International Journal of Science Education, Part B
Communication and Public Engagement

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rsed20

The double-dip: quality discrepancies in out-of-school time STEM programs

Rebecca K. Browne , Patricia J. Allen & Gil G. Noam

To cite this article: Rebecca K. Browne , Patricia J. Allen & Gil G. Noam (2021): The double-dip: quality discrepancies in out-of-school time STEM programs, International Journal of Science Education, Part B

To link to this article: https://doi.org/10.1080/21548455.2020.1866787

Published online: 11 Jan 2021.

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION, PART B
https://doi.org/10.1080/21548455.2020.1866787

The double-dip: quality discrepancies in out-of-school time STEM programs

Rebecca K. Browne, Patricia J. Allen and Gil G. Noam

The PEAR Institute, McLean Hospital, and Harvard Medical School, Belmont, MA, USA

ABSTRACT

We report on national trends in STEM program quality using the Dimensions of Success (DoS), an empirical observation tool that provides a common definition of STEM program quality. We analyzed ratings for 12 dimensions of quality obtained from 452 DoS observations performed in 452 STEM-focused OST programs across 25 U.S. states by certified DoS observers. When plotted on a graph, the averages for the 12 quality dimensions display a 'double-dip' – a phrase that has been used in practice to communicate OST STEM strengths (higher ratings) and challenges (lower ratings). Nationally, OST programs excelled in quality indicators related to features of the learning environment, including preparation, materials, and space, as well as relationships. However, programs demonstrated less consistent evidence for quality in dimensions related to STEM knowledge and practices, including STEM content learning, inquiry, and reflection (dip #1), as well as areas related to supporting youth voice and STEM relevance (dip #2). This 'double-dip' persisted regardless of region, locale, season, and participant age or gender, though certain program and participant characteristics changed the magnitude of the scores. Ongoing professional development efforts are needed to address persistently challenging areas that are essential for building children's STEM skills, content knowledge, and fluency. Key words: Informal education, STEM, research trend, professional development.

Introduction

Opportunities for young people to learn about science, technology, engineering, and math (STEM) in the U.S. have dramatically changed over the past decade. Young learners have increasing access to engage in STEM outside the classroom, including afterschool programs, museums, libraries, parks, and summer camps (Krishnamurthi et al., 2013). Recent estimates suggest that over 10 million youth participate in U.S. out-of-school time (OST) programing, with a large portion of these youth consisting of historically underrepresented or underserved populations, including girls, youth of color, and youth from low-income families (Afterschool Alliance, 2014). Additionally, there has been an increase in the number of programs focused exclusively on STEM, catalyzed in part by high impact reports underscoring the societal and economical importance placed on STEM National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007). Importantly, STEM learning in OST programing is now viewed as an essential strategy in STEM education (Krishnamurthi et al., 2013), with research indicating that STEM engagement in informal learning environments promotes interest and persistence in STEM among youth (Allen et al., 2020; Chittum et al., 2017; Funk & Hefferon, 2016; Maltese & Tai, 2010).

CONTACT Gil G. Noam gil_noam@harvard.edu 1075 Pleasant Street, Belmont, MA 02478

Consensus study reports on informal STEM environments by the National Research Council support the potential for OST programing to strengthen STEM education on a national scale (National Research Council, 2015). Notably, these reports emphasize that the quality of STEM learning experiences in OST programs are integral for maximizing impact on youth STEM learning and persistence. In other words, increasing availability, access, and attendance is necessary but not sufficient; to improve STEM literacy and engagement among all youth, the field must expand its focus on the quantity of informal STEM learning opportunities to include the quality of informal STEM learning opportunities (Allen et al., 2020; Mahoney et al., 2010; NRC, 2015).

Yet, there are still many questions that remain about the quality of OST STEM programs being offered to youth and how to best support program quality improvement. Until recently, conceptualizing, defining, assessing, discussing, and systematically improving OST STEM program quality represented a significant challenge for the OST field, particularly because informal learning environments vary widely in terms of setting, duration, and content, and there were no valid measurement tools specific to OST STEM settings (Noam et al., 2017; Shah et al., 2018). The needs of researchers and practitioners to understand the quality of STEM-focused OST programs led to the creation of a measure known as the Dimensions of Success (DoS), an observation tool that provides a common, evidence-based language to communicate levels of quality specific to STEM activities conducted in informal learning environments (Shah et al., 2018). The DoS framework, which the observation tool is based in, includes 12 dimensions addressing key pedagogical principles spread across four broad domains including Features of the Learning Environment, Activity Engagement, STEM Knowledge and Practices, and Youth Development in STEM (Shah et al., 2018).

The DoS framework has been adopted by many OST STEM programs across the U.S. to guide their continuous quality improvement efforts. For example, several national cohort initiatives use the DoS framework to promote the collection of data and communication of findings to practitioners to improve the quality of informal STEM activities, such as the Afterschool and STEM system-building initiative (Allen et al., 2019) and the STEM Learning Ecosystems Community of Practice (Allen et al., 2020; Traill et al., 2015; Traphagen & Traill, 2014). The wide adoption of DoS allows for the aggregation of OST STEM program quality data from communities across the U.S. into one common national database. Analysis of a national database can inform practitioners, researchers, and funders alike about the common strengths and challenges of the field. Previous research using the DoS observation tool has found that, on average, programs tend to excel in areas related to the features of the learning environment – including preparation, materials selection, and space utilization – as well as areas related to demonstrating positive relationships with youth (Allen et al., 2019; Shah et al., 2018). However, these same programs tend to show less consistent evidence of quality in areas more directly related to STEM learning and practices – including STEM content learning, engaging youth in STEM practices, and reflection – as well as dimensions related to youth ownership of learning and the connection of STEM learning to real world occurrences.

When the data collected from recent DoS studies are visualized – with average ratings for the 12 dimensions (across the four DoS domains) plotted on a graph – we observe what looks like a 'double-dip;' a phrase that has been used in practice to communicate the common strengths and challenges found in OST STEM programing. When scanning the plotted averages from left to right (from dimension 1–12), the quality of STEM programing is rated very high for the first six dimensions (within the Features of the Learning Environment and Activity Engagement domains), but then quality ratings dip beginning at the STEM Content Learning dimension (within the STEM Knowledge and Practices domain) – this signals the first dip, or the first three areas for improvement. Quality ratings then peak again with high ratings for the Relationships dimension, but then a second dip in quality occurs for the Relevance and Youth Voice dimensions – this signals the second dip, or two additional areas for improvement. We discovered that visualizing and discussing the data in this way helps communicate programing strengths and challenges to non-

research audiences, including practitioners, funders, policymakers, and other stakeholders of STEM learning communities.

