Received: 27 January 2017 | Revised: 11 July 2017 | Accepted: 17 September 2017

DOI: 10.1002/sce.21327

WILEY

RESEARCH ARTICLE

Improving STEM program quality in out-of-school-time: Tool development and validation

Ashima Mathur Shah[1] | **Caroline Wylie**[2] | **Drew Gitomer**[3] | **Gil Noam**[1]

[1]The PEAR Institute: Partnerships in Education & Resilience, Harvard Medical School & McLean Hospital, Belmont, MA, USA

[2]Educational Testing Service, Princeton, NJ, USA

[3]Graduate School of Education, Rutgers University, New Brunswick, NJ, USA

Correspondence
Ashima Mathur Shah, The PEAR Institute: Partnerships in Education & Resilience, Harvard Medical School & McLean Hospital, 1075 Pleasant St., Belmont, MA 02478, USA.
Email: pear@mclean.harvard.edu

Abstract

In and out-of-school time (OST) experiences are viewed as complementary in contributing to students' interest, engagement, and performance in science, technology, engineering, and mathematics (STEM). While tools exist to measure quality in general afterschool settings and others to measure structured science classroom experiences, there is a need for reliable measures of STEM program quality in OST settings such as afterschool programs, summer camps, and museum or science center programming. In this paper we present the development of the Dimensions of Success (DoS) tool, which defines twelve key components of informal, exploratory STEM programming that goes beyond the school day. Additionally, we present a validity argument that includes reliability evidence for the DoS tool based on two studies: Study 1 (n = 284 observations) and Study 2 (n = 56 observations). Our findings suggest that the coherence of the constructs and validity evidence, as well as the training and certification procedures in place for DoS, make it an important tool to understand the quality of STEM experiences for youth beyond the school day. A tool like DoS has several implications, including the ability to make national comparisons across programs, create aggregate databases, improve program quality and professional development, as well as to link program quality to student-level outcomes.

KEYWORDS

afterschool, informal science, observation tool, out-of-school time, quality assessment, STEM, validity and reliability study

1 | INTRODUCTION

Despite much variability, overall, US student performance in science, technology, engineering, and mathematics (STEM) is not strong when compared internationally (OECD, 2013). There are significant differences in mathematics

and science proficiency across subgroups: for example, boys scoring higher than girls and students from higher-income families scoring higher in STEM subjects (National Science Board, 2012). Issues of access and equity in STEM education have led to the strong emergence of afterschool and summer programs rooted in youth development across the United States. These environments are increasingly viewed as both complementary and supplemental to school learning (Pierce, Bolt, & Vandell, 2010).

In order to support out-of-school time (OST) staff to successfully provide programming in an area where they may not have training or appropriate support, it is important to provide clear definitions of quality for science teaching and learning in this informal environment. Observation tools serve a dual purpose of providing language to define quality learning experiences in settings that value youth-development over academic achievement, and also to offer a consistent way to measure and compare quality across a wide range of programs. In order to understand how these OST programs are supporting students in STEM learning, there is a need to develop measurement tools specifically addressing STEM and these informal experiences. Our work presented in this paper focused on developing an observation framework for STEM learning in OST and building the validity evidence for the observation tool and rubrics.

While there are many tools available to measure the quality of classroom learning both during the school day and in general afterschool or summer programs, there are limited tools that are designed specifically for STEM programming in OST. Staff in STEM programs have traditionally determined which outcomes to focus on and how to measure them, often using project-specific surveys. Recent attempts to better define outcomes for the field include an initiative by the Afterschool Alliance that used experts to define a set of student-level outcomes for afterschool STEM programming (Afterschool Alliance, 2013). As outcomes are defined, assessments that can measure these outcomes across settings can be useful for providing systematic descriptions of informal STEM programs.

Student-level surveys or assessments allow for tracking of students' progress or interest in a particular educational intervention like an afterschool program. Observations can support a better understanding of how program inputs influence these outcomes by looking at the quality of the experiences that promote interest and/or learning. By understanding what leads to those student-level outcomes, we can distill key components so they can be replicated. Observation data can also help staff, administrators, or funders gain an understanding of how activities are enacted in a program in ways that go beyond self-reports or checklists administered upon the completion of programming. They allow us to look closely at the interactions among facilitators, students, and resources. In this way, we gain insights into the methods used by facilitators to guide learning, the ways in which students are engaging with materials and ideas, and the quality of discourse and learning processes.

The development of an informal science observation protocol builds on and extends on work using observations in educational settings. Most widely used are general protocols designed to examine classroom qualities across all classrooms and grades (e.g., Danielson, 2011). Inherent in these scoring protocols are assumptions particular to formal educational settings, particularly the relatively stable and ongoing participation of students and teacher in the classroom. When such tools are used, the observer collects evidence from a single observation, but interprets that evidence as a signal of classroom routines that have been developed over longer periods of time. These protocols often prescribe sampling multiple observations to collect evidence that is representative of interactions between a teacher and her students over the course of a year. The unit of analysis is almost always the individual classroom or the teacher (see Bell et al., 2012).

Some protocols describe instructional interactions in abstract terms designed to be applied equally well in English language arts, science, or social studies classrooms (e.g., Danielson, 2011). Although protocols may value the teaching of reasoning, for example, there is little guidance about what such reasoning looks like in different disciplinary classrooms. There are however, a number of protocols that have been developed specifically for formal classrooms in mathematics (Hill, Schilling, & Ball, 2004) and science (Horizon Research, Inc., 2000; Minner & Delisi, 2012; Schultz & Pecheone, 2014). Such protocols have been typically used as research tools. These protocols each focus on evidence of STEM reasoning and disciplinary practices, albeit each with different emphases and scoring procedures. Some focus on specific teacher actions (e.g., Minner & Delisi, 2012), while others make more global evaluations of some

instructional segment (e.g., Schultz & Pecheone, 2014). Each of the protocols, however does adhere to the same assumptions of formal education structures as the general protocols noted previously.

A significant body of research exists for observations in formal education settings (e.g., Bell et al., 2012; Kane & Staiger, 2012; Pianta, Hamre, Haynes, Mintz, & La Paro, 2009) and describes issues and findings related to the validity of such approaches. These studies focus on the reliability of judgments, sources of variance attributed to factors that include classrooms, teachers, occasions, as well as other factors.

Much less studied are observation protocols for informal education settings. However, Yohalem and Wilson-Ahlstrom (2009) reviewed observation tools used systematically in informal education settings to measure quality indicators such as staff facilitation processes, activity content and structure, level of engagement, relationships with peers, and program resources. Most of these tools, however, treated all informal activities, whether sports, recreation, arts, or science, in the same way regardless of content.

The assumptions underlying observations of formal settings vary in important ways from informal settings. The assumptions of relatively stable participation of teachers and students are not warranted. OST programs can occur over very differing, and often brief, timeframes. And of course, the assumptions of what constitutes a formal learning environment versus an informal one are quite different. For all of these reasons, the target of inference for informal observations and other evaluations are typically the program rather than the individual teacher/classroom.

1.1 | Study goals

Generic observation tools for informal settings are not able to capture the dynamics of STEM learning experiences with great depth. Therefore, since 2010, our team at The PEAR Institute: Partnerships in Education and Resilience decided to take advantage of the advances in the study of observing interactions among teachers and students, observations of STEM classrooms, and observations of informal settings to develop a hybrid observation tool that is tuned specifically to the unique characteristics of informal STEM settings. To achieve this goal, we used existing frameworks and literature to create an initial tool, pilot it with programs, go through multiple rounds of refinement, and then study its psychometric properties. In this paper we report on our first study, conducted in collaboration with the Educational Testing Service (ETS), where we refined and analyzed the use of the tool to build the initial validity argument. Study two was a smaller follow-up study to examine whether our improved training approaches led to a more reliable instrument. As noted in the introduction, afterschool science programming can benefit from a tailored observation tool. It is well-documented the student interest in science declines as students get older (e.g., Aschbacher, Ing, & Tsai, 2014; Dabney et al., 2012; Potvin & Hasni, 2014), and afterschool programming can offer the time and space for tackling that fall in engagement. Due to a youth-development perspective, many afterschool environments provide access to facilitators and mentors that know how to connect with students and attend to their social emotional needs (e.g., Durlak & Weissberg, 2007). Further, these informal environments also provide a context for re-sparking lost student interest in science and building confidence and trust in their abilities. We are sharing this work with the intention that it serves as a model for others hoping to develop robust tools that can capture interactions during non-school based science learning experiences, that it offers an approach for creating and testing an accessible framework for a field where staff are not always trained for this type of teaching and learning, and that it offers one way to develop an efficient and clear common language to guide cycles of formative assessment and feedback with staff.

