension within the *STEM Content Knowledge & Practices* domain were negatively correlated with change in *math proficiency* and *science proficiency* as well as change in *math confidence*, *science confidence*, and *social skills*.

Figure 14. Change in Science-Related Attitudes and 21st-Century Skills Across Four DoS Domains Based on Lower, Average, and Higher Program Quality



DISCUSSION

The PEAR Institute and IMMAP, in partnership with state STEM network partners, successfully implemented this national evaluation and research plan that links STEM program quality measurement to student and facilitator outcomes. This project was highly collaborative, made possible with great effort and coordination across two funders, two research teams, one consulting team, 11 state afterschool networks, 160 afterschool STEM programs, and nearly 1,600 youth in Grades 4–12. This achievement serves as a proof point that it is possible to gather evidence of STEM learning in afterschool using common tools on a national scale. It also demonstrates the ability, need, and willingness of the afterschool STEM world to join the data-supported movement in education. The states valued having a common framework for evaluation because it lends a common language that can be used to communicate results within and among states, especially so that afterschool educators can learn from one another.

The primary aim of this nationwide initiative was to determine how the support provided by funders and state networks impacted STEM practices and 21st-century skills among youth across the United States. This is critical, as evidence indicates a steady decline in STEM interest as well as motivation to pursue STEM careers among youth (Osborne, Simon, and Collins, 2010; OECD, 2010). However, the out-of-school time field has been very resistant—for good reason—to rely on academic performance measures, especially as many afterschool STEM programs are youth development-based and do not teach directly to performance. Rather, afterschool programs teach other areas that we hypothesize can influence academic outcomes. Thus, our approach was to introduce the Common Instrument Suite (CIS) and Dimensions of Success (DoS), two data-creating tools that were conceived from a youth-development perspective and supported by academic literature. For instance, previous research has shown that gains in student outcomes such as STEM interest can improve science literacy (Dabney et al., 2011) and can predict academic achievement (Hughes, Luo, Kwok, & Loyd, 2008; Shiefele, Krapp, & Winteler, 1992), college readiness and acceptance (Wang & Holcombe, 2010), and STEM course enrollment and career acquisition (Wang, 2013; Watt et al., 2012). The CIS promotes youth voice by asking students their thoughts and feelings about STEM, such as interest and career motivation, as well as 21st-century skills, such as critical thinking, perseverance, and relationships. The DoS observation tool was designed to assess the quality of STEM activities and includes youth development as one of its core domains, particularly relationships, relevance, and youth voice.

The results from this large-scale evaluation are very encouraging. The project utilized a number of innovative methodological features from planned missing data collection protocols (Little & Rhemtulla, 2013) to the use of visual analog scales instead of Likert scaling (Gorrall, Curtis, Little, & Panko, 2016) to the principle component auxiliary variable approach to missing data imputation (Howard, Rhemtulla, & Little, 2015) to the retrospective pretest-posttest design (Allen & Nimon, 2007; Schwartz & Sprangers, 2010). Most important for this evaluation is the validity of the retrospective pretest-posttest design. At issue is the fact that we did not implement a traditional pretest, partly because of resource limitations but also because the traditional pretest approach has been criticized for its lack of sensitivity to detect change (Miller & Elder Hinshaw, 2012; Sprangers & Hoogstraten, 1989).

ON THE VALIDITY OF THE RETROSPECTIVE PRETEST-POST TEST DESIGN

In a forthcoming paper on the retrospective pretest-posttest design, we detail many of the concerns associated with traditional pretests, particularly when the constructs of interest are noncognitive in nature. Noncognitive may refer to "soft skills," such as attitudes, motivation, and relationships, whereas cognitive skills are often related to "hard skills" like performance and intellect. The constructs in this evaluation center on interest and self-beliefs about STEM-related activities. Such constructs are likely to have biased responses at pretest for a number of reasons. First, as described earlier, the frame of reference of the respondent is unclear (Nieuwkerk & Sprangers, 2009). Are the responses in comparison to the self at a prior time point, at the current time point, or compared to others? Such ambiguous frames of reference lead to what is termed the "response-shift bias" (Howard, 1980; Schwartz, Sprangers, Carey, & Reed, 2004). In addition, with little experience with STEM-related activities or only weak informal experience, the true beliefs that a respondent has about science interests and self-beliefs can be ill informed at the pretest and thus also be prone to response-shift bias. There are also concerns related to repeated testing and lack of anonymity, since repeated administrations require tracking students from the pretest to posttest (Bray, Maxwell, & Howard, 1984; Moore & Tananis, 2009).

The retrospective pretest-posttest design minimizes or removes these concerns (Howard & Dailey, 1979). For example, the design forces the respondent to focus on his or her self at a particular point in time. In this case, we primed respondents to think back to December 2015 and probed for memories of that time (see instructions example in method section). Thus, the frame of reference for the respondent is fixed (Drennan & Hyde, 2008). In addition, retrospective surveying occurs after youth participate in STEM activities, increasing awareness of the self. Lastly, because there is only one surveying time point, retest effects and test-reactivity are not an issue.

FUNDED STUDIES THAT HAVE USED THE RETROSPECTIVE PRETEST-POSTTEST DESIGN

Numerous studies from federally funded efforts have utilized the retrospective pretest-posttest design. These studies include examinations of educational, social, and health science program outcomes. Funded by National Institute on Drug Abuse, Moberg and Finch (2007) examined program outcomes of high school students (n = 321) recovering from a substance use disorder across 18 high schools in seven states (California, Colorado, Minnesota, Pennsylvania, Tennessee, and Texas). Moberg and Finch argued that the retrospective pretest-posttest design was the only alternative design to use. Using the retrospective pretest-posttest design, they found a significant reduction in substance use as well as in mental health symptoms. In addition, students were found to have positive attitudes about the therapeutic value of the schools but were less enthusiastic about the educational programs.

Pratt, McGuigan, and Katzev (2000) were funded by Oregon Healthy Start Evaluation, and they evaluated longitudinal data of mothers (n = 307) with first-born infants who participated in a home-visitation, child-abuse prevention program. In this study, a traditional pre- and post-design and a retrospective pretest-posttest design were both implemented. The data was collected when the infant was 1 to 7 days old (pretest) and 6 months old (posttest). A seven-item self-report measure was used to assess maternal knowledge of child development, confidence in parenting, etc. Results indicated that all seven items on the measure showed a significant improvement on the retrospective pretest-posttest design; however, only four items showed a significant improvement on the pre- and post- design (i.e., there was an underestimation of program effect). Pratt et al. (2000) also found the presence of a response-shift bias between the mean scores

on pretest items and the retrospective pretest items. Further examination revealed evidence of response-shift bias on the three items that failed to show significant change on the pretest-posttest design. The researchers concluded the retrospective pretest-posttest methodology provided a legitimate assessment of program outcomes compared to the traditional pretest-posttest design.

Kreulen, Stommel, Gutek, Burns, and Braden (2002) conducted a longitudinal nursing intervention study funded by the National Cancer Institute. The study investigated substitutability of retrospective pretest ratings for pretest ratings due to logistic problems of collecting pretest data. Women (n = 251) receiving breast cancer treatment rated their concerns regarding their perceived health status satisfaction. The data were collected during structured telephone interviews (three times after the start of the intervention). They compared three waves of self-report data both prospectively and retrospectively at 2-week intervals during a 6-week intervention period. The results indicated a moderate level of agreement between the two measurement approaches. The satisfaction scores measured retrospectively were similar to those measured prospectively in means and variances. The results of regression analysis demonstrated that recall satisfaction ratings reflect true score variation in prior current satisfaction ratings though this relationship is only moderately strong. Kreulen et al. (2002) concluded that retrospective assessment might be preferred to actual pretest measures for measuring changes because it can provide a way to monitor bias and evaluate change more sensitively.

INTERPRETING EFFECT SIZES FOR OVERALL CHANGES IN SCIENCE-RELATED ATTITUDES AND 21ST-CENTURY SKILLS

To understand the magnitude of the impact of the program evaluation, effect sizes were calculated on overall changes in attitudes and group differences. We used a measure called Cohen's *d* to measure the effect sizes of the overall changes in attitudes. When we have a normal distribution of responses, we can calculate the amount of shift in the scores of all the individuals in the distribution using the *d* calculation. Cohen's *d* quantifies the amount of change in the mean level of the distribution on a standardized scale. This standard metric allows us to calculate what proportion of students rate themselves higher (at the posttest) than the mean score at the pretest period. When there is no change, the two distributions of scores overlap completely. Hattie (2009) provides guidelines for interpreting Cohen's *d*. Hattie's critical *d* value for concluding a positive result is a *d* value of .2 or greater. An intervention shows "promise" when the *d* value is between .1 and .2. In the analysis of the overall changes in STEM-attitudes and 21st-century skills, the effect sizes in this evaluation varied from around .16 to above .60 in many cases. For effect sizes like we see in Florida, for example, a value of .30 means that 62% of the respondents at the posttest scored higher than the mean from the pretest. This positive effect indicates a 12% increase over chance. For effect sizes such as we see in Nebraska, a *d* of .66, for example, means that 73% of the respondents are above the mean of the pretest. This level of a superior effect indicates a 23% increase above chance.