In the present study, we aimed to expand the field's understanding of STEM program quality by examining the 'double-dip' phenomenon to determine whether the trend can be replicated in a larger national sample – including more OST programs, states, and regions – and whether certain participant or program factors – including region, locale, season, grade/gender of program participants, among other variables – can change the direction or the magnitude of the trend. We begin with a review of the DoS framework and recent studies that support the validity and reliability of the measure. We next describe our methods and results from our analysis of national trends in OST STEM program quality based on data collected using the DoS observation tool. We conclude with a discussion of key findings – based on our analysis of differences in OST STEM program quality by region, locale, season, and grade/gender of program participants – and recommendations for researchers and practitioners based on national strengths and areas for improvement identified in this study.

The Dimensions of Success tool

The DoS observation tool was specifically designed to assess the quality of informal STEM learning environments and was developed and studied with funding from the National Science Foundation (NSF) in 2007 along with research partners at Educational Testing Service (ETS) and practitioners of Project Liftoff. Building from prior work in the field, DoS was designed to align with two of the leading frameworks for OST STEM program quality: The NSF's *Framework for Evaluating Impacts of Informal Science Education* Projects (Friedman, 2008) and the National Research Council's (NRC) report, *Learning Science in Informal Environments: People, Places, and Pursuits* (NRC, 2009). The NSF framework defines five categories for assessment:

- Awareness, knowledge, or understanding of STEM concepts, processes, or careers
- Engagement of interest in STEM concepts, processes, or careers
- Attitude toward STEM-related topics or capabilities
- Behaviors related to STEM concepts, processes, or careers
- Skills based on STEM concepts, processes, or careers

The NRC's report identifies six strands the describe what learners do cognitively, socially, developmentally, and emotionally when they engage with science in informal environments (NRC, 2009). The six strands are:

(1). Experience excitement, interest, and motivation to learn about phenomena in the natural and physical world
(2). Come to generate, understand, remember, and use concepts, explanations, arguments, models, and facts related to science
(3). Manipulate, test, explore, predict, question, observe, and make sense of the natural and physical world
(4). Reflect on science as a way of knowing; on process, concepts, and institutions of science; and on their own process of learning about phenomena
(5). Participate in scientific activities and learning practices with others using scientific language and tools
(6). Think about themselves as science learners and develop an identity as someone who knows about, uses, and sometimes contributes to science (NRC, 2009, pp. 294- 295)

As depicted in Table 1, the DoS observation tool, in its final form, is comprised of twelve dimensions arranged in four domains; *Features of the Learning Environment, Activity Engagement, STEM*

Knowledge and Practices, and Youth Development in STEM. Each dimension is accompanied by a four-point rubric indicating increasing levels of quality. A rating of 3.0 (reasonable evidence) is used as the criterion threshold for STEM program quality in any of the dimension.

The *Features of the Learning Environment* domain captures the logistics and preparation of an activity, whether the materials are appealing and appropriate, and how the learning environment creates a suitable space for informal STEM learning.

The *Activity Engagement* domain requires 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.

The *STEM Knowledge and Practices* domain defines how informal STEM activities are helping youth understand STEM concepts, make connections, and participate in the inquiry practices 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 indicators of quality in an informal STEM learning environment as defined by the NSF and NRC (Shah et al., 2018).

DoS is used in two primary ways: (1) as an internal assessment tool to promote continuous program quality improvement, and (2) as a research and evaluation tool. Typically, after each observation, feedback is communicated to either the facilitator or program manager discussing both activity strengths and areas for improvement, adjustments are made to programing, and programing is then observed again. When used in this capacity, the tool helps to provide a common language that program staff can use to discuss their activities, describing where they excel, and where they can improve. DoS can also be used by external researchers and evaluators to track quality by program, network, or state, over time. The data collected with DoS can be used to make decisions by funders and policy makers.

Previous research has demonstrated the psychometric properties of DoS (Shah et al., 2018). Two separate studies tested the validity of the tool by examining the descriptive statistics to determine the use of the full scoring scale; internal consistency as measured by Cohen's kappa; the inter-rater agreement levels between observer pairs scoring the same activity; a factor analysis to examine the factor structure of the 12 DoS dimensions; and a preliminary G-study analysis. Notably, results found the inter-rater agreement for the 12 dimensions had Cohen's kappas ranging from .73 to .94 and percentage agreement ranging from 95% to 100% based on the current training and certification methods (Shah et al., 2018). Furthermore, DoS has shown similar, and sometimes stronger, levels of agreement between raters than the levels of agreement reported for observation tools used in formal settings (Bell et al., 2014; Shah et al., 2018).

A large-scale study involving 158 STEM-focused OST programs that received training and support from 11 state afterschool networks found that youth participating in higher quality programing, as determined using DoS ($n = 250$ observations total), reported significantly greater change in STEM attitudes and beliefs – including STEM engagement, career interest and knowledge, and identity – than peers participating in lower quality programing (Allen et al., 2019). A smaller scale study of four OST programs conducted by Fenton et al. (2019) used DoS to study changes in STEM program quality following a professional development intervention (i.e. the Click2Science approach, with a minimum of five hours of professional development, including two 90-minute face-to-face, hands-on trainings, two 30-minute staff meetings, two 30-minute coaching sessions, and two two-hour self-directed web lessons). Certified DoS observers performed 12 observations before and after staff training, and results showed improved STEM program quality following

Table 1. Domains, dimensions, descriptions, and examples of quality STEM programing for the Dimensions of Success (DoS) Framework.

Domain	Dimension	Rubric Description	Examples of Quality
Features of the Learning Environment (FLE)	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 available, logical sequence, flexibility, smooth transitions
	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.	Appropriate and appealing
	Space Utilization	Focuses on the extent to which the program space is utilized in a manner that is conducive to STEM learning in an out-of-school time environment.	Conducive to STEM learning with minimal distractions
Activity Engagement (ActEng)	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.	Students doing activities, following directions
	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.	Students understand activity goals and time is used to support learning
	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.	Opportunities for hands-on activities so students do the cognitive 'minds-on' work
STEM Knowledge and Practices (STEMKP)	STEM Content Learning	Focuses on the extent to which youth are supported to build understanding of science, math, 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.	Accuracy of content presented in activities and evidence of student learning
	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.	Students using inquiry practices of STEM professionals (e.g. scientists, mathematicians, engineers)
	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.	Opportunities for students to reflect and engage in sense-making about activities
Youth Development in STEM (YDSTEM)	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.	Degree of positive, respectful interactions among students and facilitators
	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.	Students and facilitators explicitly connect activities to real-world, other subjects, STEM careers, etc.
	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.	Students' opinions and ideas are heard, and they have opportunities to make decisions

the intervention as well as a positive association between DoS quality ratings and youth self-reported STEM engagement (Fenton et al., 2019). Together, these studies not only provide further support for the validity of the DoS observation tool, but they demonstrate the practical importance of observation tools, and specifically DoS, as it helps build a common language for discussing quality both within and across STEM programs. Defining and assessing OST STEM program quality serves both to improve programing and strengthen youth outcomes.