1.2 | The Dimensions of Success tool

In 2007, we created a prototype for a formal assessment tool, Dimensions of Success (DoS), to assess quality indicators of STEM programming in OST (Dahlgren & Noam, 2009). The early version of the DoS tool was based on five broad categories of potential impacts presented in the Framework for Evaluating Impacts of Informal Science Education Projects (Friedman, 2008):(1) Awareness, Knowledge, or Understanding; (2) Engagement or Interest; (3) Attitude; (4) Behavior; and (5) Skills. The dimensions of the tool represent features of the informal learning environment that are associated

with these outcomes (e.g., Shernoff, 2010). The category of attitudes was not included in the dimensions, as it seemed better measured by student surveys versus observations. Later enhancements to the definitions of constructs in the tool drew on the six-strand framework describing what learners do cognitively, socially, developmentally, and emotionally when they engage with science in informal learning environments (National Research Council [NRC], 2009, p. 294). In 2014–2015, additional revisions were made that clarified the types of science and engineering practices that could be easily observed during an observation, as this was a point of difficulty for trainees who were learning to use the tool. To this end, language was better aligned with the newly released Next Generation Science Standards (NGSS Lead States, 2013).

The final twelve dimensions that comprise DoS in its current form are arranged in four domains and presented in Table 1. The four domains are *Features of the Learning Environment*, *Activity Engagement*, *STEM Knowledge and Practices*, *and Youth Development in STEM*. This structure serves as an organizational device for training and communication of the rubric to support understanding of related conceptual sets. Scoring judgments, however, are made at the dimension level.

The three dimensions of the *Features of the Learning Environment* domain capture the logistics and preparation of the activity, whether the materials are appealing and appropriate for the learning goals, and how the learning environment creates a suitable space where students can explore science informally.

The three dimensions of the *Activity Engagement* domain require observers to describe how the activity engages students: for example, the dimensions examine whether or not all students have access to the activity, whether activities are moving toward STEM concepts and practices purposefully or superficially, and whether or not the activities are hands-on and designed to support students to think for themselves versus being given the answer.

The *STEM Knowledge and Practices* domain defines how the informal STEM activities are helping students understand STEM concepts, make connections, and participate in the inquiry practices (e.g., collecting data, using scientific models, building explanations, etc.) that STEM professionals use and determines whether students have time to make meaning and reflect on their experiences.

Finally, the *Youth Development in STEM* domain assesses how student-facilitator and student–student interactions encourage or discourage participation in STEM activities, whether or not the activities make STEM relevant and meaningful to students' everyday lives, and how the interactions allow youth to make decisions and have a voice in the learning environment and community. Together, these four domains capture key components of a STEM activity in an informal afterschool or summer program.

In 2007 and 2008, initial field tests of DoS were conducted in programs serving approximately 1,700 children from grades K–12 in urban, suburban, and rural settings (Dahlgren, Larson, & Noam, 2008; Dahlgren & Noam, 2009). Programs varied in location (e.g., school, museum, community center) and focus (e.g., career awareness, college preparation, hands-on science exploration).

In 2011, formal study of the psychometric properties began with researchers and in close collaboration with a state-level afterschool network. The team brought together expertise in OST, classroom observation, the evaluation of instruments, data collection in networks, and access and communication with afterschool programming for sites for the pilot study.

1.3 | Tool structure

For each DoS dimension (see Table 1 below), there are two pages of information: one with definitional guidance for observers and the other with the four levels of the rubric. The definitional guidance includes a description, elaboration, and commentary. The dimension description illustrates the main focus of the dimension (as seen in Table 1), while the elaboration describes the relevance of the dimension to the OST context and what might distinguish dimensions that may seem closely related on face value. The commentary focuses on scoring specific information to help an observer distinguish between one rubric level and another.

Each dimension is rated on a 4-point scale. The four points are defined by the degree to which there is evidence to support the essential features of the respective dimension:

TABLE 1 The DoS domains and dimensions

Domain	Dimension	Rubric description
Features of the learning environment	Organization	Focuses on the extent to which the facilitator delivers the observed activities in a way that reflects appropriate planning and preparation, through having the necessary materials readily available, being ready to accommodate to changing situations, and having smooth transitions to prevent time loss and chaos in the learning environment.
	Materials	Focuses on the extent to which the activities make use of materials that are appropriate for the particular youth in a program, aligned with intended STEM learning goals, and appealing to youth.
	Space Utilization	Focuses on the extent to which the program space is utilized in a manner that is conducive to STEM learning in an OST environment.
Activity engagement	Participation	Focuses on the extent to which the youth have equal access to the activities offered. Participation refers only to general participation (access to materials, prompting to participate and contribute, etc.) in the activities and does not consider the degree to which the youth are participating in STEM thinking/reasoning or inquiry practices.
	Purposeful Activities	Focuses on the extent to which activities are structured so that youth clearly understand the goals of each activity, and the connections between them; it also examines the degree to which the facilitator uses his/her time productively to best support youth understanding of STEM learning goals.
	Engagement with STEM	Focuses on the extent to which youth are engaging in hands-on activities that allow them to actively construct their understanding of STEM content. It also looks at whether or not the activities leave youth as passive recipients of knowledge from the facilitator or as active learners who interact directly with STEM content so they do the cognitive work and meaning-making themselves.
STEM knowledge and practices	STEM Content Learning	Focuses on the extent to which youth are supported to build understanding of science, mathematics, technology, or engineering concepts through STEM activities. Observers must consider the accuracy of STEM content presented during activities, the connectedness of STEM content presented during activities, as well as evidence of youth uptake of accurate STEM content based on their questions, comments, and opportunities to demonstrate what they learned.
	Inquiry	Focuses on the extent to which activities support the use of STEM practices. These STEM practices are usually used in the service of helping youth learn the science content more deeply. Stronger quality involves youth participating in STEM practices in authentic ways (versus superficially going through the motions of inquiry) to pursue scientific questions, address a design problem, collect data, solve an engineering task, etc.
	Reflection	Focuses on the extent to which activities support explicit reflection on the STEM content in which the youth have been engaged. This dimension also refers to the degree to which the quality of youth reflections is superficial or meaningful and connection-building.
Youth development in STEM	Relationships	Focuses on the extent to which the facilitator has positive relationships with the youth and other facilitators as well as the extent to which youth have positive relationships with each other.
	Relevance	Focuses on the extent to which the facilitator makes connections between the STEM activity and the youth's lives and personal experiences, other subject areas, or a broader context.
	Youth Voice	Focuses on the extent to which the STEM activities encourage youth to have a voice by taking on roles that allow for genuine personal responsibility and having their ideas, concerns, and opinions acknowledged and acted upon by others.

TABLE 2 Engagement with STEM dimension rubric

Level 1 Evidence absent	Level 2 Inconsistent evidence	Level 3 Reasonable evidence	Level 4 Compelling evidence
There is minimal evidence that the youth are engaged in hands-on activities in which they can explore STEM content.	There is weak evidence that the youth are engaged in hands-on activities in which they can explore STEM content.	There is clear evidence that the youth are engaged in hands-on activities in which they can explore STEM content.	There is consistent and meaningful evidence that youth are engaged in hands-on activities in which they can explore STEM content.
The activities mostly leave youth in a passive role, where they are observing a demonstration or listening to the facilitator talk (minimal hands-on opportunities).	Youth engage in hands-on activities; however, there is limited evidence that the hands-on activities encourage youth to engage with STEM content in meaningful ways ("hands-on, minds-off").	There are some opportunities for youth to engage in hands-on activities that allow them to actively explore STEM content. Some parts of the activities still leave youth as passive observers while the facilitator does all the cognitive work. OR Activities are hands-on and minds-on (at level 4) for less than half of the youth.	There are consistent opportunities for youth to actively explore STEM content by engaging in hands-on activities, where youth do the cognitive work themselves and the facilitator maintains the role of facilitator versus teller.