As reported above in the Results section and detailed in Appendix B, the size of the effects for overall changes in science attitudes and 21st-century skills was typically in the superior range. Given that this project spanned 11 states and 160 diverse programs, effects sizes reported here range from promising to superior. Nebraska, for example, was consistently at the top when it comes to showing strong effects of the intervention ((four constructs with *d*'s $\geq$ .6), while Florida showed much more modest effects (all nine constructs had *d*'s $\geq$ .3). Moreover, the differences in the effects also varied depending on which outcome was rated. And these differences

between outcomes also varied by state. These varied patterns suggest that states differed in their implementation of capacity building, which had different points of emphasis, and in the demographic makeup of the evaluation sample. This pattern also indicates that both ways of assessing change are sensitive to these variations in the measured outcomes.

In terms of effect sizes seen in the literature, the results reported here are quite promising. Some published evaluations using the retrospective pretest-posttest design showed weak effects, while others showed stronger effects, comparable to some of the states that we report on here. Across 42 measures in a collection of 16 studies that contained traditional pretest-posttest and/or retrospective pretest-posttest methodologies, an omnibus Cohen's *d* was calculated to examine the effectiveness of both methodologies. On the traditional pretest-posttest self-report design, Cohen's *d* values range from -1.59 through 4.85, with the aggregated Cohen's *d* being equal to.39. In other words, slight differences were detected between pretest and posttest scores. For the retrospective pretest-posttest, Cohen's *d* ranged from .01 to 2.81, and the aggregated Cohen's *d* was 1.031. Moreover, these findings demonstrated the attenuation of change scores when traditional pretest-posttest self-report designs are utilized. The retrospective pretest-posttest method has consistently demonstrated the ability to overcome design limitations of traditional self-report methods.

DIFFERENCES BY GENDER, GRADE, PROGRAM TYPE, AND PROGRAM DURATION

Using the retrospective pretest-posttest design, gender was not found to play any role in student STEM-related outcomes, which is in contrast to many published findings showing that boys have significantly more positive STEM-related attitudes than girls (Desy, Peterson, & Brockman, 2011; Weinburgh, 1995). However, gender had a small but statistically significant effect on students' perceived quality of relationships with adults and peers. Compared to male students, female students reported higher quality relationships with peers and adults at both the beginning and end of programming (based on retrospective pretest and posttest scores, respectively). This outcome is consistent with literature describing gender differences in perceived quality of relationships (Fabes et al., 2014); however, more research is needed to understand how these differences in perceptions may influence afterschool program dynamics or future academic and career success, especially in STEM fields. An examination of student outcomes by grade level indicated that there were no differences in STEM-related attitudes and 21st-century skills based on year in school, regardless of whether a student is in Grade 4 or Grade 12. This is in contrast to previously published research studies reporting a decline in STEM-related attitudes as students get older. For instance, youth in middle school and high school have been found to think less positively about STEM and to show less interest in obtaining a STEM career compared to youth in elementary school (VanLeuvan, 2004; Potvin & Hasni, 2014).

An examination of youth outcomes by program type indicated that enrollment in a school-based, center-based, or other type of programs had a minimal but significant effect on seven of the student outcomes measured (*science interest*, *science identity*, *science career knowledge*, *science career interest*, *relationships with adults*, *perseverance*, and *critical thinking*). Here, students enrolled in school-based programs reported more positive gains than students enrolled in center-based and other program types. However, less than 1% of the accounted-for effects in the difference score between the retrospective posttest and the retrospective pretest was attributed to the effects in program type. Additionally, it is important to note that the majority of programs participating in this evaluation identified as school-based.

Finally, program duration influenced youth's ratings. Students participating in afterschool STEM for a longer period of time reported better outcomes with larger effect sizes than students participating for a shorter period of time. Specifically, students participating in afterschool STEM activities for four weeks or longer reported significantly greater STEM-related attitudes (*STEM interest*, *STEM career interest*, *STEM career knowledge*, *STEM identity*, and *STEM activity participation*) and 21st-century skills (*critical thinking*, *perseverance*, and *relationships with peers and adults*) than students participating for less than four weeks. Close to 8.0% of the accounted- for effects in the difference score was attributed to the effects of time spent in program. Self-reported program duration is a reasonable index of dosage and in this context, dosage matters.

POTENTIAL LIMITATIONS AND FUTURE DIRECTIONS FOR STUDENT-LEVEL EVALUATION

Possible threats to the validity of a retrospective pretest-posttest design are experimenter effects and social desirability (Drennan & Hyde, 2008; Finkelstein, Quaranto, & Schwartz, 2014; Hill & Bentz, 2005). Experimenter effects occur when participants know the hypothesis of the experimenter and when the experimenter is present, the responder responds in the direction of the hypothesis. This effect is not likely because we did not have an experimenter present. Students were given a Kindle tablet with a URL and simple instructions to tell us how they feel. Social desirability is also less likely because the survey is anonymous and both the retrospective and the current rating would likely have the same degree of socially desirable responding such that the difference between the two ratings would not be impacted. Nevertheless, the authors do acknowledge the concerns about the retrospective design, including memory-related problems (e.g., memory distortion, selective perception, and poor memory, especially among children and adolescents), social desirability, and impression management and response bias. The authors detail issues pertaining to both the retrospective pretest-posttest and traditional pretest-posttest designs in a forthcoming paper.

Another limitation of this project is that we were unable to measure actual skill in science due in part to the diversity of skills that are represented and also due to limits in resources. Measuring actual skill changes in science would necessitate a traditional pretest-posttest design because the outcome measure of interest is actual aptitude. Future work should examine the change in skills and the degree to which changes in beliefs and attitudes predict and are associated with changes in skills.

Another future direction for the analysis of this intervention is to drill deeper into differences by program type to examine the degree of variability across programs and examine potential predictors of the differences by program type. Another future direction would be to employ a latent variable modeling approach to correct the estimates of effect size for measurement error (Bollen & Lennox, 1991; Little, 2013; MacCallum & Austin, 2000). When measurement error is removed, the size of each effect that we reported here will go up depending on the amount of error in the scale used.

STEM FACILITATION AND PROGRAM CHARACTERISTICS

There are several important insights gained from the responses of the facilitators who lead STEM activities outside of school. First, facilitators were very positive in their assessment of student gains in terms of proficiency and confidence in math and science as well as in social skills. More than 90% of facilitators who responded to the survey reported positive gains in their students across these five outcomes. Second, the majority reported feeling more confident (~ 88%), competent (~90%), and interested (~92%) in facilitating afterschool STEM. Third, more than 91% of facilitators felt that professional development experiences had a positive impact on their STEM facilitation abilities. Fourth, facilitators serving urban and suburban communities perceived greater improvement in their students' proficiency and confidence in math in comparison to rural communities, whereas facilitators in suburban communities also perceived greater improvement in their students' science proficiency and confidence in comparison to rural and urban communities. These findings suggest that the financial and resource support given to afterschool programs helped increase the confidence, abilities, and interest of the STEM facilitators. It also showed that facilitators would like even more professional development in the future.

STEM PROGRAM QUALITY

Another important goal was to examine levels of STEM program quality across the states by using the DoS observation tool and to determine if there is a link between STEM program quality and student outcomes. DoS observations were performed at the same afterschool programs as the students who completed the CIS surveys. The results demonstrated a strong relationship between the quality of STEM activities and students' self-reported gains in science-related attitudes, including *science interest*, *career interest and knowledge*, and *science identity*, as well as 21st-century skills such as *critical thinking*, *perseverance*, and *relationships*. Specifically, student ratings of change in science-related attitudes and 21st-century skills were significantly lower in programs with lower STEM program quality ratings compared to programs with higher STEM program quality ratings. These findings serve to substantiate the linkage between program quality and student outcomes and underscore the importance of focusing on quality improvement to enhance student gains.

A closer look at the student outcome trends within each of the four DoS domains (*Features of the Learning Environment*, *Activity Engagement*, *STEM Knowledge and Practices*, *Youth Development in STEM*) suggest important considerations for program leaders and practitioners. Levels of *science career interest*, *science career knowledge*, and *science identity* reported by students were most highly related to the quality ratings of the *STEM Knowledge and Practices* domain. In other words, while a lower quality rating in any of the four domains was associated with less positive outcomes among students, this effect was much more dramatic for programs with lower quality ratings in the *STEM Knowledge and Practices* domain. Importantly, programs with the highest ratings in the *STEM content learning*, *inquiry* and *reflection* dimensions reported the most positive gains in science-related attitudes and 21st-century skills measured using the CIS. However, it is important to highlight that this DoS domain has proven to be the most challenging for afterschool STEM programs to master, which underscores the need for further professional development in *STEM content learning*, *inquiry*, and *reflection*.

A related goal was to examine the relationship between facilitator beliefs about their own STEM facilitation and changes in their students' confidence and proficiency in math and science as well as in social skills. There were several significant relationships detected. For instance, facilitators of programs given higher quality ratings in the *Youth Development* and *Features of the Learning Environment* domains tended to rate their students' science and math proficiency more positively than facilitators whose programs received lower quality ratings in these domains. Interestingly, in contrast, facilitators whose programs received lower ratings for *STEM content learning* tended to rate their students' science and math proficiency more positively than facilitators whose programs received higher quality ratings. This may suggest that facilitators receiving the highest ratings in *STEM Content Learning* are underestimating their impact on their students' social skills or their proficiency or confidence in math and science, or vice versa. It is also plausible that facilitators receiving lower quality ratings are interpreting the quality of *STEM content learning* differently, leaving some to believe their students are improving when in reality the facilitators are not doing well with including quality STEM content in their activities. The *STEM content learning* dimension falls within the *STEM Knowledge and Practices* domain—the same domain in which most programs receive low ratings. This information serves to emphasize not only the important role of youth development, activity engagement, and learning environment, but also the need to target professional development around improving *STEM content learning*, *inquiry*, and *reflection*, the three dimensions in the *STEM Knowledge and Practices* domain.