Study goals and hypotheses

Since the development of DoS nearly a decade ago, the tool has been used to observe thousands of OST STEM activities across the country. A significant amount of data has been collected to allow for the creation of a national database to aide in the communication of program strengths and challenges. There are now enough data to inform the field about the quality of OST STEM programs being offered to youth on a national scale and how best to advise practitioners and policymakers on strategies to support program improvement. Our primary aim of this study was to determine whether the double-dip persists in a larger, national dataset inclusive of programs with various levels of training and support – including many programs connected to larger systems that receive significant levels of support around quality. Our secondary aim is to explore program and participant factors that may be associated with quality ratings to inform continuous improvement, and guide funding and policy-making decisions, especially decisions that direct best practices for STEM education and strengthen college and career readiness. We define program characteristics as U.S. region (i.e. Northeast, Midwest, South, West), locale (i.e. city, suburban, town, rural), and season/time of year (i.e. school year programing, summer programing), and we define participant characteristics as the grade and gender of youth participating in activities (e.g. proportion of boys to girls participating in an activity).

Based on previous research (Allen et al., 2019; Shah et al., 2018), and the fact that OST STEM programs are rooted in youth development (Shah et al., 2018), we expected the double-dip to emerge, with those areas focused on environmental factors and relationships scoring higher than both STEM-specific areas and areas focused on youth ownership of learning and connections to real world occurrences. We expect the double-dip to emerge despite the growing number of programs connected to larger systems of support, as improvements in lower scoring dimensions likely requires the systematic roll out of training on a large scale. Additionally, we hypothesized that there would be differences in quality based on the grade range of students being served – given that the methods and strategies for meeting the learning needs of children vary developmentally – and the time of year that programing takes place – given that the nature of school and summer programing is different in terms of program offerings, youth experiences, professional development approaches, and the dosage and duration of programing. By identifying national trends in OST STEM program quality, and associations between different program and participant characteristics and quality outcomes, we can make actionable recommendations to the field, including areas for professional development, opportunities to expand STEM learning for youth, and funding.

Methods

This section describes participating programs, the DoS measure, procedures, and statistical analyses used to examine national trends in STEM program quality observations, including variation in program quality by specific program and participant characteristics.

Program participants

Program quality data was drawn from a national database consisting of a total of 893 DoS observations conducted at 452 STEM-focused OST programs between October 2013 and September

2018 which were voluntarily submitted to the lead creators of DoS. October 2013 was selected as the starting point for data inclusion, as this is when the DoS protocol and certification process were in its current, validated form (Shah et al., 2018). Observations were conducted at programs that provide informal STEM instruction to youth in grades K-12 (e.g. afterschool programs, summer camps, science centers or museums). The programs that were observed represent 25 states and the four U.S. regions defined by the U.S. Census Bureau: Northeast (22.2%), including Massachusetts, New Jersey, New York, Pennsylvania, and Vermont; Midwest (31.9%), including Indiana, Iowa, Kansas, Michigan, Missouri, Nebraska, and Wisconsin; South (33.8%), including Alabama, Florida, Kentucky, Maryland, Oklahoma, South Carolina, Texas, Tennessee, and Virginia; West (11.8%), including Alaska, California, Oregon, Utah, and Wyoming. Furthermore, programs were located in cities (46.0%), suburbs (16.2%), towns (11.1%), and rural areas (4.7%) across the country. Program size varied widely, from 1 to 80 participants.

Each of the four STEM content areas were represented across program observations: Science (47.8%), Technology (8.6%), Engineering (19.5%), and Math (4.0%). The remaining observations were of activities covering multiple STEM content areas (20.1%). STEM curriculum usage varied widely between programs. Curriculum usage can be defined as activities using a written program of STEM objectives, STEM content, STEM activities or learning experiences, and STEM resources developed or adopted by the program that build upon previous activities or learning experiences. Examples of STEM curriculum used by participating programs included Lego Robotics, NASA curriculum, and program specific curricula. Some programs did not follow a written STEM curriculum or activity plan but rather planned activities on an ad hoc basis. For example, several observations note the facilitator found the activity online.

Measure

Dimensions of Success (DoS): DoS an evidence-based observation tool that captures 12 dimensions of STEM program quality in OST learning environments along four organizing domains, was used to collect data at participating programs (see Table 1). As previously mentioned, the DoS observation tool has been rigorously field-tested and shows strong psychometric properties. Rigorous training and certification are required to perform DoS observations. Qualitative data from field notes are quantified by the observer using a standard rubric on a 4-point scale from low (1, evidence absent) to high (4, compelling evidence), with a rating of 3 (reasonable evidence) representing the criterion threshold for quality. The DoS tool measures the quality of STEM activities based on live observations when all the interactions of students, materials, space, and facilitators are at play. The protocol does not attend to analyses of lesson plans or discussions with facilitators (Shah et al., 2018).

Procedure

Observations were conducted by certified DoS observers as either part of a large-scale national evaluation or internal program quality improvement efforts. Observers completed a training and certification process that consisted of a two-day online or in-person training led by the lead developers of DoS. During these sessions, trainees were introduced to the twelve DoS dimensions. Video examples of informal STEM activities conducted with youth in grades K-12 were used to provide opportunities for trainees to practice taking field notes, providing evidence, and assigning ratings for each of the 12 DoS dimensions. Verbal and written feedback was provided by DoS trainers throughout the two-day training. After the two-day training, trainees were required to complete a video calibration process in which they viewed between one and three videos and provided evidence and ratings for DoS dimensions. Trainees then attended a 1-hour calibration call and received written feedback on their ratings and evidence from DoS trainers. Finally, trainees were

required to submit two practice field observations for review. Upon satisfactory completion of two observations, trainees were certified in the DoS observation tool for two years.