TABLE 3 Sample rating and evidence for engagement with STEM dimension

Rating	Sample observer evidence
2	All of the students are excitedly making their balloon rockets move across the string (very active, hands-on). They set them up multiple times and take turns, but their comments are limited to how fun the activity is: "Whoa, this is so cool!" "Look, mine is a super rocket!" "Can I go next? Mine is going to be awesome!" One student offered, "I think mine is going the fastest this time," and the facilitator responded, "That's cool," without any prompting for what might make it go faster, etc. No deeper questioning or engagement is prompted by the facilitator beyond helping the students to follow the procedure for making the balloon rocket work.

- *Compelling Evidence* (Level 4): defined by the existence of compelling and consistent evidence supporting the presence of practices and/or interactions defined by the dimension;
- *Reasonable Evidence* (Level 3): defined by the presence of clear evidence of the dimension, although the evidence is less consistent than the evidence that would merit a score of 4;
- *Inconsistent Evidence* (Level 2): suggests the presence of weak evidence supporting the dimension definition; and
- *Evidence Absent* (Level 1): reflects minimal or no evidence in support of the dimension definition.

Each level of evidence for a particular dimension is described in the tool (see Table 2 for an example) and further illustrated in the extensive training process through video exemplars.

Observers not only assign a numerical rating, but they also write detailed evidence justifying the score. Table 3 illustrates what an observer might write as a summary of the evidence he or she recorded during the observation that provides justification for the final score decision.

1.4 | Using DoS in the field

Afterschool programs need to have access to reliable and valid ways of measuring the quality of STEM programming specifically designed for OST environments. This need for standardization is reinforced given findings that many afterschool programs use homegrown surveys, observation tools, or written assessments to internally monitor their progress, thereby preventing cross-program comparisons or aggregation of data to report trends across the field

(Dahlgren, Noam, & Larson, 2008). There may be concerns that defining and measuring quality for these settings may lead to more prescriptive programs that lack the flexibility, creativity, and outside-the-box thinking that is so valued in informal learning environments. While the goal of this effort is to standardize the constructs and their operationalization, as well as how settings are observed, it is also deliberate in not prescribing or privileging particular informal learning structures over others. Indeed, it is our contention that high-quality STEM experiences, as defined by the protocol, can be realized in the full range of informal learning structures. We worked from the premise that standards, when built on a foundation of flexibility, choice, and informal learning approaches, can help define and push for higher quality in the field without sacrificing variation and innovation.

The key characteristic of the observation tool is that DoS ratings and written evidence to support those ratings are based solely on the live observations of the specific STEM activities. The protocol does not attend to, for example, analyses of lesson plans or discussions with facilitators. The DoS tool measures the quality of an activity when all the interactions of students, materials, and facilitators are at play so that observers can give feedback that is grounded in the actual conversations that occur in the space and the actual responses of students as they experience the activity. In addition to providing feedback that can be used as formative assessment for continuous improvement at a program site, the DoS scores can be used to look at trends over time or as a measure in an evaluation. For example, some organizations may choose to pair program quality data with student outcome data to evaluate their STEM program offerings. Larger state networks can look at trends across their state or within particular regions, as well as make comparisons within the state or to national norms, as DoS is used across the country.

While there are many types of informal STEM learning experiences offered, DoS is designed to be used to observe programs that concentrate on OST or afterschool science programming that has pre-planned activities, a designated facilitator/teacher/leader, and some type of structure (e.g., an afterschool science club that meets in the cafeteria twice a week, a community center's afterschool program that has a 30-minute science block each day, or a museum's or science center's camp or weekend programming with pre-planned activities and curricula). DoS is not designed for free-choice environments where students are interacting with exhibits and guiding themselves through a series of experiences.

1.5 | DoS versions

During Study 1, DoS observers conducted observations in 15-minute blocks followed by ten minutes to rate the evidence for each dimension. The observers began timing the first block as the lesson began, took a10-minute break to score, and resumed taking notes for the second block at the 25-minute mark in the lesson, repeating the pattern to start note-taking for the third block at the 50-minute mark. This is a similar observation structure to that used in formal education settings (e.g., Pianta et al., 2009). The rationale for this segmentation was to limit the amount of classroom interactions that an observer has to judge at any one time. Additional blocks of observation and ratings then occurred until the activity was over.

Observer feedback during Study 1 consistently pointed to the struggle to capture details of the full activity when there were constant starts and stops using the block method. For example, evidence of student content learning may gradually build throughout an activity and only capturing a 15-minute snapshot does not truly provide an understanding of the quality of the activity as a whole. Additionally, key aspects of the activity were missed during the 10-minute scoring period. Therefore, in Study 2, the protocol for conducting an observation was revised to have observers first take detailed field-notes that captured evidence for the entire activity as a whole, followed by the use of the rubrics to summarize evidence and select rubric levels.

As more observers used DoS in the field and provided feedback with regard to confusing wording or unclear explanations, minor wording revisions were made to clarify the constructs and quality levels and to reduce rater drift; however, there were no substantial changes to the format or content.

The revised version of the DoS tool, along with the improved protocol for completing the observations, addressed ways to enhance observer scoring based on literature summarizing sources of error when doing observations (Bell et al., 2014). In their chapter, the authors analyzed key sources of error when doing observations. First, observers can

make judgment errors due to their content knowledge, professional training, or experience using the protocol (training and ongoing calibration is recommended). Also, the more observers use an observation tool, the more they can start to stray (or drift) from the criteria based on their own personal interpretations or biases of the constructs. They also may agree less in how they draw on evidence to make judgments between score points (Bell et al., 2014, pp. 57–58). Finally, there can be errors due to youth behaving differently due to the presence of an observer in the room. To minimize these sources of error, efforts were made to continually check in with observers, re-calibrate, and update the requirements for observers. DoS certification trainings were revised to create more rigorous checks for reliability, and observers in Study 2 were required to complete a more in-depth training, calibration, and certification process.

2 | STUDY 1

2.1 | Methods

2.1.1 | Study 1 sample

In Study 1, observations using DoS were conducted in two geographical regions: the Midwest and New England (see Table 4). Due to the logistics of scheduling observations and maintaining relationships with programs locally, the New England region had one team of observers, and the Midwest region had a different team of observers for each state (Missouri, Ohio, and Kansas). In New England, a call for observers was posted on science education and informal science discussion threads and local university graduate student employment offices. Applicants' resumes were reviewed and in-person interviews were required. The minimal requirements included a Bachelor's degree, preferably in the areas of education, science education, assessment, or a related field. Most applicants were graduate students, teachers, or those who worked with science organizations. In the Midwest, recruitment and interviews took place through university graduate student offices, through recommendations by our local partners who identified active staff in a variety of childhood and family services locations as well as science museums, and through existing coaches and observers that already existed in the afterschool system. Both the New England and Midwest teams communicated frequently to ensure that the same protocols, program selection, and observation procedures were being used. The total of 284 observations took place in a range of organized, facilitator-led program settings and structures including school-based afterschool programs, science clubs at community organizations (e.g., Boys and Girls Clubs of America, YMCAs, etc.), museums and science centers programming, and other community outreach and nonprofit organizations.

Of the 284 observations, 59% were completed at OST programs during the school year, and 41% were conducted at summer programs. Both school year and summer programs were observed in both regions. Within a given summer program, multiple observations often occurred due to the fact that multiple STEM activities or types of camps were running at the same summer program site. The STEM options offered at each program varied in terms of the content focus (e.g., physics, robotics, engineering design, filmography, arts and science, etc.), length, use of homemade or commercially available curricula, frequency of meetings, etc. Programs were recruited through flyers sent out to afterschool online list-serves, information distributed at educational conferences, and direct emails to STEM programming leadership located through online program searches. Program staff agreed to participate in the research study and were given a report of their scores in return. Fifty percent of the programs observed ran for only one to two weeks, and only 5% of the programs ran for the entire year.