These findings have important implications for program leaders who train and support facilitators as they implement STEM programming. It is important to choose professional development activities that encourage teaching practices that best support STEM learning instead of focusing solely on the organization or space use of an activity or how to choose appropriate curricular materials. Program leaders may reduce some of this burden on facilitators by providing vetted and research-based curricula to facilitators versus asking them to create their own curricula. They can also provide the DoS Program Planning Tool (DoS-PPT), which is a complementary resource to the DoS observation tool. The DoS-PPT provides prompts and activity design suggestions to help facilitators plan both what they want to do in their activity and what teaching approaches they want to employ. For example, the planning tool encourages facilitators to list specific questions they would like to ask to increase cognitive engagement and determine how they will set-up the materials to ensure all students can participate. As facilitators use the planning tool through each cycle of planning activities, the goal is to continue to increase their quality scores during observations using the DoS observation tool.

Specifically, for the *STEM Knowledge and Practice* domain, the DoS-PPT can be used to help facilitators design experiences that are not just "hands-on" but that engage youth in authentic practices of STEM professionals. In other words, facilitators learn how do we encourage students to observe, collect data, analyze, report, and build arguments like STEM professionals. This is something that many trained science classroom teachers struggle with as well, as it involves very careful guidance and resistance to the idea of "giving away the answer." Therefore, afterschool facilitators, who often have little to no training in leading STEM, need this additional support. The findings from the present work indicate that facilitators' ability to help students grapple with STEM concepts, practices, and knowledge in a meaningful way can make a great impact on student-reported science interest and identity.

DETAILS OF PROJECT INNOVATIONS

As mentioned, this evaluation brought a number of innovations to afterschool STEM evaluation world. First, we employed a retrospective pretest-posttest survey method, which allows students to self-report change in science attitudes while minimizing response-shift bias that is inherent to traditional pretest-posttest survey designs (Lam & Bengo, 2003). In particular, response-shift bias can occur when there is a change in understanding or perception about concepts between pretest and posttest ratings. For instance, an intervention can change the way individuals perceive themselves with regards to how they feel about something or how much they know about something, which in turn changes their frame of reference. Thus, the comparison of results from two time points is less valid when a perceptual shift occurs in between.

To better illustrate the potential response-shift bias in science-related attitudes and 21st-century skills (which is negated by the use of the retrospective designs), IMMAP calculated response-shift bias on prior CIS data collected between 2014 and 2015 by The PEAR Institute from afterschool programs in two system-building states, Indiana (n = 11 programs, 411 students) and Massachusetts (n = 17 programs, 343 students). The results from this work showed that there was response-shift bias for scales such as *science interest*, *science career interest*, *critical thinking*, *perseverance*, and *relationships*. In other words, there was a dissonance between students' actual ratings at the beginning of the program and their ratings for how they think they felt at the beginning of the program after having experienced the program. These data support our rationale for using retrospective design, namely that results from traditional pretest-posttest designs are less valid than results from the retrospective design because the students' internal frames of reference shift during the course of STEM programming.

A second innovation was the decision to rely on electronic data capture using Wi-Fi-enabled tablets (or other available smartphones, tablets and computers), which can dramatically improve efficiency of data collection as well as the quality of data. Each of the state afterschool networks received three to four tablets per program recruited (ranging from 27 to 66 tablets per state, or 539 tablets total), and programs were also instructed that students could use any available smartphones, tablets, or computers available. There were concerns about programs having reliable access to the Internet. However, we have found that very few programs had technology-related issues, and both staff and students were highly adaptable to the use of technology. Feedback from programs was positive, such as how the use of technology was a good match for programs designed to teach STEM, and the tablets helped engage students because they were new and exciting. The use of web-based surveys provided at least four additional benefits: (1) they saved substantial time in terms of survey dissemination and data entry, (2) they saved substantial resources so that programs did not have to print or mail surveys, (3) they allowed for advanced programming that systematically reduced the total number of items per student (planned missing data design) and that helped students focus on one question at a time, and (4) they allowed for the use of visual analog scales that are more sensitive to change that traditional Likert-based response scales. The successful wide-scale use of technology in this evaluation should encourage the field to consider moving toward electronic data capture for all future efforts. In the case of this study programs were able to keep the tablets for educational use and for future evaluative projects.

SUMMARY OF CHALLENGES FACED AND OVERCOME

As with any major evaluation effort, there were a number of challenges related to data collection and analysis that can serve improved future work. It is important to note that all of these challenges were resolved by a network support team that included PEAR, IMMAP, and Mainspring Consultants—none hindered or prevented the positive completion of project. There is an understandable learning curve involved with the use of data-creating tools—especially when programs are joining a particularly large-scale effort with many moving parts for the first time and they had the more challenging task of educating youth.

On the technical end, the most challenging issue this project faced was related to errors around assigning survey identification numbers (ID#s), which are required to ensure the integrity of the data collected. For instance, when completing surveys, many students did not enter survey ID#s, entered survey ID#s incorrectly, or were mistakenly assigned the wrong survey ID#s. As a result, extra time was required to resolve user issues and to clean and process the data. Related to data collection, there were occasional issues with program Internet access that resulted in the rescheduling or cancelation of surveying, though this was rare. Overwhelmingly, programs embraced electronic data capture.

In terms of recruitment, some state networks had trouble reaching the sample size goal of at least 15 students per program. This was unavoidable in some cases, such as in programs in rural areas, but the benefit of including different program types was more important to ensure the representativeness of the sample. A related issue was recruitment of DoS observers who would be willing to travel across large states to observe programs; stipends were effective to a point, but establishing a large network of trained observers across all regions in the states would help to mitigate this problem.

Regarding inclusion, the work was limited in scope in terms of the grade range of students participating. While the academic literature points to middle school as a pivotal point in the development of science interest and identity, we have received feedback from state network leaders that many STEM programs are beginning to enroll younger children (below Grade 4), and they want to collect data to understand this group of students. Nevertheless, there are many challenges around surveying students this young. The field needs to develop innovative methods to evaluate STEM learning in students ages 8 and below since this is an important emerging demographic in the afterschool STEM universe.

In terms of communication, some state networks indicated that they preferred to remain the main point of contact for reminders and problem solving, so any issues at the individual-program level had to be resolved through an intermediary. On the logistical end, the delivery of electronic devices also provided some issues, as many state networks preferred delivery of notepads to individual programs across their large states. It would be helpful to have a centralized registration database so that program information, contact information, and mailing addresses of all programs participating within each state are up-to-date. This would facilitate the distribution of information and materials to programs partnering with state afterschool networks.

Despite these hurdles, which are characteristic of any large-scale evaluation effort, we believe that this project will make an important contribution to quality improvement for afterschool STEM programming within the Mott-Noyce STEM Initiative. We have reached a wider audience than expected, which makes us very optimistic about future efforts. We were able to build on four years of previous cooperation with state networks, which made this phase of the work especially productive. The collaborative relationship with the state networks and their leads was essential for success of this large-scale effort.

FINAL THOUGHTS AND RECOMMENDATIONS

This evaluation demonstrates that this large-scale initiative to effect positive change in youth outcomes, including science interest and identity, science career orientation and motivation, and 21st-century skills, shows success. All states exhibited superior effects for one or more youth outcomes, with approximately 65–85% of youth reporting positive gains across the 11 state afterschool networks. Given the selection of states to approximate the national census data, which included rural, urban, and suburban programs, the demographic diversity of the sample, and the consistency in findings, one can anticipate similar results in states that share the same level of support, professional development, and leadership across the United States. It is important to note that some states showed modest effects in certain outcome areas that were found to be superior for other states, but this likely reflects the fact that states were in different phases of system-building implementation and had different strategies for supporting and training programs. It probably also had to do with the quality of implementation, the experience of staff, and other unmeasured factors. Notably, the survey and observation tools used were sufficiently sensitive to capture these expected state-level differences. Follow-up inquiry with states is warranted because it is important to know what strategies the states with the most superior quality and outcomes are using to make their efforts more successful than others. There is a great deal that can be learned from the length of time that has been invested in system-building, including the infrastructure that states are building to support quality afterschool STEM programming.

A substantial finding is the link between DoS program quality ratings and CIS student outcomes; youth participating in programs with higher quality ratings in STEM report greater improvements in science-related attitudes and 21st-century skills. Also, facilitators who lead programs with higher quality ratings believe they are having a greater impact on their students' social skills and proficiency and confidence and math and science. Lastly, the more confident and able facilitators feel about leading STEM, the more confident and able they believe their students to be math and science. Nevertheless, additional work is needed to develop a logic model to assess an "if-then" causal relationship. We cannot yet show the theory of change because the strategies and activities within each state are unique. Though all states have common system-building elements since they belong to cohorts, states are in different phases of system-building implementation, and each approach is tailored to the specific assets and needs of the states and the states' partners.