Program participation in DoS observations was voluntary. Observations were conducted by both observers affiliated with the program and external evaluators. DoS observers arrived at participating programs 10–15 min early in order to introduce themselves to the facilitator(s) and determine a place to sit and observe, while causing minimal distractions to the activity. Observers did not interact with the facilitator(s) or youth while the activity was being conducted. Observations performed by certified observers ranged from 30 to 120 min depending on the length of the STEM activity. Fields notes taken during the observation are used to form ratings and evidence for each of the 12 DoS dimensions. All qualitative (field notes and evidence) and quantitative (ratings) data, as well as demographic information about the observed program and participating youth (e.g. grade and gender) were voluntarily submitted to the lead developers of DoS via an online data collection link and added to the national database. The researchers and lead developers of DoS were not involved in the coordination of observations or collection of data but consulted with observers upon request to ensure observation guidelines were followed. Observations were most frequently conducted on one to two occasions toward the middle to end of programing, depending on each program's STEM activity schedule and the ability of observers to commute to programs located across each state. Data were reviewed by researchers to ensure the data met the standards set by the research team (i.e. individuals submitting observations were currently certified and evidence was submitted for all 12 ratings).

All procedures were reviewed and approved by the institutional review board at our research institution.

Data analysis

To ensure an equal distribution of observations across program sites, we analyzed a subset of the national database that met our inclusion criteria. Specifically, one observation was included per program site. For programs with multiple observation submissions, we selected the observation with the greatest level of qualitative detail (using evidence directly reported by observer) to support the ratings provided. We ensured that the ratings closely followed and addressed each of the components of the 12 DoS rubrics and used both examples and direct quotes from facilitators and students to support ratings. These requirements resulted in the inclusion of 452 observations of the same number of programs in the present analyses.

All statistical analyses were performed on the raw observation data that was summarized in numeric form (from 1 to 4). Nonparametric analyses were used for all of the nonnormally distributed dimensions of STEM program quality to test for differences among the independent variables (i.e. region, year type, gender ratios) based on the dependent variables (i.e. ratings for the 12 DoS dimensions). Alpha was initially set as $p < 0.05$, and Bonferroni corrections were applied to correct for multiple comparisons where appropriate. When examining differences by the independent variables, the frequencies for each program and participant characteristic were examined to ensure the distribution of characteristics were relatively even across each grouping.

Results

Overall trends

A Friedman test revealed a statistically significant difference in observed STEM program quality across the 12 DoS dimensions (χ^2 (11) = 1288.22, $p < 0.001$) (see Figure 1(a and b)). Significance values were adjusted using Bonferroni's correction for multiple tests. *Post hoc* analyses showed that four of the dimensions were consistently higher, on average, particularly the three dimensions

within the Features of the Learning Environment domain (i.e. Organization, Materials, and Space Utilization) and one dimension within Youth Development in STEM domain (i.e. Relationships). Five of the dimensions were consistently lower, on average, including the three dimensions within the STEM Knowledge and Practices domain (i.e. STEM Content Learning, Inquiry, and Reflection) and two dimensions within the Youth Development in STEM domain (i.e. Relevance and Youth Voice dimensions).

Differences by region

A Kruskal–Wallis test was used as a one-way test of variance to examine differences between the four U.S. Census regions (i.e. Northeast, Midwest, South, and West). There were significant regional differences in STEM program quality ratings for seven out of 12 DoS dimensions (see Table 2). Kruskal–Wallis multiple comparisons tests were used for *post hoc* analysis, which revealed that STEM activities observed in Western and Southern programs typically demonstrated the highest ratings of STEM program quality, with differences between the regions varying by dimension (see Table 2). However, while Southern and Western programs demonstrated higher STEM program quality scores on average, these averages were still below the criterion threshold set for STEM program quality (i.e. rating of 3.0) for both regions in the following dimensions, respectively: STEM Content Learning, Reflection, Relevance, and Youth Voice. The finding that there were below-criteria ratings, on average, for these four dimensions (within the STEM knowledge and Practices and Youth Development in STEM domains) within each region is consistent with the patterns observed across all four regions (reported as Overall Trends, above).

Differences by locale

Differences among locale (i.e. City, Suburban, Town, Rural) were examined using a Kruskal–Wallis test. Locale was defined using National Center for Education Statistics (NCES) classifications and criteria, where 'City' ($n = 208$) includes small, midsize, and large territories in an Urbanized Area and inside a Principal City; 'Suburban' ($n = 73$) includes small, midsize, and large territories outside a Principal City and inside an Urbanized Area; 'Town' ($n = 50$) includes remote, distant, and fringe territories inside Urban Clusters that are some distance from an Urbanized Area (e.g. Town – Remote is more than 35 miles from an Urbanized Area); and 'Rural' ($n = 21$) includes remote, distant, and fringe census-defined rural territories that are some distance from an Urban Cluster (e.g. Rural – Remote is more than 25 miles from an Urbanized area and more than 10 miles from an Urban Cluster). The results showed that there were significant main effects of locale on STEM program quality in the Engagement with STEM dimension (χ^2 (6) = 12.69, $p < .048$) and the Youth Voice dimension (χ^2 (6) = 13.81, $p < .032$) by locale. However, *post hoc* analyses using Kruskal–Wallis multiple comparisons tests revealed no significant differences between locales, suggesting that the main omnibus test produced a 'false alarm' – potentially due to insufficient power given smaller sample sizes within the four locale groups, with a particularly small sample size for rural programs (Chen et al., 2018).

Year type differences

A Mann–Whitney U test was conducted to examine differences in STEM program quality based on the time of year activities took place. The results showed that there were significant differences in quality ratings between STEM activities that took place during the summer (i.e. activities facilitated between June 16th and August 31st, $n = 101$) compared with STEM activities that took place during the school year (i.e. activities facilitated between September 1st and June 15th, $n = 343$). The results indicated that summer programing had significantly higher ratings in select DoS dimensions than