TABLE 4 Distribution of programs, observations, and observers by region

Region	Number of observations *N* (%)	Number of programs *N* (%)	Maximum number of observations per program	Mean number of observations per program	Number of observers *N* (%)	Number of observations per observer	
						Maximum	Mean
Midwest	167 (59%)	23 (40%)	48	7	26 (68%)	29	6
New England	117 (41%)	34 (60%)	12	3	12 (32%)	34	10

2.1.2 | Data collection procedure

Prior to data collection, all observers completed a two-day online or in-person training led by the lead developers of DoS. During these sessions, trainees were introduced to the twelve dimensions of the DoS instrument, the process of taking field notes, and building evidence for a rating. Each dimension was reviewed, with specific attention given to the description, elaboration, and commentary in addition to the rubric's specific language. The training provided scaffolding for the trainees, gradually increasing the complexity of the scoring process over time, starting initially with a focus on a single dimension, to later scoring cases for multiple, and eventually, all 12 dimensions. Benchmark video example cases (scored prior to training by multiple members of the research team) were used to illustrate particular score points. Trainees also reviewed written descriptions of STEM activities, which provided additional examples of high and low practices on each dimension. In this way, trainees were exposed to a range of activities and examples of each level for all 12 dimensions. When the training was completed, all observers independently scored two videos of science activities and had to demonstrate scoring proficiency before they were cleared to move on to live scoring.

For each observation, two certified DoS observers arrived early to a program site to introduce themselves to the youth and the facilitator, and to find an appropriate place to observe in the space that would minimize any distractions that might be caused by their presence. Observers used notebooks or laptops to take field-notes in 15-minute blocks of the activity. After each 15-minute block, they would spend ten minutes using the DoS rubrics to assign ratings for each dimension. Each 25-minute block (observing and scoring) represented a segment.

To establish rater reliability, two observers watched the same STEM activity simultaneously and communicated about start and end times for each segment so that their scores reflected the same observed part of the activity. Observers did not discuss their ratings or evidence during or after the observation; they were only asked to share their impressions after submitting their observation data.

Most observations ranged from 30 to 90 minutes, depending on the length of the STEM activity from start to finish. The modal number of segments was two. If the afterschool program continued with a music or sports activity, only the STEM activity time was observed. Observers were politely instructed to not interfere with the activity in any way as they observed the lesson in real-time. For example, they did not assist the facilitator in teaching the activity or attempt to ask questions or elicit responses from students about their learning.

2.1.3 | Data analyses and results

Building a validity argument (Kane, 2006) is an ongoing process to examine the sequence of claims that are made based on observation scores. The data from Study 1 represent a first step in doing this whereby we examine quality of scores and the extent to which we are capturing generalizable characteristics. In Table 5 we lay out the questions to be addressed and the related analyses.

One way to examine the data is to determine the extent to which the full score range is applied across the observations. Do observers use different scale points to characterize activities? Table 6 provides the means and standard deviations of dimension scores across the 284 observations in Study 1, averaged across segments and observers. One

TABLE 5 Validity argument for DoS

Inference	Question	Analysis
Scoring	Do observers use different scale points to characterize activities?	Descriptive statistics to determine the use of the full scoring scale
	How reliable are observers in their ratings?	internal consistency as measured by Cohen's Kappa and inter-rater agreement levels between observer pairs scoring the same activity
	How well do the scores support the predicted domain structure of the DoS?	Exploratory factor analysis to examine the factor structure of the 12 DoS dimensions
Generalization	To what extent do claims about a program generalize across different topics or different settings?	A preliminary G-study analysis

TABLE 6 Mean and standard deviation of rater judgments

	Mean	Standard deviation
Materials	3.36	0.60
Organization	3.29	0.61
Space utilization	3.23	0.62
Relationships	3.19	0.71
Participation	3.17	0.67
Purposeful activities	2.84	0.78
STEM content learning	2.47	0.83
Engagement with STEM	2.46	0.81
Inquiry	2.32	0.79
Youth voice	2.19	0.67
Reflection	1.86	0.74
Relevance	1.84	0.73

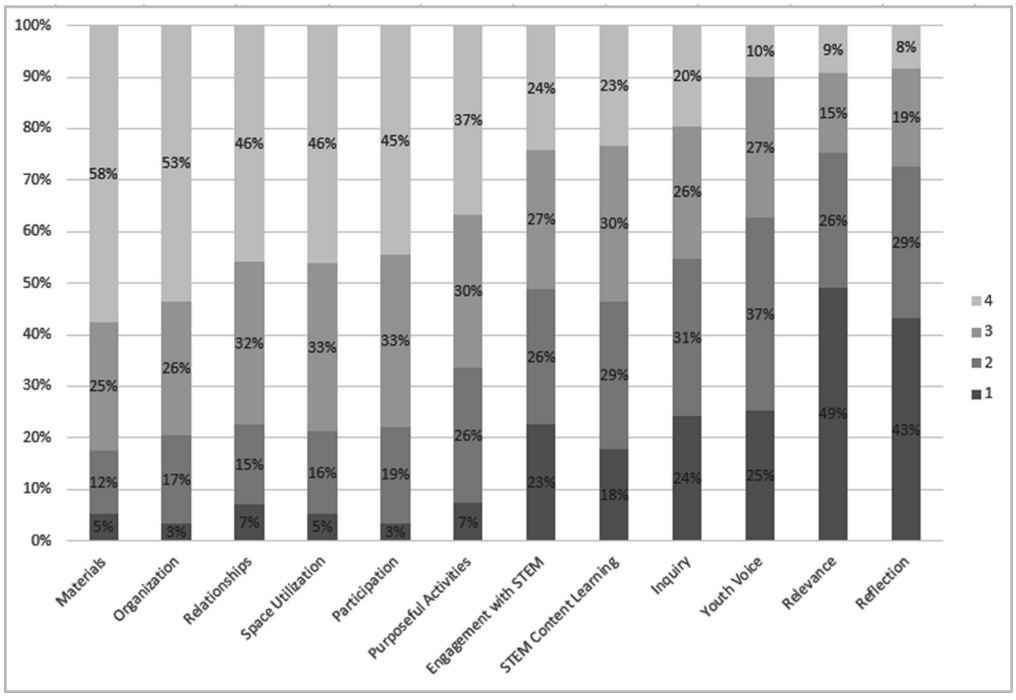

FIGURE 1 Stacked bar chart showing proportion of segment-level scores for each dimension

finding that is apparent is that certain dimensions are scored more highly than others. Activities seem to be relatively well organized, with materials available, space well used, relationships strong, and students actively participating. However, the relevance to science is limited as is any evidence of reflection. The second finding from examining the standard deviations is that dimension scores vary substantially across activities, even when averaged across observers and segments.

Figure 1 illustrates the distribution of scores across all observed segments (15-minute observation blocks) on each of the 12 dimensions, ordered from the dimension with the highest proportion of score 4s to the lowest. For each

TABLE 7 Measures of observer agreement across the 284 observations

Scale	Percent exact agreement	Percent exact or adjacent agreement	Quadratic weighted Kappa	Correlation
Materials	50.4	90.6	.30	.30
Participation	49.6	89.7	.31	.31
Engagement with STEM	48.7	83.8	.33	.33
Organization	48.7	92.3	.28	.29
Space utilization	47.0	86.3	.16	.17
Purposeful activities	47.0	86.3	.27	.27
Reflection	46.2	89.7	.44	.44
Relevance	45.3	86.3	.44	.47
Relationships	43.6	86.3	.41	.42
Youth voice	43.6	88.9	.19	.21
STEM content learning	43.6	85.5	.37	.37
Inquiry	40.2	88.9	.44	.45

dimension, all score points were used, and noticeably the whole score range was used, rather than scores clustering around the center of the score scale (2s or 3s). There are differences in the score distribution across the dimensions. For example, while only 23% of observed segments were scored 1 or 2 for Organization, 73% of observed segments were scored in these two lowest score categories for Relevance.

Additional empirical evidence relevant to the scoring inference comes from examining the score data for inter-observer reliability in order to gather support for the consistency with which observational data are translated to scores through the application of the rubrics. Table 7 above presents several measures of observer agreement at the segment level. *Percent Exact Agreement* refers to the instances when both observers awarded the same score to a particular segment during an observation period. *Percent Exact or Adjacent Agreement* refers to instances when two observers awarded either exact scores or scores that differed by only one point. For example, a score of 1 from one observer and a score of 2 from a second observer would be counted as adjacent agreement. The quadratic weighted for Kappa, a measure of inter-observer agreement that takes into account agreement happening by chance, is presented as well. The correlations between the pairs of scores are also presented.

There is a suggested set of guidelines for interpreting Kappa values from Landis and Koch (1977). They suggested that Kappa values between .81 and 1 indicate perfect agreement, values between .61 and .80 indicate substantial agreement, values between .41 and .60 indicate moderate agreement, values between .21 and .40 indicate fair agreement, and values between 0 and .20 indicate slight agreement. Using these guidelines, we have moderate agreement levels for four of the dimensions, fair agreement for six, and only slight agreement for the remaining two dimensions (Youth Voice and Space Utilization). It is important to note, however, that these indices of agreement are in line with, and sometimes stronger than, agreement levels that have been observed in studies in formal settings (e.g., Bell et al., 2014).