For instance, some states have excelled in communication and policy, whereas others have excelled in quality building, due to differences in factors like the partnerships, resources, and infrastructure available in each state. However, the evidence described here using a triangulation method, which includes self-report of youth and facilitators, as well as observations of the quality of settings and activities, is robust and shows that real change is happening to boost quality and youth outcomes. Nevertheless, there needs to be more research, training, collaboration, and technical assistance to continue this positive trend. While there were many higher quality programs identified, it is clear that the work to improve afterschool STEM is not done. There are many programs with areas that require improvement—but there is growth potential. There were several clear outcomes captured in this work that make it possible to introduce a number of significant recommendations.

1. **Leverage leaders' strengths:** An effective strategy for further improving program quality and youth outcomes is to continue to support the growing community of system-builders in their efforts to address key system components: partnership and leadership development, quality building and professional development opportunities, communication and policy, and evaluation and data collection. The states' goals are to map the landscape of afterschool and STEM efforts; to engage key stakeholders; to prioritize and act through communication, policy, and professional development; and to measure the supply, quality, and impact of STEM programming in afterschool. It is important to note that states can decide on their own process for accomplishing these steps in whichever order they choose, and logically many have begun by playing to their own unique strengths. Thus, states excelling in specific elements of system-building should be encouraged to teach others through communities of practice. For instance, some states have been doing better in the quality capacity building components, whereas others are better in communication and outreach. However, the strengths and challenges of each state network was not quantified in the present work. A future study is needed to examine the specific strategies and activities of the state's STEM efforts in relationship to program quality and outcomes. We recommend that future work focus on identifying, quantifying, and leveraging the strengths of each of the states (and clusters of strong programs) to lift up all programs across all states. It is important to recognize that states can be very different (e.g., regionally, demographically, economically, politically), but by studying states that are succeeding in greater detail, the other states can adopt or modify the strategies and partnerships to fit with their own state's unique qualities and capabilities. This recommendation is not directly supported by our research, but relates to the findings of the significance of program quality.

2. **Target professional development and quality support:** Continuous improvement through professional development for facilitators and feedback from program quality observations will strengthen such outcome areas as interest in science, career knowledge and interest, science activity participation, and skills in relationships, perseverance, and critical thinking. Importantly, this evaluation showed that program quality is statistically and significantly correlated with students' perceived outcomes in STEM. Thus the strategy pursued in this initiative to spend a good amount of time on program quality improvement before assessing outcomes paid off. It does not make logical sense to assume strong outcomes in weak programs. Having these results, it is essential to further strengthen the quality of programs. This evaluation helped us determine specific areas of need identified by facilitators and STEM quality observers. Specifically, facilitators expressed a desire for additional support in programming ideas, program management, and how to connect afterschool programming with the school day. DoS observers identified a general need for more support to improve STEM content learning, inquiry, reflection, relevance, and youth voice in the implementation of STEM activities. These outcomes were independent of curriculum used or type of setting. Importantly, the data indicate that superior outcomes can only be achieved with adequate training, support, and commitment to continuously improve based on feedback. These findings from facilitators and quality observers across the states should highlight to state-level decision makers the need for more professional development and quality support using a coordinated system-building approach. These findings should also motivate state and program leaders to incorporate quality and facilitator feedback for continuous improvement.

3. **Focus on the linkage between science learning and 21st-century skills:** In an effort to reverse the worrisome national trends in science interest, performance, and selection of STEM careers, afterschool programs have incorporated enriching science learning opportunities into their original focus on youth-development-oriented programming—but without knowing for sure if this combination would fit together or if science would be viewed as an "add-on" by students, facilitators, and families. Importantly, the present findings support the fit between youth development/21st-century skills and science activity participation; youth who reported feeling that they have stronger critical thinking skills and more positive relationships with adults and peers as a result of their program also reported more positive gains in science-related attitudes. The causal direction of this relationship is unclear at this point, but important life skills like teamwork, collaboration, flexible thinking, and grit are important for succeeding in science. Notably, program quality observations indicated that most programs exhibited clear evidence of strong relationships, meaning that programs foster nurturing interactions that create a positive supportive atmosphere. We propose that, by supporting the whole child through youth programming, afterschool programs can enhance every child's potential to learn and thrive in science. We recommend that states and programs embrace their strengths and focus on developing quality in afterschool STEM from a youth-development/21st-century skills perspective; for instance, sparking interest through hands-on activities (active engagement), providing guidance from science role models (identity and belonging), allowing youth to make decisions around the steps in an activity (youth voice and assertiveness), and encouraging thoughtful questions and application to everyday life (reflection and relevance). More work is needed to establish the links between the best practices in youth development in afterschool (such as connecting with students and creating safe learning environments) and the best practices in informal STEM learning (such as providing opportunities to learn science concepts and engaging youth in hands-on exploration and inquiry). Thus, youth development and informal science should continue to be integrated to simultaneously address the socioemotional/21st-century needs of students while also sparking their curiosity and skills in science.

4. **Encourage use of data to inform practice:** One of the greatest successes of this evaluation was the demonstration that it is possible to rally diverse states and afterschool programs around a common framework for evaluation and assessment. With 11 states and 160 programs volunteering the time of their staff and students, we have shown that it is possible to apply one set of data collection tools on a large scale to inform both research and practice. System-building benefits from a standard set of reliable and valid data collection tools because they help to establish benchmarks that measure success and to enable communication across states and programs. If states and programs chose to use different tools and methodology, it would be impossible to relate the data and learn from one another. Programs should be encouraged at the state level to continue to gather data for at least two purposes. First, programs can use data to learn from themselves and improve. Observation and survey tools used as part of this evaluation can be used to inform on everyday programming and to inform on professional development needs. Second, states can encourage programs to work together collectively to pool data that will identify strengths and challenges on a city and state level. States and programs can learn from each other and can also be motivated by the progress made by others. Importantly, by working together, it is possible to determine whether system-building is having an impact more broadly, such as improving future science achievement and science career attainment among youth. Thus, we recommend that states and programs continue to use a common framework with common data-creating tools to facilitate communication between partners, to continuously improve

practices, and to advocate for policy changes. There are systems in place. For instance, The PEAR Institute supports evaluation and assessment by consulting on methodology, setting up data collection protocols, and providing actionable feedback on a program, organization, community, city, or state level (depending on what is needed). PEAR is currently working with technology partners to create a data and technical support hub for the STEM field. This hub, "PEAR Data Central," will analyze data collected by program staff and evaluators and provide actionable reports for student and program improvement to the practitioners involved. It would also aggregate data from thousands of individual programs to allow comparisons across geographic regions, program types, dosages, quality, and other variations in STEM program characteristics.

5. **Innovate out-of-school time evaluation and assessment strategies:** This evaluation was innovative in a number of ways. We created a data collection infrastructure through the use of inexpensive portable devices that we left behind with states and programs for use as educational platforms. This is the beginning of a national data collection infrastructure in a field that is very much in need of resources. With 11 states now equipped, and having proven that this method works in even the most remote of programs, we have created a system that not only can be replicated in future studies but can also be used if we return to the same programs and states in the future. Additionally, it was possible to reduce the burden of response for facilitators and students by using an imputation method that made each questionnaire about 25% shorter without losing the overall response use for the entire list of questions. Finally, and most importantly, we were pushing against the traditional pretest-posttest design that has great limitations. We wanted to see whether the retrospective pretest-posttest design would yield differences across states and be sensitive enough to show differences in outcome depending on states, programs, and especially levels of quality or lack thereof. While no method is perfect, and we will need more studies of the kind we present here, we recommend that the field be open to more innovative methods to gain a better understanding of outcome than the traditional methods make possible. We realize this to be a controversial statement, but the fact is that the traditional method of asking a student at two time points the same questions can be alienating ("Why am I asked the same questions?") and thus demotivating the second time around. It is also time consuming to administer a survey twice, and staff have the added responsibility of tracking students' survey responses, which makes it impossible to keep the survey responses anonymous. There also remains the issue of response-shift bias described previously that needs to be addressed. It is mentioned frequently as one reason why even very strong programs might show slight, if any, growth in attitudinal measures. Now is the time to really experiment and address this issue head on. We need to bring the same level of inquiry and reflection that we hope to instill in students to our own methods for data collection. The retrospective pretest-posttest has grown sufficiently as a valid method to have made it feasible to use in this large-scale evaluation. Additionally, triangulation of student, facilitator, and program data as a method, as we have used in this evaluation, is something we and others believe is an essential part of methods use. In future studies, we recommend using objective growth indicators through embedded assessments (e.g., computer games for STEM that measure proficiency) and student product analyses. The fact that we found that the retrospective pretest-posttest results varied by states, settings, quality, and other indicators is a very exciting finding that should be capitalized on in many future studies as evaluators and researchers are allowed to experiment further. In the future we will need to pursue quasi-experimental designs and randomized controlled trials. However, the field must first use the results from this and other studies to further strengthen interventions, professional development strategies, and the choice of strong curricula.