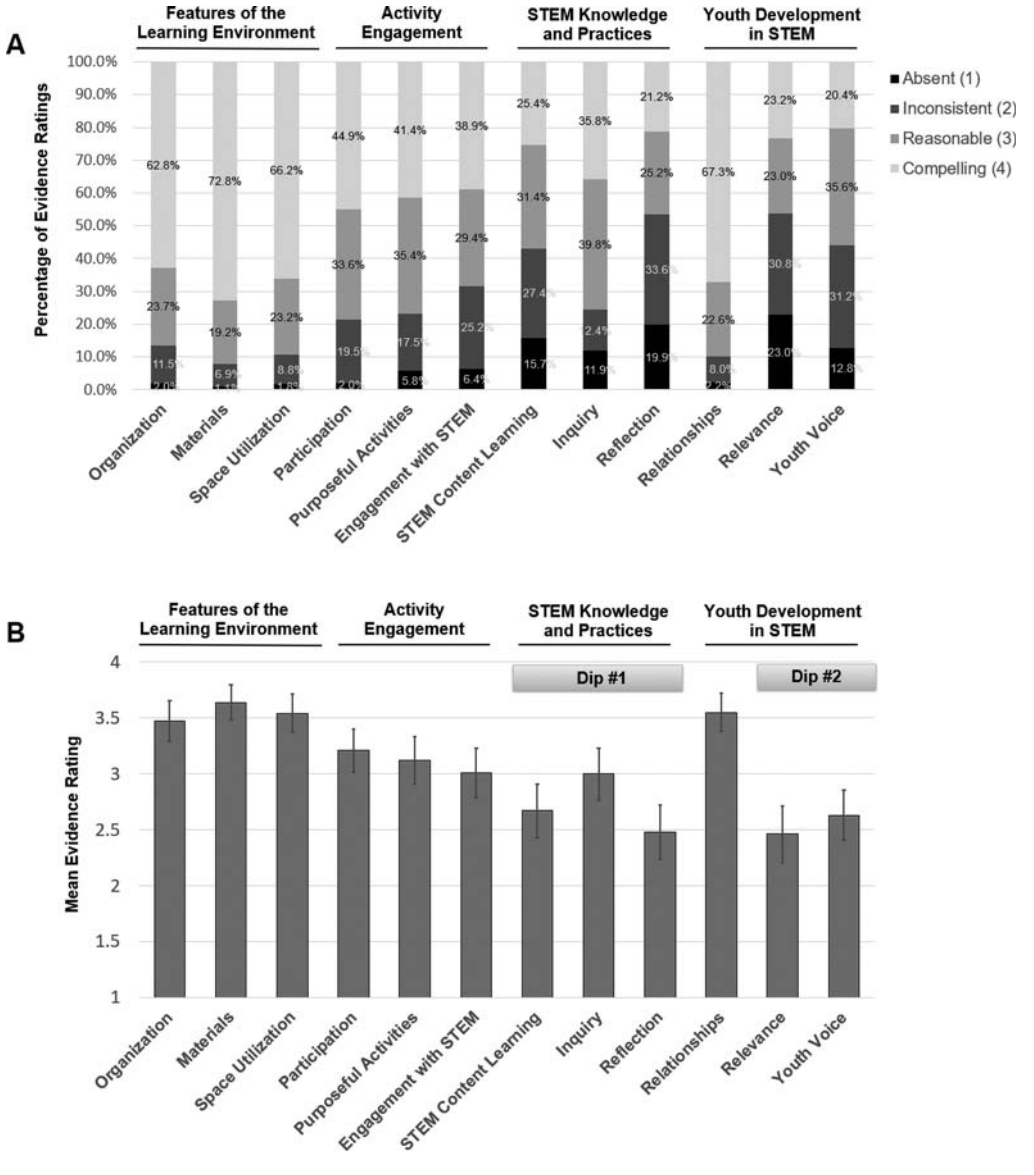

Figure 1 National Dimensions of Success (DoS) STEM program quality observation data displayed as (a) stacked bar chart showing proportion of segment-level scores for each dimension of quality, where darker colors illustrate the pattern of challenges (lower quality ratings, 1's and 2's) and lighter colors illustrate the pattern of strengths (higher quality ratings, 3's and 4's), and (b) mean (± SE) evidence ratings for each dimension of quality, where average scores that are equal to or greater than 3 represent reasonable evidence of quality (i.e. the criterion threshold for high quality using DoS). SE for each dimension was multiplied by a factor of five to increase visibility of error bars.

school year programing, including: Reflection (Summer $Mdn = 3.00$, School $Mnd = 2.00$, $U = 19{,}978.50$, $p = .015$), Relevance (Summer $Mdn = 3.00$, School $Mnd = 2.00$, $U = 22{,}775.50$, $p = . < 001$), and Youth Voice (Summer $Mdn = 3.00$, School $Mnd = 3.00$, $U = 20{,}360.00$, $p = .005$). Notably, while summer activities scored higher in these dimensions, they still fell below the criterion threshold for quality for four dimensions: STEM Content Learning ($M = 2.80$, $SD = 1.00$), Reflection ($M = 2.69$, $SD = 1.00$), Relevance ($M = 2.94$, $SD = 1.00$), and Youth Voice ($M = 2.87$, $SD = 1.01$). This pattern parallels the challenges to quality observed across all observations, reported for overall trends above, with the exception of Inquiry.

Table 2. Regional Differences in STEM Program Quality based on DoS Observations.

	Quality Levels by Region				Test Statistics		
Variable	NE	MW	S	W	Stat.	Signif.	Between-Groups
Features of the Learning Environment	***M (SD)***	***M (SD)***	***M (SD)***	***M (SD)***	χ^2 (3)	*p*	Pairwise
Organization	3.33 (0.77)	3.38 (0.91)	3.59 (0.67)	3.64 (0.59)	10.7	0.013	NE:S*
Materials	3.55 (0.63)	3.62 (0.73)	3.65 (0.67)	3.81 (0.44)	7.91	0.048	W:NE*
Space Utilization	3.48 (0.73)	3.59 (0.72)	3.53 (0.76)	3.55 (0.67)	2.18	0.535	N/A
Activity Engagement							
Participation	3.04 (0.80)	3.27 (0.78)	3.27 (0.86)	3.25 (0.85)	7.26	0.064	N/A
Purposeful Activities	3.07 (0.82)	2.99 (0.97)	3.23 (0.91)	3.26 (0.74)	6.95	0.074	N/A
Engagement With STEM	2.90 (0.92)	2.99 (0.96)	2.98 (0.94)	3.34 (0.96)	9.94	0.019	NE:W*, W:S*
STEM Knowledge and Practices							
STEM Content Learning	2.69 (0.94)	2.47 (1.06)	2.80 (1.01)	2.75 (1.05)	7.53	0.57	N/A
Inquiry	2.91 (0.84)	2.94 (1.01)	2.97 (1.08)	3.36 (0.76)	9.91	0.019	W:NE*
Reflection	2.47 (0.96)	2.44 (1.11)	2.54 (1.03)	2.42 (1.05)	0.97	0.809	N/A
Youth Development in STEM							
Relationships	3.53 (0.73)	3.42 (0.82)	3.67 (0.64)	3.60 (0.74)	10.19	0.017	MW:S*
Relevance	2.43 (0.91)	2.24 (1.09)	2.73 (1.11)	2.38 (1.18)	16.06	0.001	S:MW**
Youth Voice	2.47 (0.89)	2.49 (0.92)	2.82 (0.98)	2.81 (0.92)	14.7	0.002	S:NE*, S:MW*

Note. DoS dimensions are rated on a 4-point scale from 1 (evidence absent) to 4 (compelling evidence), where the goal is to achieve an average rating of 3 (reasonable evidence) or higher. Sample size for mean STEM program quality by region was as follows: Northeast (NE), $n = 100$, Midwest (MW), $n = 144$, South (S), $n = 153$, West (W), $n = 53$. There were statistically significant differences between regions for seven DoS dimensions. The asterisk (*) in the Pairwise column denotes the region with higher mean rating (i.e. Region A*:Region B indicates that Region A had a significantly higher mean dimension rating relative to Region B, whereas Region A:Region B* indicates that Region B had a significantly higher mean dimension rating relative to Region A.)