The final piece of empirical evidence that speaks to the quality of the translation from observed performance to observed scores is evidence about data fit. Although the dimensions presented to observers during training were organized by the four domains (see Table 1), the four-factor structure was not a good fit when a confirmatory factor analysis was completed. Therefore, an Exploratory Factor Analysis approach was used (using maximum likelihood with promax rotation as the extraction method). As Table 8 illustrates, the analysis indicated the dimensions loaded into two distinct groups. The first factor focuses on the ways in which students make meaning in STEM including Purposeful activities,

TABLE 8 Factor loadings from the exploratory factor analysis

Dimension	Component	
	1	2
STEM content learning	.859	−.046
Reflection	.815	−.328
Purposeful activities	.678	.209
Inquiry	.617	.333
Engagement with STEM	.526	.328
Relevance	.590	−.232
Youth voice	.510	.262
Space utilization	−.370	.790
Relationships	−.069	.741
Participation	.013	.707
Materials	.069	.683
Organization	.058	.678

Engagement with STEM, STEM content learning, Inquiry, Reflection, Relevance, and Youth voice. The second factor is focused more on the organization and support structure in place during the program and activities, including Organization, Materials, Space utilization, Participation, and Relationships. The exploratory factor analysis provided a tentative structure that is interpretable, given an understanding of teaching and learning experiences in informal science classrooms. However, we do not recommend reducing DoS to just these two factors given the importance of feedback at the dimension level to programs.

Dimensions that loaded on the learning environment factor were generally scored more highly on their respective scales than were dimensions associated with STEM meaning-making. This trend of dimensions that loaded on the learning environment factor (i.e., Organization, Materials, Space utilization, Participation, and Relationships) scoring higher than the dimensions that loaded on the content-related factor (i.e., Purposeful activities, Engagement with STEM, STEM content learning, Inquiry, Reflection, Relevance, and Youth voice) suggests that it is more challenging to facilitate high-quality activities on the meaning-making dimensions. Scores on the dimensions that contribute to each factor can be averaged to provide two composite scores.

Only in Study 1 did we have a sufficient number of participants to conduct a generalizability study in order to estimate variance associated with known sources of measurement error in the DoS dimension scores—that is, differences in scores due to the content being taught, the specific observation visit, the raters, or the interactions between these sources of difference. The G-coefficient was used to estimate dimension reliabilities after two observations by two observers (see Appendix). We were then able to conduct a dependability (D) study to estimate how many observations would be needed to get a more reliable estimate of the quality of a module (keeping two observers constant). These analyses (Appendix) indicated that multiple observations are needed to get a stable measure of quality and that the most stable measure was to use the two-factor structure that was identified from the Exploratory Factor Analysis to create two composite scores for the quality of the learning environment and the quality of STEM meaning-making. Interestingly, more observations were needed to understand the STEM meaning-making factor versus the learning environment factors: While four observations would likely result in a reliable estimate of the learning environment factor, even 10 observations would be insufficient for the STEM content factor, given the levels of inter-observer agreement in the current study. Given these findings, we made changes to the training, certification, and calibration process and followed up with Study 2 to examine the impact on resulting assessor reliability.

TABLE 9 Distribution of programs, observations, and observers by region

Region	Number of observations *N* (%)	Number of programs *N* (%)	Maximum number of observations per program	Mean number of observations per program	Number of observers *N* (%)	Number of observations per observer	
						Maximum	Mean
Midwest	37 (67%)	19 (59%)	7	2	15 (88%)	18	5
New England	19 (33%)	13 (41%)	4	1	2 (12%)	19	10

3 | STUDY 2

3.1 | Methods

3.1.1 | Study sample

Study 2 observations were conducted in two separate regions: the Midwest and New England (see Table 9) during summer STEM program offerings in both regions. In both areas, we recruited existing program staff who were already using the DoS tool and who had completed the updated more rigorous training process. A total of 56 observations were completed by pairs of raters (producing two separate observation data sets).

Midwest observers conducted their observations in different pairings based on availability. New England had a pair of observers that scored all activities together.

3.1.2 | Data collection procedure

Compared to Study 1, some changes were made to increase the quality and overall rigor of the certification process. All observers still were required to complete a two-day online or in-person training led by DoS trainers. The exercises were revised to include a wider range of examples and to illustrate more diverse settings and age groups. Also, new emphasis on language connected to science and engineering practices was incorporated to clarify the Inquiry construct. Trainees had to complete three (instead of two) calibration video exercises that featured more complex interactions. Also, before becoming officially certified, observers had to submit two practice observations from the field that were given detailed feedback from the training team.

Our partners in the Midwest coordinated the scheduling of observations at a range of STEM programs. Midwest observers were provided a small stipend for completing and uploading their observations to our online database. Unlike Study 1, observers for Study 2 were asked to provide a rating and detailed evidence for all twelve dimensions at the end of the observation instead of every 15 minutes. Observations ranged from 30 to 90 minutes long. As in Study 1, two observers watched the same activity simultaneously, did not interfere with the teaching, and did not discuss the activity quality until their individual ratings and evidence were completed.

3.1.3 | Data analyses and results

Similar to Study 1, we also present descriptive statistics of these observations to provide evidence for the scoring inference. Given the smaller sample size (n = 54 observations), the focus of Study 2 was on re-visiting the rater reliabilities since more rigorous training and certification steps were developed and rating segments changed from Study 1. Table 10 provides the mean and standard deviation of rater judgments across the 54 observations.

Figure 2 illustrates the percent of scores that fell in each score category for each of the 12 dimensions across the 56 observations conducted in Study 2 (organized from dimension with the greatest number of 4s to the least). Similar to what was observed in Study 1, for each dimension all score points were used, without scores clustering around the center of the score scale. Also similar to Study 1, there are differences in the distributions across the dimensions, with only 21% of the observations scored 1 or 2 for Organization compared to 70% of the observations scored in these two lowest score categories for Relevance.

TABLE 10 Mean and standard deviation of rater judgments

	Mean	Standard deviation
Space utilization	3.52	0.67
Relationships	3.39	0.76
Materials	3.37	0.90
Organization	3.35	0.87
Participation	3.01	0.80
Purposeful activities	2.77	0.99
Engagement with STEM	2.52	1.10
Youth voice	2.30	0.95
STEM content learning	2.28	1.07
Inquiry	2.28	1.21
Relevance	2.13	0.99
Reflection	2.03	0.93

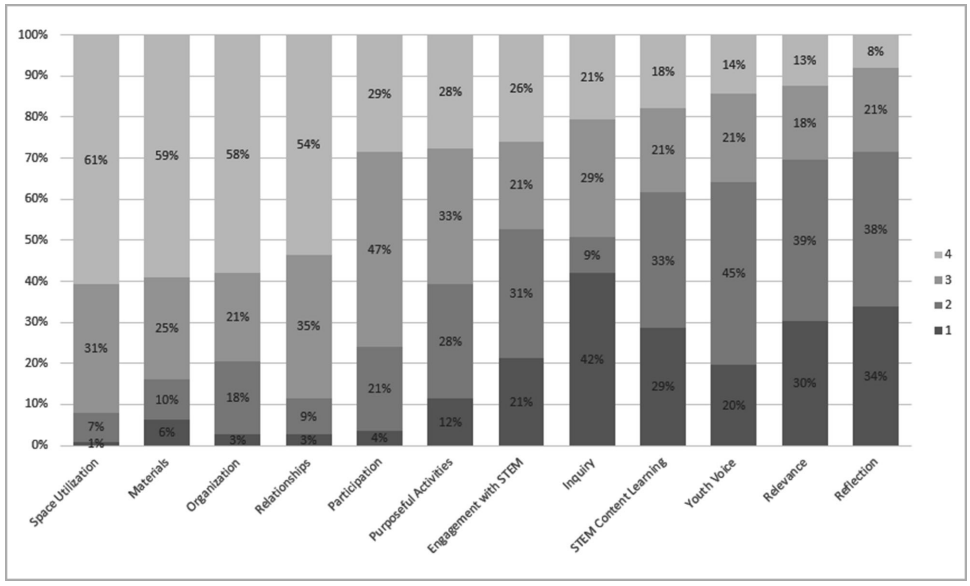

FIGURE 2 Stacked bar chart showing distribution of observation scores for each dimension

The factor structure from Study 1 suggested a learning environment factor (i.e., Organization, Materials, Space utilization, Participation, and Relationships) and a STEM meaning-making factor (i.e., Purposeful activities, Engagement with STEM, STEM content learning, Inquiry, Reflection, Relevance, and Youth voice). The bars in Figure 2 are organized with the dimensions with the greatest proportion of scores of 4 to the left. Similar to Study 1, the five learning environment dimensions (located to the left in Figure 2) all had a greater proportion of scores of 3 and 4 compared to the dimensions in Study 1 that loaded on the STEM meaning-making factor.