6. **Prioritize evaluation in the system-building process:** Our final recommendation is that the evaluative data collection and data reporting effort becomes a priority of this system-building initiative. We believe this can become a model in which large-scale projects in many different educational venues can monitor themselves over time. Implicit in the present evaluation is the use of data for multiple purposes: to track successes and challenges in each state, to use the results for management purposes (e.g., where and how to invest professional development and coaching efforts, how to assess the results of curricula and staff selection), and to expand the advocacy and policy efforts based on evidence. On a national level, these data allow this evaluation, as well as future studies, to compare states, program types, and professional development investments as they relate to STEM quality and student outcomes. This information should present the current and future funders with a targeted approach to provide support to elements of the system-building initiative that need improvement or to provide support for the roll out for those elements that have shown to be especially successful. It will also be possible to explore those states and programs that have especially strong quality and outcomes to learn from them, and to possibly take these successes and generalize them across the system.

REFERENCES

Allen, J. M., & Nimon, K. (2007). Retrospective pretest: A practical technique for professional development evaluation. *Journal of Industrial Teacher Education*, *44*, 27–42.

Aschbacher, P. R., Ing, M., & Tsai, S. M. (2014). Is science me? Exploring middle school students' STE-M career aspirations. *Journal of Science Education and Technology*, *23*(6), 735–743.

Bell, C. A., Qi, Y., Croft, A. C., Leusner, D., Gitomer, D. H., McCaffrey, D. F., & Pianta, R. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Measures of effective teaching* (pp. 50–97). San Francisco, CA: Jossey-Bass.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314.

Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. *Educational and Psychological Measurement*, *44*, 781–803.

Cribbs, J. D., Hazari, Z., Sonnert, G., & Sadler, P. M. (2015). Establishing an explanatory model for mathematics identity. *Child Development*, *86*(4), 1048–1062.

Dabney, K., Tai, R. H., Almarode, J. T., Miller-Friedman, J. L., Sonnert, G., Sadler, P. M., & Hazari, Z. (2011). Out of school time science activities and their association with career interest in STEM. *International Journal of Science Education*, *Part B: Communication and Public Engagement*, 2(1), 63-79.

Desy, E.A., Peterson, S.A., & Brockman, V. (2011). Gender differences in science-related attitudes and interests among middle school and high school students. Science Educator, *20*(2), 23–30.

Drennan, J., & Hyde, A. (2008). Controlling response shift bias: The use of the retrospective pre-test design in the evaluation of a master's programme. *Assessment & Evaluation in Higher Education*, *33*, 699–709.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fabes, R. A., Hayford, S., Pahlke, E., Santos, C., Zosuls, K., Martin, C. L., & Hanish, L. D. (2014). Peer influences on gender differences in educational aspiration and attainment. In I. Schoon & J. S. Eccles (Eds.), *Gender differences in aspirations and attainment: A life course perspective* (pp. 29–52). Cambridge, UK: Cambridge University Press.

Finkelstein, J. A., Quaranto, B. R., & Schwartz, C. E. (2014). Threats to the internal validity of spinal surgery outcome assessment: Recalibration response shift or implicit theories of change? *Applied Research Quality Life*, *9*, 215–232.

Gorrall, B. K., Curtis, J. D., Little, T. D., & Panko, P. (2016). Innovations in measurement: Visual analog scales and retrospective pretest self-report designs. *Actualidades en Psicologia*, *30*, 1–6.

Hattie, J. (2009). *Visible learning*. London, UK: Routledge.

Hill, L. G., & Bentz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation*, *26*, 501–507.

Howard, G. S. (1980). Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Evaluation Review*, *4*, 93–106.

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*, 144–150.

Howard, W. J., Rhemtulla, M., & Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioural Research*, *50*(3), 285–299.

Hughes, J. N., Luo, W., Kwok, O., & Loyd L. K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology*, *100*, 1–14.

Jones, M. G., Howe, A., & Rua, M. (2000). Gender differences in students' experiences, interests, and attitudes toward science and scientists. *Science Education*, *84*, 180–192.

Kreulen, G. J., Stommel, M., Gutek, B. A., Burns, L. R., & Braden, C. J. (2002). Utility of retrospective pretest ratings of patient satisfaction with health status. *Research in Nursing & Health*, *25*, 233–241.

Lam, T. C., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation*, *24*, 65–80.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Little, T. D., Chang, R., Gorrall, B. K., Fukuda, E., Waggenspack, L., Noam, G., & Allen, P. (in preparation). The retrospective pretest-posttest design redux: On its validity as an alternative to traditional pre-post measure. Institute for Measurement, Methodology, Analysis, & Policy, Lubbock, TX, and The PEAR Institute: Partnerships in Education and Resilience, Belmont, MA.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*, 151–162.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, 7, 199–204.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226.

Martinez, A., Linkow, T., & Velez, M. (2014). *Evaluation study of Summer of Innovation stand-alone program model FY2013: Outcomes report*. Retrieved from http://www.nasa.gov/sites/default/files/soi_stand-alone_program_model_fy2013_outcome_report.pdf

Miller, M., & Elder Hinshaw, R. (2012). The retrospective pretest as a gauge of change. *Journal of Instructional Psychology*, *39*, 251–258.

Moberg, D. P., & Finch, A. J. (2007). Recovery high schools: A descriptive study of school programs and students. *Journal of Groups in Addiction & Recovery*, *2*, 128–161.

Moore, D., & Tananis, C. A. (2009). Measuring change in a short-term educational program using a retrospective pretest design. *American Journal of Evaluation*, *30*, 189–202.

Nieuwkerk, P. T., & Sprangers, M. G. (2009). Each measure of patient-reported change provides useful information and is susceptible to bias: The need to combine methods to assess their relative validity. *Arthritis & Rheumatism*, *61*, 1623–1625.

Noam, G.G., Allen, P.J., Sonnert, G., & Sadler, P. (in preparation). Validation of the Common Instrument: A Brief measure for assessing science interest in children and youth. The PEAR Institute: Partnerships in Education and Resilience. Belmont, MA.

Noam, G., Malti, T., & Guhn, M. (2012). From clinical-developmental theory to assessment: The Holistic Student Assessment tool. *International Journal of Conflict and Violence*, *6*(2), 201–2013.

Noam, & Shah, A. M. (2013). *Game-changers and the assessment predicament in afterschool science*. Retrieved from http://www.pearweb.org/research/pdfs/Noam%26Shah_Science_Assessment_Report.pdf

Organization for Economic Cooperation and Development. (2010). PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science (Vol. I). Retrieved from http://dx.doi.org/10.1787/9789264091450-en

Osborne, J., Simon, S., & Collins, S. (2010). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079.

Papazian, A. E., Noam, G. G., Shah, A. M., & Rufo-McCormick, C. (2013). The quest for quality in afterschool science. Retrieved from http://www.niost.org/pdf/afterschoolmatters/asm_2013_18_fall/Pages%20from%20ASM_Fall2013- Papazian%20NoamShahMcCormick.pdf

Potvin, P., & Hasni, A. (2014). Analysis of the decline in interest towards school science and technology from grades 5 through 11. *Journal of Science Education and Technology*, *23(6)*, 784-802.

Pratt, C. C., McGuigan. W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. *American Journal of Evaluation*, *21*, 341–349.

Rhemtulla, M., Savalei, V., & Little, T. D. (2016). On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*, *81*, 60–89.

Richardson, J. T. E. (2010). Eta squared and partial eta squared as measures of effect size in education research. *Educational Research Review*, *6*, 135–147.

Rose, A. J., & Rudolph, K. D. (2006). A review of sex differences in peer relationship processes: potential trade-offs for the emotional and behavioral development of girls and boys. *Psychological bulletin*, *132*(1), 98.

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 183–212). Hillsdale, NJ: Erlbaum.

Schwartz, C. E., & Sprangers, M. A. G. (2010). Guidelines for improving the stringency of response shift research using the ntest. *Quality Life Research*, *19*, 455–464.

Schwartz, C. E., Sprangers, M. A. G., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology and Health*, *19*, 51–69.

Shah, A. M., Wylie, C. E., Gitomer, D., & Noam, G.G. (2016). *Development of the Dimensions of Success (DoS) observation tool for the out of school time STEM field: Refinement, field-testing and establishment of psychometric properties.* Retrieved from http://www.pearweb.org/research/pdfs/ DoSTechReport_092314_final.pdf

Sprangers, M. A. G., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of Applied Psychology*, *74*, 265-272.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.

VanLeuvan, P. (2004). Young women's science/mathematics career goals from seventh grade to high school graduation. *Journal of Educational Research*, *97*(5), 248–268.

Weinburgh, M. (1995). Gender differences in student attitudes toward science: A meta-analysis of the literature from 1970 to 1991. *Journal of Research in Science Teaching*, *32*(4), 387–398.

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, *50*(5), 1080–1121.

Wang, M. T., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement and academic achievement in middle school. *American Educational Research Journal*, *47*, 633–662.

Watt, H. M. G., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., & Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: A Comparison of samples from Australia, Canada, and the United States. *Developmental Psychology*, *48*(6), 1594–1611.