Grade level

A Mann–Whitney U test revealed that STEM activities conducted with middle to high school-aged students (Grades 6–12, $n = 117$) scored significantly higher in the Youth Voice dimensions than STEM activities conducted with elementary school-aged students (Grades K-5, $n = 122$) (Elementary $Mdn = 2.00$, Middle to High School $Mnd = 3.00$, $U = 8{,}733.50$, $p = .002$). However, despite scoring higher in Youth Voice, STEM activities conducted with middle to high school-aged youth still fell below the criterion threshold rating of 3.0 in this dimension ($M = 2.80$, $SD = .89$). Activities conducted with middle and high school aged youth also fell below the criterion threshold rating of 3.0 in STEM Content Learning ($M = 2.7$ $SD = 1.03$), Inquiry ($M = 2.93$, $SD = 1.01$), Reflection ($M = 2.44$, $SD = 1.00$), and Relevance ($M = 2.44$, $SD = 1.08$), mirroring the double-dip observed across all programs reported in the overall trends. In addition, programs serving middle and high school aged youth scored below a 3.0 in the Engagement with STEM dimension (M=2.98, SD=.97).

Gender differences

To determine if there were differences in STEM program quality based on the gender ratio of the students, observation data was divided by STEM activities where girls represented the majority of participants (i.e. 80%–100% girls, $n = 49$ observations) and the minority of participants (0%–20% girls, $n = 40$ observations). The results of a Mann–Whitney U test showed that STEM activities facilitated with a majority of girls received significantly higher ratings in three dimensions: Materials (Majority $Mdn = 4.00$, Minority $Mdn = 4.00$, $U = 1{,}165.00$, $p = .043$), Participation (Majority $Mdn = 4.00$, Minority $Mdn = 3.00$, $U = 1{,}208.00$, $p = .039$), and Relationships (Majority $Mdn = 3.00$, Minority $Mdn = 4.00$, $U = 1{,}174.50$, $p = .042$). This pattern of strengths for majority girl attended activities is consistent with the pattern of strengths observed across all observations, reported in the overall trends above. Furthermore, while activities facilitated with a minority of girls (i.e. 80%–100% boys) received significantly lower quality ratings in these three dimensions than

activities facilitated with a majority of girls, the average quality ratings for minority girl/majority boy activities exceeded the threshold criterion for quality of 3.0: Materials ($M = 3.55$, $SD = .68$), Participation ($M = 3.10$, $SD = .84$), and Relationships ($M = 3.38$, $SD = .93$), meaning that that activities were meeting STEM program quality standards for these three dimensions (Materials, Participation, and Relationships) regardless of the gender ratio of participants.

Discussion

Ensuring that OST programs facilitate high quality STEM activities is essential to maximize the impact of the field on children's STEM learning, attitudes, and skill development (Allen et al., 2019; National Research Council, 2015; Shah et al., 2018). This national study, which draws upon a large sample of observation data provided by OST STEM programs from across the U.S., enabled us to identify and communicate common strengths and challenges as defined by the DoS framework. Consistent with our predictions, we observed a 'double-dip' in program quality ratings across the 12 DoS dimensions, such that dimensions related to the learning environment and relationships rated higher than those dimensions related to STEM knowledge and practices and youth ownership and connection of learning to real world experiences. Furthermore, we identified areas of quality that differ by specific program and participant characteristics, including region, season, grade distribution of activity participants, and gender proportions of activity participants. Notably, despite differences in the magnitude of average quality ratings by program and participant characteristics, the presence and direction of the double-dip persisted. We focus our discussion on these key study findings and their implications for future research, practice, and policymaking.

National strengths and challenges

Identifying common strengths and areas for growth and communicating these outcomes in a productive, strengths-based manner is important for advancing the OST STEM field as a whole, and especially for enhancing the individual experiences of youth engaging in STEM learning (Allen et al., 2019). OST STEM programs, including those in our sample, vary widely in their focus, duration, setting, and goals, but analyzing and aggregating data using DoS, a common measure, allowed us to pinpoint several strengths and areas for improvement across a wide variety of programing. DoS dimensions focused on features of the learning environment and relationships scored higher than dimensions focused on STEM knowledge and practices, which is consistent with our hypothesis and previous research (Allen et al., 2019; Shah et al., 2018). The strengths identified should not be taken for granted, as historically, OST programs were not considered informal just in terms of their casual nature, but also in terms of preparation as no state teaching requirements or testing systems are in place. These data highlight the increasing professionalism of the OST sector for facilitating academic-focused activities in a supportive environment, but also highlight the need for more targeted professional development and training specific to STEM content learning and instruction to build knowledge and confidence among educators as well as deep and meaningful learning among youth. This does not mean that STEM in OST should be like STEM in school and the consensus is that it should not be graded and become 'high stakes.'

Exploring the strongest dimensions in closer detail, we found that most programs – nine out of every ten – met or exceeded standards for the Materials dimension, meaning they are using materials that are appealing to youth, age-appropriate, culturally sensitive, and appropriate for achieving the STEM learning goal. For added context, we provide an excerpt of evidence submitted by a certified DoS observer who visited a summer program located in an urban area in the South that served elementary school aged youth and implemented the 'Engineering is Elementary' curriculum. The following evidence supports a higher-quality rating of 4 (compelling evidence) in the Materials dimension:

> All of the materials are appropriate for supporting the STEM learning goals. The students are learning about engineering terms like force, stability, and gravity through playing with notecards and observing the stability of a paper structure. The students are having fun the entire lesson, yelling excitedly when they make the notecards stand on their own and exclaiming out loud when the paper structure falls over. They are excited and yell a little bit when they get to put the index cards together to make a triangle. The students are engaged with the materials in each stage of the lesson and eager to participate, indicating the materials are also appealing for the students.

In addition to describing what a high-quality rating in the Materials dimension encompasses, the evidence presented above highlights a crucial point: materials do not need to be expensive or 'high-tech' to be deemed high-quality. This is particularly salient as many informal STEM programs have limited funding and resources. With the appropriate curriculum, training, and/ or support, high-quality STEM activities can be enacted with simple, everyday materials such as index cards. Furthermore, programs can use their strength in Materials to help bolster other dimensions such as Participation (e.g. appealing and appropriate materials may help youth access an activity) and Purposeful Activities (e.g. learning goal appropriate materials can lead youth towards an understanding of that goal) among others.