Table 11 presents several measures of observer agreement at the observation level. Parallel to Table 7 for Study 1, *Percent Exact Agreement* refers to the instances when both observers awarded the same score to an observation, while *Percent Exact or Adjacent Agreement* refers to instances when two observers awarded either exact scores or scores that differed by only one point. The quadratic method for calculating Cohen's Kappa statistic is presented along with the correlations between the pairs of scores, Agreement levels in Study 2 should be considered with caution as in each

TABLE 11 Measures of observer agreement across the 56 observations

Scale	Percent exact agreement	Percent Exact or adjacent agreement	Quadratic weighted Kappa	Correlation
Materials	91.1	100.0	.94	.95
Relationships	89.3	96.4	.82	.83
Engagement with STEM	82.1	94.6	.82	.82
Relevance	82.1	100.0	.91	.92
Inquiry	80.4	96.4	.90	.90
Organization	78.6	98.2	.82	.83
Space utilization	78.6	100.0	.76	.77
Purposeful activities	75.0	100.0	.87	0.87
Youth voice	75.0	100.0	.86	.87
Reflection	73.2	96.4	.78	.78
Participation	71.4	98.2	.73	.75
STEM content learning	69.6	96.4	.82	.82

region, only two raters participated. Thus, we do not have as much confidence as in Study 1 that agreement levels will generalize across raters.

Nevertheless, comparing the results in Table 7 and Table 11 suggests that the changes to the training and certification procedures had a positive impact on rater agreement. For example, in Study 1 there were no dimensions where there was perfect or substantial agreement (0.81 to 1), whereas in Study 2 there were 8 dimensions that had perfect agreement, and the remaining dimensions had substantial agreement (Kappa value between 0.60 and 0.80). There were substantial improvements in agreement with new training, certification, and monitoring procedures.

4 | DISCUSSION

The purpose of Study 1 was to establish validity and reliability evidence for the DoS observation tool, while Study 2 was intended to provide additional reliability evidence after some time had passed with the tool in the field and appropriate refinements had been made. In Study 1, we observed that the full score range for each dimension was used, that there is a meaningful two-factor structure for dimensions which could be used to support composite scores (learning environment and STEM meaning-making) that provide a more stable view of activity quality, with some of the noise at the dimension level being averaged out and that pairs of observers provided similar scores across observations. There is some variation by dimension with agreement being easier to achieve on some dimensions than others. The level of acceptable agreement is very much a factor of consequence. In a low stakes, professional learning context, value is likely to come from the discussion after the observation, as much as from the specific scores. The more consequential any score is, the higher the expectation is that scores will not be a function of who the rater is. As noted the composite scores will have greater levels of agreement, and for overall program evaluation may be sufficient. The lack of agreement may also suggest that the field needs to have a better shared understanding of what these various constructs mean, in terms of providing quality STEM education (Gitomer et al., 2014).

The G-study underscored the importance of multiple observations of activities at a program in order to gain an understanding of quality with sufficient reliability. DoS can be used not only for ongoing assessment but also for program quality support. It is, therefore, important that programs that can only conduct a limited number of observations use the data to provide feedback about that particular activity's strengths and weaknesses, rather than use it for high-stakes decision-making (e.g., hiring, firing, cutting a program, etc.).

Study 1 observers generally scored within one point of each other, but exact agreement at the dimension level was not as high as it was for Study 2. The findings from Study 2 suggest that our careful design of a longer and more rigorous training process as well as an extended calibration and certification procedure led to better observer agreement. It is important to note that the constructs themselves did not change from Study 1 to Study 2, but the dimensions were defined more clearly, training materials given more details, updated exemplars to help differentiate among dimensions, and feedback from the field incorporated to make the rubrics more user-friendly overall.

In both studies the score distributions across dimensions are similar: Organization, Materials, Space utilization, and Relationships tended to score higher than Engagement with STEM, STEM content learning, Inquiry, and Reflection. This pattern reflects the literature base that documents how hard it is to master teaching approaches that support minds-on exploration, accurate explanations (beyond memorization) of scientific phenomena, and the authentic application of STEM practices (e.g., Roehrig & Kruse, 2005; Roehrig & Luft, 2004). These findings can have implications for large-scale professional development and intervention plans (see section below).

When doing work that involves training and collecting data with observations across many programs with many shifting variables, there are some limitations to discuss in the context of the results presented in this paper. We hope to address these in future work. First, we did not test the DoS scores with other validated tools such as more general quality tools or student-level outcomes instruments. There are existing tools that measure science engagement, interest, and literacy; those are all possible tools to use in future work. Evidence of relationships between DoS scores and related measures would strengthen the construct-related validity of the tool. Second, we were not able to keep consistent pairs observing together or to maintain a rotation that would allow us to see how different pairs behaved. Observations were scheduled by convenience—they were not required to adhere to the best research conditions. Finally, we do not consider the impact of program philosophy or content focus.

Future research should explore if particular types of programs lend themselves to more reliable judgments of quality. To what extent is the reliability of the instrument sensitive to differences among programs, context, raters, etc.? Another area of future exploration is whether or not reliability shifts when quality shifts. In other words, do raters agree more when they are observing activities deserving scores in the 3 or 4 range or when scores are lower at 1 or 2? Some understanding of these relationships would help further improve training and inform potential revisions to the wording that appears in the rubric. Additionally, we will need to replicate a generalizability study conducted as part of Study 1, to refine our understanding of the number of observations that are required to get stable measures of program quality.

5 | CONCLUSION

The findings from Study 1 and Study 2 suggest several implications for individual programs offering STEM learning experiences as well as for the broader field. First, the 12 DoS dimensions provide a common language for quality STEM learning experiences. With such diversity in informal learning experiences, it is an important contribution to the field to offer a framework that is general enough to be used in summer camps, afterschool programs, science center program offerings, and so on, while specific enough to lead to concrete feedback and conversation about programming needs and improvement. Further, the formal teaching profession has existing requirements for credentialing with definitions of coursework, understanding of pedagogy, and performance indicators that are expected to be successful educators. The afterschool STEM world does not have a formal credentialing process for educators, and often those leading these OST programs come from a range of backgrounds that often exclude experience in science teaching. With the need for these educators to "learn on the job" or quickly gain an understanding of pedagogical approaches for hands-on, engaging science learning, the DoS tool offers a digestible entry-point. Twelve constructs with key indicators offer a way for OST educators who may have more of a youth-development background to understand the key components of a quality learning experience and to gauge their developing skills as they use the rubric or receive feedback from an observer. Many of our observers have shared that the DoS certification training offers their first deliberate study of

what high-quality STEM teaching is, and its reliance on real-time data from the field makes it relevant and immediately meaningful.

Another important implication of this work is that it provides scores and evidence that program leaders and activity facilitators can use right away to improve the quality of their activities. Simply providing DoS scores would leave the burden of translating these scores on the observer, but since all scores must be backed by detailed evidence that ties to the rubrics, feedback and conversation-starters are built-in to the process. This is critical, as identifying strengths and weaknesses in programs for youth and improving access to high-quality STEM learning experiences is connected strongly to overall workforce and equity issues in this country. Specifically, today's youth need to be exposed both in and out of school to experiences that will prepare them for the technology-rich and data-driven issues of the future. Since youth only spend 20% of their annual waking hours in school, OST experiences can significantly increase their exposure and engagement to STEM (Banks et al., 2007). These experiences can especially impact underrepresented groups in STEM including students of color and girls. Thus, having a tool like DoS to make sure these critical opportunities are as effective as possible is valuable to the field and to the preparation of our youth.