APPENDICES

Appendix A

CORRELATIONS BETWEEN RETROSPECTIVE DIFFERENCES SCORES AND PROSPECTIVE DIFFERENCE SCORES

Table A1. Correlations Between Retrospective Differences Scores and Prospective Difference Scores

	Science Interest_RPD	Science Identity_RPD	Science Career Knowledge_RPD	Science Career Interest_RPD	Science Activity Participation_RPD	Relationships With Adults_RPD	Relationships With Peers_RPD	Perseverance_RPD
Science Interest_RPD	1.000							
Science Identity_RPD	0.896	1.000						
Science Career Knowledge_RPD	0.858	0.848	1.000					
Science Career Interest_RPD	0.906	0.884	0.868	1.000				
Science Activity Participation_RPD	0.796	0.753	0.727	0.768	1.000			
Relationships With Adults_RPD	0.859	0.801	0.782	0.835	0.705	1.000		
Relationships With Peers_RPD	0.815	0.802	0.780	0.811	0.679	0.758	1.000	
Perseverance_RPD	0.909	0.873	0.836	0.882	0.759	0.835	0.804	1.000
Critical Thinking_RPD	0.839	0.806	0.757	0.824	0.687	0.781	0.769	0.807
Science Interest_PD	0.056	0.042	0.048	0.049	0.070	0.038	0.025	0.035
Science Identity_PD	0.050	0.034	0.030	0.040	0.088	0.038	0.015	0.036
Science Career Knowledge_PD	0.028	0.020	0.025	0.019	0.073	0.016	0.000	0.016
Science Career Interest_PD	0.041	0.025	0.035	0.037	0.072	0.024	0.014	0.024
Science Activity Participation_PD	0.085	0.073	0.069	0.070	0.124	0.059	0.059	0.061
Relationships With Adults_PD	0.056	0.054	0.030	0.021	0.063	0.017	0.030	0.028
Relationships With Peers_PD	0.002	0.030	0.040	0.015	-0.026	-0.006	0.019	-0.011
Perseverance_PD	0.012	0.005	0.005	0.007	0.016	-0.002	-0.001	0.004
Critical Thinking_PD	-0.018	-0.033	-0.006	-0.026	-0.020	-0.027	-0.032	-0.036

	Critical Thinking_RPD	Science Interest_PD	Science Identity_PD	Science Career Knowledge_PD	Science Career Interest_PD	Science Activity Participation_PD	Relationships With Adults_PD	Relationships With Peers_PD	Perseverance_PD	Critical Thinking_PD
	1.000									
	0.037	1.000								
	0.026	0.806	1.000							
	0.002	0.699	0.851	1.000						
	0.024	0.855	0.882	0.832	1.000					
	0.059	0.680	0.822	0.823	0.810	1.000				
	0.032	0.505	0.426	0.287	0.371	0.280	1.000			
	0.026	0.303	0.226	0.071	0.237	0.086	0.556	1.000		
	0.006	0.561	0.622	0.570	0.520	0.413	0.533	0.379	1.000	
	-0.024	0.627	0.556	0.461	0.559	0.340	0.524	0.586	0.637	1.000

Appendix B

CIS ANALYSIS RESULTS FOR TABLES FOR RETROSPECTIVE PRETEST-POSTTEST

Table B1. Paired Sample *t*-Test Results and Proportion of Changes Across 11 States (For students in Grades 4–12)

Florida (*n* = 122)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	64.964	17.859	70.003	20.179	4.076	p < 0.001	0.264	69.7%	1.6%	28.7%
Science Identity	55.457	20.672	59.305	20.674	3.676	p < 0.001	0.186	64.8%	3.3%	32.0%
Science Career Knowledge	50.146	20.038	55.785	20.268	4.254	p < 0.001	0.280	68.0%	4.9%	27.0%
Science Career Interest	56.742	19.415	61.507	19.998	3.789	p<0.001	0.242	66.4%	1.6%	32.0%
Science Activity Participation	39.957	21.234	44.594	22.372	4.430	p<0.001	0.213	70.5%	2.5%	27.0%
Relationships With Adults	68.396	18.172	71.287	19.329	2.251	0.026	0.154	63.1%	2.5%	34.4%
Relationships With Peers	76.182	16.119	78.597	16.271	2.106	0.037	0.149	59.0%	6.6%	34.4%
Perseverance	70.697	18.212	74.833	18.305	2.907	0.004	0.226	63.1%	4.1%	32.8%
Critical Thinking	71.419	19.101	77.071	18.631	4.489	p < 0.001	0.300	66.4%	0.8%	32.8%

Iowa (*n* = 137)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	59.617	20.105	71.369	16.949	9.216	p < 0.001	0.632	83.9%	2.2%	13.9%
Science Identity	51.534	23.028	59.950	21.178	7.838	p < 0.001	0.380	70.8%	6.6%	22.6%
Science Career Knowledge	43.433	22.519	52.594	21.867	8.653	p < 0.001	0.413	75.9%	7.3%	16.8%
Science Career Interest	50.350	22.005	60.040	21.055	8.669	p < 0.001	0.450	80.3%	4.4%	15.3%
Science Activity Participation	35.096	21.232	44.714	22.005	8.750	p < 0.001	0.445	78.8%	5.8%	15.3%
Relationships with Adults	62.723	19.758	71.182	17.479	6.749	p < 0.001	0.453	70.1%	5.1%	24.8%
Relationships with Peers	69.769	19.200	76.875	16.365	6.300	p < 0.001	0.398	65.0%	7.3%	27.7%
Perseverance	66.512	19.154	76.725	14.849	8.399	p < 0.001	0.596	78.8%	4.4%	16.8%
Critical Thinking	69.301	19.008	78.629	13.849	7.975	p < 0.001	0.561	74.5%	4.4%	21.2%

Table B1. Continued

Indiana (*n* = 169)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	59.709	22.259	71.028	20.336	9.689	p < 0.001	0.531	81.7%	4.7%	13.6%
Science Identity	49.451	24.918	58.915	23.530	8.864	p < 0.001	0.391	80.5%	3.6%	16.0%
Science Career Knowledge	46.243	24.342	58.953	21.115	10.223	p < 0.001	0.558	85.2%	2.4%	12.4%
Science Career Interest	49.927	24.839	60.600	22.929	9.290	p < 0.001	0.447	79.9%	4.1%	16.0%
Science Activity Participation	36.859	23.074	46.430	23.124	10.097	p < 0.001	0.414	81.7%	3.0%	15.4%
Relationships With Adults	63.626	22.449	72.407	19.250	7.871	p < 0.001	0.420	78.7%	3.6%	17.8%
Relationships With Peers	71.067	21.697	77.359	18.131	6.064	p < 0.001	0.315	66.9%	5.9%	27.2%
Perseverance	66.245	22.913	76.671	18.165	8.206	p < 0.001	0.504	74.6%	7.1%	18.3%
Critical Thinking	67.442	22.855	77.764	17.784	7.876	p < 0.001	0.504	73.4%	5.3%	21.3%

Kansas (*n* = 90)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	58.976	20.179	67.010	20.856	6.108	p < 0.001	0.392	72.2%	2.2%	25.6%
Science Identity	47.185	21.856	54.316	23.307	5.338	p < 0.001	0.316	70.0%	4.4%	25.6%
Science Career Knowledge	41.867	20.026	51.096	21.552	6.780	p < 0.001	0.444	77.8%	4.4%	17.8%
Science Career Interest	48.418	20.926	55.852	23.006	5.422	p < 0.001	0.338	73.3%	4.4%	22.2%
Science Activity Participation	33.745	19.616	41.730	22.954	6.277	p < 0.001	0.374	74.4%	1.1%	24.4%
Relationships With Adults	67.111	18.797	72.448	18.992	4.361	p < 0.001	0.282	68.9%	6.7%	24.4%
Relationships With Peers	73.585	16.579	78.476	17.618	3.392	0.001	0.286	66.7%	3.3%	30.0%
Perseverance	66.967	17.997	75.089	18.182	5.568	p < 0.001	0.449	73.3%	3.3%	23.3%
Critical Thinking	69.185	18.419	76.357	19.355	5.519	p < 0.001	0.380	76.7%	3.3%	20.0%

Table B1. Continued

Massachusetts (*n* = 220)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	57.621	19.491	66.550	20.408	8.343	p < 0.001	0.447	75.9%	2.3%	21.8%
Science Identity	47.201	23.051	54.328	23.421	7.701	p < 0.001	0.307	71.4%	3.2%	25.5%
Science Career Knowledge	40.948	21.879	51.511	22.162	10.263	p < 0.001	0.480	80.9%	2.7%	16.4%
Science Career Interest	48.794	21.973	56.638	23.143	7.911	p < 0.001	0.348	71.4%	3.6%	25.0%
Science Activity Participation	35.591	20.727	42.369	21.981	7.411	p < 0.001	0.317	70.5%	3.2%	26.4%
Relationships With Adults	59.815	20.795	65.942	21.044	6.073	p < 0.001	0.293	65.5%	5.9%	28.6%
Relationships With Peers	73.609	18.273	78.064	17.320	5.328	p < 0.001	0.250	62.3%	6.8%	30.9%
Perseverance	65.246	21.127	72.128	19.827	6.252	p < 0.001	0.336	68.2%	3.2%	28.6%
Critical Thinking	66.349	20.063	74.416	18.778	7.590	p < 0.001	0.415	71.8%	3.2%	25.0%