We also found that most OST programs excelled in promoting positive relationships between children and their peers as well as between children and the adults leading activities. This finding may be explained by the fact that OST settings are often relationship focused – as they encourage teamwork and collaboration during STEM activities – and many OST facilitators often have a background in youth development practices – such as adults acting as mentors to help youth cope with the social and emotional challenges they face (Fredricks et al., 2014; Lyon et al., 2012; Shah et al., 2018). Many programs consider the social environment – trusting and supportive relationships between staff and youth – as one of most important features of an engaging program (Fredricks et al., 2014; Naftzger et al., 2018). For instance, the primary goal of many afterschool programs is to provide youth with a safe space outside of school – during the hours of 3:00pm–6:00pm – to foster academic learning and social-emotional development in youth and provide support to families in the community (Afterschool Alliance, 2014). For added context from the present study, we provide an excerpt of evidence submitted by a certified DoS observer who visited an afterschool program for girls in the South conducted during the school year that used the Operation Smart curriculum. The following evidence supports a high-quality rating of 4 (compelling evidence) in the Relationships dimension:

> Interactions among youth and between the facilitator and youth are consistently positive, creating a warm and friendly learning environment. The youth are noted several times to be going around the room and sharing their playdough with each other. Girls are noticed to be helping each other and are encouraged by the facilitator to help each other. Facilitator is consistent with tones, mannerisms and behaviors with the girls. The girls seem comfortable approach facilitators and answering questions, by raising their hand or shouting out answers.

On the other hand, examining growth areas in closer detail revealed that many programs did not adequately or consistently implement best practices related to the Reflection, Relevance, STEM Content Learning, and Youth Voice dimensions, respectively. Practitioners and policymakers may want to prioritize its focus on Reflection and Relevance first, as one out of every two activities observed did not meet the standard for quality set by DoS. Furthermore, there is evidence to suggest that these dimensions of quality can support growth in STEM Content Learning and Youth Voice. Thus, we primarily focus this discussion on Reflection and Relevance, providing evidence for this common challenge as well as evidence that it is possible to meet the high standards set by DoS.

Reflection has been previously identified as an area for improvement in U.S. OST STEM programs (Allen et al., 2019; Shah et al., 2018). Lower quality ratings in the Reflection dimension indicates that youth are not engaging in reflection processes that develop deeper understanding of STEM content and connections between STEM content ideas. For added context, consider the following example of evidence submitted by a certified DoS observer who visited an afterschool

program in the Northeast serving middle school aged youth that focused on life sciences. The following evidence supports a lower-quality rating of 1 (evidence absent) in the Reflection dimension:

> Facilitators provided no opportunity for students to connect ideas across activities or reflect on what they were doing in the activities, why they were doing them, and the relevance to DNA. Students could not use the experience to make sense of the activities they were doing.

The above evidence is representative of the lower scoring observations and suggests that it is not merely the depth of student reflection that is impacting ratings, but the lack of reflective prompts used by the facilitator.

Providing opportunities for reflection is an essential component in allowing youth to process and organize their learning (Davis, 2000, 2003; Shah et al., 2018). Reflection is essential if students are to become aware of themselves as competent and confident learners and doers in the realms of science and engineering (NGSS Lead States, 2013). Furthermore, reflection during activities provides facilitators with a clear picture of youths' levels of understanding, and thus allows them to make informed decisions around how to scaffold youth learning to be sure concepts are fully and accurately understood. A focus on improving the Reflection dimension would likely improve other related dimensions, and in particular, STEM Content Learning. By providing ample opportunities for youth to engage in reflection, youth are able to solidify their learning, begin to make connections amongst STEM content ideas, and articulate their learning, thus bolstering the rating in this dimension. While the Reflection dimension represents a challenge for the field generally, many activities excelled in this area. For example, the following evidence supports a higher-quality rating of 4 (compelling evidence) in the Reflection dimension based on an observation submitted by a certified DoS observer who visited a summer program in the West conducting a STEM activity that was investigating the technology used to build hovercrafts.

> The Learn & Grow students were engaged at reflection time and the facilitator was not always the one who initiated the facts. Student #157 stated that the more air you put inside the balloon the longer the lift off. 'It's like a real hovercraft but smaller.' Student #78 stated it takes force to move objects which is why you have to push the CD to get to go forward after the air from the balloon lifts it up.

Similar to Reflection, the Relevance dimension has been previously identified as an area for improvement in many U.S. OST STEM programs (Allen et al., 2020; Shah et al., 2018). Lower quality ratings in the Relevance dimension indicate that youth are not being consistently prompted to connect what they are learning to the real world (e.g. community concerns, personal interests) to help them see that STEM is meaningful and important to their lives. Lower quality ratings are received when activities are conducted in isolation, preventing youth from relating STEM content outside the learning setting. For example, a certified DoS observer visited a youth program whose facilitator was leading an engineering activity that required youth to build the tallest, free standing, tower possible with a solid base. The following evidence supports a lower-quality rating of 1 (evidence absent) in the Relevance dimension:

> The facilitator was not observed in making attempts to connect the activity to youths' lives and/or broader context. There was no reference to the building they were in, the tallest building they've ever seen. Careers where they might use this information, etc.

Connecting activities to broader contexts makes otherwise intangible content more personally meaningful and can help foster a sense of STEM identity and belonging with the community (Allen et al., 2019; Aschbacher et al., 2014; Vincent-Ruz & Schunn, 2018). For example, a recent longitudinal study found that 9th grade girls have significantly higher expectations to major in STEM fields, compared to non-STEM fields, when they have positive perceptions of the social relevance of science (Blanchard Kyte & Riegle-Crumb, 2017).

Furthermore, the data collected as part of this study show that facilitating activities with high levels of relevance is possible. The following evidence supports a higher-quality rating of 4 (compelling evidence) in the Relevance dimension based on evidence submitted by a certified DoS

observer who visited a summer program in an urban area of the South conducting an activity on environmental engineering.

> The facilitator and students are actively involved in discussing the relevance of the activity to their broader interest, the environment. The students are encouraged to think of the feathers used in the experiment as the feathers on a real bird impacted by the oil spill. The students are taught about the oil spill in the Gulf of Mexico before performing the experiment, and most appear to relate the experiment to the real impact of oil on marine wildlife: 'Birds have to dive into the water and they get clean under the oil, but when they come back up, the oil gets into their feathers.' Many students appear concerned about the impact of the oil spill on the health and wellness of the birds who hunt for food in the ocean.