Finally, another way DoS contributes to the field is by providing multiple entry-points to quality improvement. If a program is just starting and needs to understand the baseline workings of their program, the scores and evidence can be used to decide on the most pressing concerns first. If a program is more developed and has experienced staff, the tool can be used to refine and push quality to a higher level, or to gather data that can help a program show their value to funders or other partners that may enhance their success. Whether the data leads to a short-term goal of moving from a 2 to a 3 on a particular dimension, or loftier goals of bringing all staff to scores of 3 or higher on all activities offered, DoS scores and evidence can be used in ways that meet diverse program needs.

The importance of translating findings at the program level and at a larger scale is underscored by the continuing need to "make the case" for afterschool STEM funding. DoS allows the field to collect data across a range of diverse settings in afterschool/summer programs and to report quality using a common measure (beyond their homemade surveys) that allows them to see how they are doing compared to a national average or other similar programs. The DoS data is stored (without identification markers beyond geographic region and program descriptors) to continually grow a nation-wide database. This allows programs to write reports that can show funders and other stakeholders their progress on particular dimensions compared to national norms. Having these data available and continually updated will help address federal pressure to develop evidence standards for determining that programming is moving toward high-quality STEM outcomes. Additionally, private funders looking to invest in afterschool networks want proof that their contributions are affecting the lives of youth and engaging them with science in ways beyond what the school day can provide.

Next steps to take DoS even further include addressing the need for supporting resources that can be used in complementary ways with DoS. Programs require systemic support for implementing high-quality STEM activities. For example, as facilitators receive scores, they can use those scores and feedback to improve their activity plans. A cycle then continues: enactment, reflection/feedback based on DoS scores, planning based on DoS scores and definitions of high quality, and repeated observation. This cycle will be best scaled and supported by a range of tools that maintain the DoS dimensions as the backbone. We have already piloted a DoS Program Planning Tool, DoS Feedback Report, and Coaching Guide to help support the translation and use of DoS data at the site, program, district, or state levels and beyond. Further work of creating video libraries of high- and low- quality practice for afterschool STEM programs, connecting to existing professional development resources, and supporting capacity for STEM with the often changing and shifting world of afterschool programs is needed.

The results of both studies suggest that the DoS tool can be used to assess aspects of practice for OST STEM learning. The training provided for observers in these studies was sufficient to produce reasonable levels of observer reliability, but the generalizability study suggested that caution is needed when using the results from only two observations of STEM activities to make high-stakes claims about the quality of programming. STEM learning experiences can spark, inspire, and maintain student interest and competency in future STEM coursework, STEM careers, or even general appreciation of STEM in their lives. As STEM offerings in OST programming increase rapidly, it is important to focus on quality and use valid and reliable tools to provide meaningful data to the field.

ACKNOWLEDGMENTS

The development and study of the DoS tool as presented in Study 1 was supported by funding from the National Science Foundation (Award Number 1008591). We would also like to thank the Noyce Foundation and Charles Stewart Mott Foundation for funding Study 2 to further strengthen our evidence base for this tool. The views expressed in this paper are those of the authors and do not represent the views of the Educational Testing Service (ETS) or Rutgers University.

ORCID

Ashima Mathur Shah http://orcid.org/0000-0002-3163-4650
Caroline Wylie http://orcid.org/0000-0001-7378-9733
Drew Gitomer http://orcid.org/0000-0002-2452-7953

REFERENCES

Afterschool Alliance. (2013). *Defining youth outcomes for STEM learning in afterschool*. Washington, DC. Retrieved from www.afterschoolalliance.org/STEM_Outcomes_2013.pdf

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

Banks, J., Au, K., Ball, A., Bell, P., Gordon, E., Guitierrez, K., … Zhou, M. (2007). *Learning in and out of school in diverse environments: Life-long, life-wide, life-deep*. Seattle, WA: The LIFE Center, University of Washington, Stanford University, SRI International and Center for Multicultural Education, University of Washington. Retrieved from life-slc.org/docs/Banks_etal-LIFE-Diversity-Report.pdf

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2–3), 62–87.

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.

Dabney, K. P., Tai, R., Almarode, J., Miller-Friedmann, J., Sonnert, G., Sadler, P., & Hazari, Z. (2012). Out-of-school time science activities and their association with career interest in STEM. *International Journal of Science Education, Part B*, *2*(1), 63–79.

Dahlgren, C. T., Larson, J. D., & Noam, G. G. (2008). I*nnovations in out-of-school time science assessment: Peer evaluation and feedback network in metropolitan Kansas City summer METS initiative.* Report for Ewing Marion Kauffman Foundation, Kansas City, MO.

Dahlgren, C. T., & Noam, G. (2009). Helping children reach their potential through quality programming: The importance of evaluation in out-of- school time science, technology, engineering and mathematics. A presentation at The First National Conference on Science and Technology in Out-of- School Time, Chicago, IL. Retrieved from www.projectexploration.org/watershed/.

Dahlgren, C. T., Noam, G. G., & Larson, J. D. (2008). *Findings for year one data for the Informal Learning and Science Afterschool Study*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Danielson, C. (2011). *The framework for teaching evaluation instrument*. Princeton, NJ: The Danielson Group.

Durlak, J. A., & Weissberg, R. P. (2007). *The impact of after-school programs that promote personal and social skills*. Chicago, IL: Collaborative for Academic, Social, and Emotional Learning. Retrieved from https://www.casel.org/wp-content/uploads/2016/06/the-impact-of-after-school-programs-that-promote-personal-and-social-skills.pdf

Friedman, A. (Ed.). (2008). *Framework for evaluating impacts of informal science education projects*. Report from a National Science Foundation Workshop. Retrieved from http://www.aura-astronomy.org/news/EPO/eval_framework.pdf

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1–32. Retrieved from https://www.tcrecord.org/library/abstract.asp?contentid=17460

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*(1), 11–30. https://doi.org/10.1086/428763

Horizon Research, Inc. (2000). *Inside the classroom observation and analytic protocol*. Chapel Hill, NC: Author.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Policy and Practice Brief. MET Project. Bill & Melinda Gates Foundation.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Minner, D., & DeLisi, J. (2012). *Inquiring into Science Instruction Observation Protocol (ISIOP) user manual*. Waltham, MA: Education Development Center, Inc.

National Research Council (NRC). (2009). *Learning science in informal environments: People, places, and pursuits*. Washington, DC: The National Academies Press. Retrieved from https://doi: 10.17226/12190

National Science Board. (2012). Science and engineering indicators *2012*. Arlington, VA: National Science Foundation (NSB 12-01). Retrieved from https://www.nsf.gov/statistics/seind12/c1/c1h.htm

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

OECD (2013). Lessons from PISA 2012 for the United States, Strong Performers and Successful Reformers in Education, OECD Publishing (pp. 19–53). Retrieved from https://www.oecd.org/pisa/keyfindings/PISA2012-US-CHAP2.pdf

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2009). Classroom Assessment Scoring System (CLASS), secondary manual. Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

Pierce, K. M., Bolt, D. M., & Vandell, D. L. (2010). Specific features of after-school program quality: Associations with children's functioning in middle childhood. *American Journal of Community Psychology*, *45*(3–4), 381–393. https://doi.org/10.1007/s10464-010-9304-2

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education*, *50*(1), 85–129.

Roehrig, G. H., & Kruse, R. A. (2005). The role of teachers' beliefs and knowledge in the adoption of a reform-based curriculum. *School Science and Mathematics*, *105*(8), 412–422.

Roehrig, G. H., & Luft, J. A. (2004). Constraints experienced by beginning secondary science teachers in implementing scientific inquiry lessons. *International Journal of Science Education*, *26*(1), 3–24. https://doi.org/10.1080/0950069022000070261

Schultz, S. E., & Pecheone, R. L. (2014). Assessing quality teaching in science. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 444–492). New York, NY: John Wiley & Sons.

Shernoff, D. J. (2010). Engagement in after-school programs as a predictor of social competence and academic performance. *American Journal of Community Psychology*, *45*(3), 325–337. https://doi.org/10.1007/s10464-010-9314-0

Yohalem, N., & Wilson-Ahlstrom, A., with Fischer, S., & Shinn, M. (2009). Measuring youth program quality: A guide to assessment tools (2nd ed.). Washington, DC: The Forum for Youth Investment.