Maryland (*n* = 172)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	64.825	19.447	72.545	18.906	6.851	p < 0.001	0.402	71.5%	5.2%	23.3%
Science Identity	56.835	23.462	63.037	22.306	5.793	p < 0.001	0.271	66.9%	2.9%	30.2%
Science Career Knowledge	52.175	22.581	62.743	19.740	8.710	p < 0.001	0.498	77.3%	3.5%	19.2%
Science Career Interest	56.350	22.402	64.385	21.038	7.551	p < 0.001	0.370	73.3%	5.8%	20.9%
Science Activity Participation	41.904	22.130	50.052	21.897	7.958	p < 0.001	0.370	75.0%	3.5%	21.5%
Relationships With Adults	66.642	20.128	73.023	18.767	6.196	p < 0.001	0.328	66.3%	8.1%	25.6%
Relationships With Peers	74.369	17.660	78.014	15.360	3.792	p < 0.001	0.220	54.7%	7.0%	38.4%
Perseverance	71.753	19.990	78.364	17.290	5.399	p < 0.001	0.354	68.0%	4.1%	27.9%
Critical Thinking	71.864	18.477	77.930	15.775	5.536	p < 0.001	0.353	62.2%	6.4%	31.4%

Table B1. Continued

Michigan (*n* = 99)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	63.728	17.773	73.042	17.105	7.132	p < 0.001	0.534	77.8%	2.0%	20.2%
Science Identity	52.127	22.643	60.919	22.603	7.435	p < 0.001	0.389	76.8%	6.1%	17.2%
Science Career Knowledge	51.132	21.488	64.131	20.043	9.329	p < 0.001	0.626	85.9%	2.0%	12.1%
Science Career Interest	54.949	21.323	63.692	21.268	6.960	p < 0.001	0.411	78.8%	2.0%	19.2%
Science Activity Participation	39.049	20.368	49.199	21.930	8.818	p < 0.001	0.480	82.8%	3.0%	14.1%
Relationships With Adults	63.278	18.155	70.445	16.568	5.454	p < 0.001	0.412	72.7%	0.0%	27.3%
Relationships With Peers	72.985	18.433	78.598	15.588	5.523	p < 0.001	0.329	66.7%	2.0%	31.3%
Perseverance	69.041	18.507	77.800	15.554	6.641	p < 0.001	0.512	80.8%	3.0%	16.2%
Critical Thinking	71.696	18.022	79.794	14.553	6.156	p < 0.001	0.494	73.7%	7.1%	19.2%

Nebraska (*n* = 115)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	59.607	19.744	71.272	19.092	10.174	p < 0.001	0.601	88.7%	3.5%	7.8%
Science Identity	51.061	21.403	61.070	21.120	9.808	p < 0.001	0.471	86.1%	3.5%	10.4%
Science Career Knowledge	45.304	20.657	57.628	20.350	10.926	p < 0.001	0.601	87.0%	0.0%	13.0%
Science Career Interest	52.713	21.033	62.881	20.833	9.167	p < 0.001	0.486	82.6%	2.6%	14.8%
Science Activity Participation	37.876	20.911	47.173	22.049	8.180	p < 0.001	0.433	79.1%	1.7%	19.1%
Relationships With Adults	65.205	18.293	73.913	16.833	7.766	p < 0.001	0.495	79.1%	3.5%	17.4%
Relationships With Peers	75.454	16.725	82.803	13.750	6.717	p < 0.001	0.480	74.8%	3.5%	21.7%
Perseverance	68.121	17.996	78.269	15.163	7.397	p < 0.001	0.610	78.3%	7.0%	14.8%
Critical Thinking	68.819	17.533	79.469	14.694	9.260	p < 0.001	0.658	80.0%	5.2%	14.8%

Table B1. Continued

Oregon (*n* = 134)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	58.066	19.184	66.990	20.229	7.945	p < 0.001	0.453	76.1%	6.0%	17.9%
Science Identity	45.515	20.993	51.427	22.580	5.912	p < 0.001	0.271	72.4%	4.5%	23.1%
Science Career Knowledge	40.136	20.170	48.847	21.850	7.201	p < 0.001	0.414	77.6%	4.5%	17.9%
Science Career Interest	46.767	20.518	53.864	22.746	6.507	p < 0.001	0.328	73.1%	0.7%	26.1%
Science Activity Participation	33.521	19.274	41.713	21.845	8.035	p < 0.001	0.398	70.9%	6.0%	23.1%
Relationships With Adults	62.056	17.367	69.226	17.232	6.573	p < 0.001	0.414	70.9%	5.2%	23.9%
Relationships With Peers	71.236	15.334	75.241	14.920	4.014	p < 0.001	0.265	62.7%	1.5%	35.8%
Perseverance	64.566	17.977	72.888	17.480	6.783	p < 0.001	0.469	69.4%	4.5%	26.1%
Critical Thinking	65.216	18.571	72.323	18.203	6.096	p < 0.001	0.387	72.4%	3.0%	24.6%

Pennsylvania (*n* = 161)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	56.672	19.881	68.999	19.064	9.147	p < 0.001	0.633	78.3%	4.3%	17.4%
Science Identity	44.897	21.228	55.123	21.227	8.519	p < 0.001	0.482	72.7%	3.7%	23.6%
Science Career Knowledge	41.128	21.304	52.853	20.305	9.069	p < 0.001	0.563	78.3%	1.2%	20.5%
Science Career Interest	46.665	20.816	57.571	21.076	8.798	p < 0.001	0.521	77.6%	1.2%	21.1%
Science Activity Participation	32.428	17.560	44.082	21.395	9.319	p < 0.001	0.595	77.6%	5.6%	16.8%
Relationships With Adults	63.569	19.397	73.494	17.952	7.086	p < 0.001	0.531	77.0%	3.1%	19.9%
Relationships With Peers	70.138	19.012	77.567	16.142	5.902	p < 0.001	0.421	73.9%	1.2%	24.8%
Perseverance	64.362	21.311	76.189	16.115	8.286	p < 0.001	0.626	75.8%	1.9%	22.4%
Critical Thinking	64.688	20.709	76.168	17.082	8.331	p < 0.001	0.605	76.4%	3.1%	20.5%

South Carolina (*n* = 180)

Variable	Retro Pre		Post		*t*-test		Effect Size	Proportion of Changes		
	M	*SD*	*M*	*SD*	*t*	*p*	*d*	*Positive*	*Neutral*	*Negative*
Science Interest	60.174	19.782	72.121	17.121	10.337	p < 0.001	0.646	77.8%	3.3%	18.9%
Science Identity	51.438	22.634	60.842	20.239	9.261	p < 0.001	0.438	73.9%	3.9%	22.2%
Science Career Knowledge	44.618	24.104	57.276	21.750	10.452	p < 0.001	0.551	82.2%	2.2%	15.6%
Science Career Interest	51.297	22.459	62.180	20.668	10.208	p < 0.001	0.504	77.8%	7.8%	14.4%
Science Activity Participation	35.995	23.215	46.461	23.885	10.490	p < 0.001	0.444	83.3%	2.2%	14.4%
Relationships With Adults	66.968	19.530	76.945	14.950	8.517	p < 0.001	0.574	71.1%	7.2%	21.7%
Relationships With Peers	74.406	17.740	81.561	14.740	6.810	p < 0.001	0.439	62.2%	8.9%	28.9%
Perseverance	68.171	20.874	79.673	15.395	9.418	p < 0.001	0.627	71.7%	8.9%	19.4%
Critical Thinking	68.418	19.679	79.924	15.524	10.154	p < 0.001	0.649	77.2%	4.4%	18.3%

Table B2. ANOVA Results for Gender Effects (For students in Grades 4–12)

Variable	Female (N=733)				Male (N=866)									Effect Size	
	Retro	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*F (Main)*	*p*	*F (Inx)*	*p*	*df*	*Partial η2 (Main)*	*Partial η2 (Inx)*
Science Interest	56.714	19.727	67.163	19.786	63.103	19.517	72.443	18.460	1.678	0.195			1	0.001	
Science Identity	48.145	22.581	56.553	22.391	51.868	22.798	59.360	22.120	0.110	0.740			1	η2 > 0.001	
Science Career Knowledge	44.055	22.209	54.852	21.756	45.780	22.393	56.387	21.271	0.141	0.708			1	η2 > 0.001	
Science Career Interest	49.859	21.916	59.159	22.241	51.976	22.063	60.490	21.538	0.266	0.606			1	η2 > 0.001	
Science Activity Participation	35.018	21.092	43.988	22.77	37.797	21.146	46.422	22.082	0.001	0.979			1	η2 > 0.001	
Relationships With Adults	66.458	18.756	74.081	17.485	62.362	20.366	69.752	18.996	6.172	0.013	1.516	0.218	1	0.004	0.001
Relationships With Peers	75.584	17.312	80.863	15.485	70.672	18.615	76.365	16.496	5.959	0.015	0.091	0.763	1	0.004	η2 > 0.001
Perseverance	67.921	19.733	76.976	16.667	66.771	20.419	75.471	17.738	0.000	0.991			1	η2 > 0.001	
Critical Thinking	67.121	19.518	76.331	17.262	69.366	19.729	77.821	16.752	0.079	0.779			1	η2 > 0.001	

Table B3. ANOVA Results for Grade Levels (Grades 4–12)