How to cite this article: Shah AM, Wylie C, Gitomer D, Noam G. Improving STEM program quality in out-of-school-time: Tool development and validation. *Sci Ed*. 2018;102:238–259. https://doi.org/10.1002/sce.21327

APPENDIX: DETAILS OF THE GENERALIZABILITY STUDY

Ideally, we would like to see scores vary in ways that capture true differences between modules. We would hope that smaller variance would be due to factors associated with who conducted the observations. In order to estimate how much scores vary due to the specific factors that contribute to any particular observation, we conduct a generalizability analysis. The analysis estimates variance associated with known sources of measurement error in the DoS dimension scores. The design specified was (o:m) × v, which allowed us to estimate five sources of variance that come from the facets of module, visit (two observations were conducted for each module observed), and observer along with interactions of those facets with each other. The percentage of variance for each source is reported as the source's variance estimated divided by the sum of all the variance estimates. The statistical model applied in this study decomposes score variance in the following way, where *m* is the module effect, *v* is the occasion effect (or visit in this context), *o* is the observer effect, *mv* is the interaction between module and visit, and the last term in the equation is the residual.

$$\sigma^2\left(X_{mov}\right) = \sigma^2_m + \sigma^2_v + \sigma^2_{o,mo} + \sigma^2_{mv} + \sigma^2_{ov,mov,e}$$

For these analyses, the module variance (*m*) is the estimated stable variation in scores that is due to variation in modules across the other sources of variance. The variance due to visits (v) takes into account the proportion of variance in module scores due to visits—that is, how differently scores were between visits 1 and 2 for given modules. The variance for observers nested within a module (o:m) accounts for how much of the variability was due to variation among observers within a module.

As noted earlier, observers were nested within sites. The goal was to ensure that the pairs of observers were evenly distributed across the possible pairings. In practice, only the design of the data collection in the New England geographic region allowed us to examine for the pairing and distribution of observers across the observations and variance in scores. The other regions had too many observations by the same observer at the same sites to make the analysis possible; this was out of our control, as some programs grouped their activities in such a way that observers could only make one trip and had to do multiple observations in one day.

Due to study design, we also cannot separately estimate observer effects or observer by module effects with the remaining data, as one of the constraints of the design was that different sets of observers rated different sets of modules. All we can estimate is the variability in the scores observers assign within a module, which we would like to be as small as possible. The variance due to the module by visit interaction (mv) indicates differences in the relative order of modules from visit to visit. If this is not small, then that means that *some* modules were scored quite differently from visit 1 to visit 2—that is, some modules may have scored high on visit 1 and low on visit 2. Finally, the visit by observer nested within module interaction (ovmp) is the residual, or the unexplained, variance in this design.

The last row in Table A1 provides the G-coefficient for each dimension. The G-coefficient can be considered a reliability estimate for how reliable or dependable the measurement procedures for a particular dimension are likely to be after two observations by two observers. Table A2 presents the generalizability results for the two composite dimensions that were identified from the exploratory factor analysis.

The G-coefficient is calculated as:

$$\epsilon\rho^2 = \frac{{\sigma^2}_m}{{\sigma^2}_m + {\sigma^2}_\delta}$$

where $\sigma^2(\delta)$ is the relative error, which is calculated as:

$${\sigma^2}_\delta = \frac{{\sigma^2}_{o,mo}}{n'_o} + \frac{{\sigma^2}_{mv}}{n'_v} + \frac{{\sigma^2}_{ov,mov,e}}{n'_o n'_v}$$

Similar to internal consistency reliability coefficients such as coefficient alpha and KR-20, values of the G-coefficient range from 0 to 1. Values closer to 1 are more desirable and indicate greater dependability (reliability). The larger the proportion of module variance relative to error variance, the larger the G-coefficient will be.

Table A1 presents the G-study results by dimension for the New England data. There is variation across the dimensions in terms of how much variation in scores is attributable to the modules. The module variance component ranged from 4% to 29%. Variance due to differences between visits was very small across dimensions. The variance due to observers nested within modules ranged from 0% to 36%. The module by visit interaction was generally small across dimensions in both sites, except for Space Utilization and Relationships. Finally, the residual, or unexplained, variance was generally high across all dimensions, exceeding 50% for nine of the dimensions.

The G-coefficients range from .09 (Reflection) to .58 (Youth voice). A reasonable rule of thumb for this coefficient is .70, but none of the dimensions met this threshold. It is possible to take the estimates of the G-coefficients and conduct a dependability (D) study to estimate how many observations would be needed to get a more reliable estimate of the quality of a module (keeping two observers constant). From those estimates, conducting four observations rather than just two per module would likely result in reliable estimates for three dimensions (Materials, STEM content Learning, and Youth voice). Increasing the number of observations to seven would likely result in getting a reliable estimate for quality with respect to the Organization dimension. For the dimensions where the G-coefficient starts off below .40, estimating the impact of even ten observations is not likely to result in reliable judgments.

TABLE A1 New England G-study results

	Organization		Materials		Space utilization		Participation		Purposeful activities		Engagement STEM with STEM		STEM content learning		Inquiry		Reflection		Relationships		Relevance		Youth voice	
	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%
m	.080	17%	.171	26%	.057	9%	.086	15%	.098	16%	.132	17%	.245	29%	.169	19%	.022	4%	.154	20%	.092	11%	.185	28%
o:m	.012	2%	.051	8%	.067	11%	.143	25%	.058	9%	.133	17%	.068	8%	.236	27%	.253	41%	.121	15%	.292	36%	.000	0%
v	.000	0%	.000	0%	.001	0%	.000	0%	.000	0%	.000	0%	.000	0%	.003	0%	.002	0%	.003	0%	.037	5%	.053	8%
mv	.025	5%	.000	0%	.117	19%	.086	15%	.000	0%	.036	5%	.022	3%	.000	0%	.026	4%	.176	22%	.074	9%	.000	0%
ov:m	.356	75%	.427	66%	.364	60%	.268	46%	.453	74%	.490	62%	.511	60%	.468	53%	.314	51%	.330	42%	.315	39%	.423	64%
G-coefficient	>.43		.56		.24		.32		.41		.39		.59		.42		.09		.40		.26		.58	

Note: m is the variance due to module differences; v is the variance due to visits (or observation occasion); o:m is the variance due to observers nested within a module; mv is the variance due to the module by visit interaction; and ov:m is the variance due to visit by observers nested within module interaction.

TABLE A2 G-Study results using the two-factor structure

	New England			
	Learning environment factor median		STEM meaning making factor median	
	VC	%	VC	%
m	.123	29%	.062	13%
o:m	.041	10%	.081	17%
v	.000	0%	.001	0%
mv	.043	10%	.056	12%
ov:m	.216	51%	.278	58%
G-coefficient	.56		.31	

Note: m is the variance due to module differences; v is the variance due to visits (or observation occasion); o:m is the variance due to observers nested within a module; mv is the variance due to the module by visit interaction; and ov:m is the variance due to visit by observers nested within module interaction.

In the earlier analyses, we explored the structure of the observation data and reported that a two-factor structure was identified from the exploratory factor analysis: (1) a learning environment-related factor (Organization, Materials, Space utilization, Participation, Relationships) and (2) a STEM content-related factor (Purposeful activities, Engagement with STEM, STEM content learning, Inquiry, Reflection, Relevance, Youth voice). The same G-study analysis was repeated for these two factors, using the median score than was awarded across the set of dimensions associated with each factor. REF _Ref364578574 \h * MERGEFORMAT

Table A2 presents the variance estimates for each factor. Although none of the G-coefficient estimates met the .70 threshold, the learning environment factor estimate was the highest. Similar to the previous dimension data, the G-coefficient was used to estimate the likely reliabilities associated with increasing the number of observations. Four observations would likely result in a reliable estimate of the learning environment factor, while even ten observations would be insufficient for the STEM content factor, given the levels of inter-observer agreement in the current study. Thus, if a site's capacity only allows for 1–2 observations to gain an understanding of quality, it is important to use the data to provide feedback and to discuss the activity's strengths and weaknesses instead of using it for high-stakes decision-making (e.g., hiring, firing, cutting a program, etc.).

In conclusion, the data from the G-studies show significant variation both across dimensions and the two factors. We currently do not have sufficient data to further explore these differences, but at a minimum would recommend no fewer than four observations of a program. We fully expect that given the refinements in training and monitoring, more consistent observer judgments are possible.