Variable	4th (n=364)				5th (n=353)				6th (n=411)				7th (n=271)	
	Retro	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*Retro*	*SD*
Science Interest	60.011	19.467	69.440	18.459	60.653	20.334	71.184	19.944	59.909	20.734	69.438	20.082	59.559	19.166
Science Identity	51.053	21.874	58.359	20.710	50.650	23.008	59.161	22.611	50.061	23.997	57.48	23.474	48.327	21.607
Science Career Knowledge	44.616	22.504	55.170	20.248	45.836	22.443	56.597	22.015	44.885	23.207	54.819	22.685	42.936	20.700
Science Career Interest	51.744	21.008	60.085	20.146	52.109	21.961	61.716	21.877	51.215	23.305	59.392	23.382	48.164	20.742
Science Activity Participation	37.064	21.278	45.918	21.647	37.462	20.933	46.678	22.460	36.517	22.39	44.731	23.421	33.545	18.989
Relationships With Adults	65.336	19.625	71.895	18.446	68.343	19.221	75.524	17.614	62.874	19.803	70.777	17.987	59.729	19.933
Relationships With Peers	73.675	17.975	78.592	16.029	74.585	18.334	81.016	15.574	72.632	18.605	77.766	16.498	69.881	17.494
Perseverance	67.057	19.998	75.360	17.513	70.145	21.290	79.111	17.694	66.652	19.747	75.423	17.017	63.854	19.683
Critical Thinking	67.142	20.080	75.993	17.336	69.955	20.673	79.395	17.891	67.856	19.541	76.214	16.854	67.035	18.530

Table B4. ANOVA Results for Effects of Program Type

Variable	School-based (n = 1109)				Center-based (n = 428)				Other (n = 62)						
	Retro	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*F*	*df*	*p*
Science Interest	60.192	19.579	70.79	18.563	59.834	20.707	68.406	20.629	62.208	19.181	67.462	20.929	6.694	2	0.001
Science Identity	49.869	22.527	58.273	21.919	50.275	23.707	57.587	23.262	54.612	20.137	57.857	22.117	4.040	2	0.018
Science Career Knowledge	44.783	22.248	56.316	21.064	45.379	23.028	45.379	23.028	45.980	18.520	52.23	20.855	6.979	2	0.001
Science Career Interest	50.417	21.858	59.939	21.443	52.021	22.923	59.818	23.17	54.532	17.682	59.257	20.379	3.861	2	0.021
Science Activity Participation	35.453	21.013	44.634	22.131	38.856	21.794	47.042	23.361	39.554	17.776	45.345	20.706	1.617	2	0.120
Relationships With Adults	62.655	19.739	71.139	18.149	67.78	19.469	73.463	19.052	68.148	18.242	70.508	18.84	3.973	2	0.020
Relationships With Peers	71.812	18.448	77.76	16.122	75.817	17.319	80.626	15.836	72.824	17.463	75.183	18.403	2.670	2	0.070
Perseverance	66.463	20.360	76.265	16.799	69.134	19.723	76.145	18.276	69.558	17.294	74.406	18.492	4.083	2	0.017
Critical Thinking	68.194	19.663	77.773	16.372	68.591	19.91	75.924	18.048	69.125	18.016	74.172	19.855	6.740	2	0.001

* denotes program type with stronger effect

			8th (n=115)				9th - 12th (n=85)							Effect Size
	Post	*SD*	*Retro*	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*F*	*df*	*p*	*Partial η2*
	69.464	18.180	60.082	19.359	68.93	19.685	62.252	18.488	73.783	18.183	1.073	5	0.374	0.003
	56.346	21.111	51.641	23.103	58.754	23.912	48.646	22.767	59.790	23.032	1.445	5	0.205	0.005
	53.881	20.355	47.914	20.953	58.443	21.091	46.155	23.34	60.284	22.242	1.639	5	0.147	0.005
	57.455	20.769	53.097	22.401	60.916	22.695	48.477	23.022	60.066	23.372	1.157	5	0.328	0.004
	41.767	20.839	39.168	21.023	46.916	22.886	36.251	21.822	48.874	24.013	2.176	5	0.054	0.007
	67.454	19.097	65.273	18.593	72.197	18.344	62.095	19.658	72.988	19.086	1.897	5	0.092	0.006
	75.547	15.959	74.422	18.294	78.648	18.055	71.879	17.699	79.049	14.612	2.196	5	0.052	0.007
	73.133	16.474	69.636	18.004	76.993	16.891	67.442	19.864	79.432	16.662	1.152	5	0.331	0.004
	75.970	15.509	69.783	18.632	77.538	16.252	71.250	18.561	80.322	17.108	1.412	5	0.217	0.004

	Effect Size*	
	Partial η2	*Tukey's HSD*
	0.008	School*-Center; School*-Other
	0.005	School*-Other
	0.009	School*-Center; School*-Other
	0.005	No significant differencs found
	0.002	
	0.005	School*-Other
	0.003	
	0.005	No significant differencs found
	0.008	School*-Center; School*-Other

Table B5. ANOVA Results for Effects of Program Duration

Variable	< 1 Week (n = 310)				1-3 Weeks (n= 272)				4-8 Weeks (n = 346)				> 8 Weeks (n = 670)			
	Retro	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*	*Retro*	*SD*	*Post*	*SD*
Science Interest	57.307	20.106	61.006	21.222	62.460	19.103	69.078	19.257	58.356	19.102	70.733	17.674	61.509	20.245	74.213	17.600
Science Identity	45.582	22.306	47.427	22.886	52.169	22.718	57.383	22.747	48.285	22.281	58.694	20.996	52.405	22.900	62.960	20.741
Science Career Knowledge	41.531	21.988	44.502	21.531	47.529	21.911	55.056	21.715	43.249	21.655	57.378	19.803	46.420	22.750	60.231	20.412
Science Career Interest	45.670	21.459	48.230	22.608	53.915	21.49	60.305	21.677	49.510	21.179	60.394	20.269	53.044	22.455	64.838	20.397
Science Activity Participation	31.410	20.520	34.129	21.541	39.091	21.409	45.693	22.377	34.835	20.440	45.438	21.852	38.691	21.270	50.248	21.342
Relationships With Adults	63.900	19.729	65.694	20.217	67.172	17.434	72.347	17.128	63.185	19.623	72.450	17.819	63.771	20.618	73.956	17.818
Relationships With Peers	73.850	18.093	74.397	18.086	75.439	16.114	78.914	15.274	73.202	17.545	80.100	14.778	71.323	19.223	79.265	16.043
Perseverance	68.985	19.646	71.075	19.630	71.264	17.368	76.946	16.156	65.482	19.694	77.250	15.815	65.822	21.296	77.646	16.858
Critical Thinking	67.986	21.389	71.722	20.234	71.54	17.454	77.51	16.096	66.521	18.977	78.255	14.894	68.123	19.914	78.936	16.247

* denotes program type with stronger effect

	Group difference			Effect Size	
	F	*df*	*p*	*Partial η2*	*Tukey's HSD*
	51.307	3	p < 0.001	0.088	<WK:1-3* ; <WK:4-8*; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	57.713	3	p < 0.001	0.098	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	71.730	3	p < 0.001	0.119	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	56.506	3	p < 0.001	0.096	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	49.609	3	p < 0.001	0.085	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	33.903	3	p < 0.001	0.060	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	27.577	3	p < 0.001	0.049	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	35.326	3	p < 0.001	0.062	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*
	29.656	3	p < 0.001	0.053	<WK:1-3* ; <WK:4-8* ; <WK:>8WKS* 1-3:4-8* ; 1-3:>8WKS*

Appendix C

DIMENSIONS OF SUCCESS (DoS) RESULTS

Table C1. Average Sum Score for Programs Defined as Low, Average, or High Quality for Each DoS Domain

	Dimensions of Success: Separate Domains			
Quality Rating	**Features of the Learning Environment**	**Activity Engagement**	**STEM Knowledge & Practices**	**Youth Development in STEM**
Lower	8.0 (Range: 1.0-9.0)	5.9 (Range: 3.0-7.0)	4.2 (Range: 1.0 - 5.5)	5.5 (Range: 3.5-6.5)
Average	10.7 (Range: 9.1-11.5)	9.3 (Range: 7.1-11.0)	7.7 (Range: 5.6-9.50)	8.6 (Range: 6.6-10.0)
Higher	12.0 (Range: 11.6-12)	11.6 (Range: 11.5-12.0)	10.6 (Range: 9.51-12.0)	11.2 (Range: 10.5-12.0)

It is important to understand the meaning of the quantitative data collected using the DoS tool. The table above provides benchmark information for each of the four DoS domains so that individual state networks or programs can understand if their ratings fall within the range of lower, average or higher program quality. These ranges are based on the 252 observations performed in programs across 11 states. Note that it is more challenging for some domains to receive high scores than others, which accounts for the difference in ranges that define lower, average, and higher quality. For instance, a sum score of 8 would be low for *Features of the Learning Environment* but average for *STEM Knowledge & Practices*.

Note that each DoS domain has three dimensions, and thus there are three possible ratings per dimension (i.e., rating for Dimension 1 + rating for Dimension 2 + rating for Dimension 3 = sum score for Domain X). For instance, receiving three perfect ratings of 4 for organization, materials, and participation equals a sum score of 12 for *Features of the Learning Environment*, which falls in the higher quality rating range. If you have multiple observations, you would take the average of each dimension separately before calculating the sum score for the given domain (thus the need for decimal points in the chart.

Figure C. Dimensions of Success (DoS) Ratings for 12 Dimensions by State