By strengthening the relevance of an activity, programs may also strengthen the Youth Voice dimension. Creating connections between the activity at hand and the outside world can create a sense of buy in from youth. When activities are personally meaningful, youth may be more likely to take ownership in their work and seek directive power in their learning.

Program characteristics

To generate hypotheses around the relationship between the 'double-dip' in quality and various program factors, we examined whether region, locale, and the time of year factored into the pattern of quality ratings observed. Interestingly, we found regional differences such that programs located in the West and South showed more consistent evidence of quality than other regions, in the Organization, Materials, Engagement with STEM, Inquiry, Relationships, Relevance, and Youth Voice dimensions. It is unclear, based on available data, why these regional patterns emerged, but differences may be related to the level of training, resources, and support that programs receive within individual states and communities. For example, there are several national initiatives that focus on improving systems that support the quality of OST STEM programing, including the system-building state afterschool networks (Mott Foundation and STEM Next, 2018) and the STEM Learning Ecosystems Community of Practice (Allen et al., 2020; Traill et al., 2015; Traphagen & Traill, 2014). States and communities that participate in these initiatives tend to have different areas of focus and different levels of investment when it comes to professional development and quality (Allen et al., 2019; Mott Foundation and STEM Next, 2018). Further research is needed to contextualize these regional differences, as well as other program characteristics, but these data suggest that regional investigations can provide meaningful insight to inform research, policy, and practice.

Another notable point when considering regional differences is that southern and western programs still exhibited the same pattern of challenges observed at the national level – the 'double-dip.' Higher quality, on average, does not equate to sufficient quality, and regions and networks must continue to learn from and support one another by sharing curriculum, professional development opportunities, among other resources. The U.S. federal government has recognized the development and enhancement of strategic partnerships through systems as one of the key pathways to success in its latest five-year strategic plan for STEM education (National Science & Technology Council, 2020), and the use of evidence-based, shared measures can greatly aid partnership development by improving communication around STEM program quality and outcomes within and across communities (Grack Nelson et al., 2019).

One other notable finding for program characteristics includes timing: STEM activities conducted during the summer months exhibited higher levels of quality, on average, than activities observed during the school year. These results may be explained by the greater variety of programing offered to youth during the summer. During the school year, youth are less likely to have options to choose from and enroll in what is available despite levels of interest. During the summer, families are more likely to consider a wider variety of programing outside of school settings and provide youth more choice around what they participate in. These programs often have varied funding sources, including industry, state, city, and parent funding. It is possible that these sources of funding influence program quality scores. Furthermore, summer programing may be more

flexible, with fewer time restraints, allowing for youth to take a more directive role in their learning. Additionally, summer programing occurs in a wide variety of places beyond school buildings, including community parks and recreation centers, beaches and trails, libraries, museums, and more (National Summer Learning Association, 2018). As with regional differences, more contextual information and additional research is needed to provide more meaningful insight into patterns of quality. Despite differences in levels of quality by time of year, activities that take place during the summer still exhibit the double-dip.

Participant characteristics

Interestingly, the demographic composition – particularly gender and age – of the youth participating in programing was associated with levels of program quality. The finding that activities enacted with older youth scored higher in Youth Voice than those activities conducted with youth in elementary school is consistent with developmental theory (Mitra, 2004; Noam & Triggs, 2018). It is plausible that facilitators are more comfortable giving youth in middle and high school more directive power and leadership roles in their learning, as compared to those youth in elementary school, because they require less behavioral management, are more cognitively mature, but also because adolescents have a stronger desire to express themselves independently. Elementary school-aged children require more support and direction to make decisions and direct their learning within facilitator-set confines. It is likely that programs serving elementary-aged youth would score higher in Youth Voice with additional guidance and/or professional development on developmentally appropriate ways to give youth decision making powers in OST settings. Young students also need to experience leadership and a sense of agency in their learning and a significant lack has been revealed.

There were also gender differences detected for two dimensions – Participation and Relationships – that are also consistent with previous literature, which indicates that girls perform better and are more engaged in STEM when gender differences are not highlighted (Danaher & Crandall, 2008; Spencer et al., 1999). Other recent research has suggested that facilitators of science activities spend more time interacting with male students than their female peers (Shumow & Schmidt, 2013), suggesting that girls may be more engaged in an activity, or by the facilitator, when fewer boys are present. It is less clear why there were differences in ratings in the Materials dimension, but is it plausible that a greater level of effort is made in the design and selection of materials for programs that serve a majority of girls (such as all girl programs), which may have a specific mission to attract and engage girls to address the issue of gender gaps in STEM fields.

Limitations and future directions

This paper focuses on the importance of identifying and communicating strengths and challenges to advance the field of OST STEM, and the authors apply the same continuous improvement lens to this study. While we replicated the 'double-dip' finding and further demonstrated that it persists across a variety of program and youth level characteristics, we acknowledge that there are limitations of the present study that need to be addressed. We consider these in the context of ongoing and future efforts to advance practice and research on OST STEM program quality.

Notably, the present findings are based on a sample of convenience; programs and observers submitted data voluntarily to the lead developers of DoS. It is possible that programs submitting data are more interested in being observed because they themselves place a greater emphasis on STEM program quality improvement or are involved in systems-building work. If this sampling bias exists (i.e. that the data come from a more supported group of programs or a group of programs that more intentionally focus on quality), we expect that the current results are over-estimating levels of STEM quality for one or more dimensions, suggesting that the magnitude of the double-dip – and the challenges facing the field – may be greater than this study reveals.

Additionally, more information is needed at the systems, program, and youth levels to better understand what factors influence quality, to develop a model that can be used to study the complex relationships between the many program, educator, and youth factors related to quality, and to support the creation of a representative sample of programs.

While we performed this study to provide a deeper look into program quality observations across the U.S., we have several projects planned or underway to help the field understand how various factors influence OST STEM program quality. We are working to create a representative sample of program quality – using randomized sampling techniques that consider youth, educator, program, and community/systems data – to remove selection bias and to obtain a more precise estimate of quality. There are many sources of information at the youth level (e.g. attitudes, skills, attendance, demographics, familial support), the educator level (e.g. self-efficacy, confidence, hours of training or professional development, demographics), the program level (e.g. curriculum, budget, resources, staffing, materials), and the systems level (e.g. state or ecosystem support, connections with schools) that can be used to understand and improve program quality. In addition, we are working to align DoS with educational priorities, including improving the rubrics by increasing measurable practices of diversity, equity, inclusion, and access as these are paramount to understanding the quality of any given program for all learners. To this end, we envision the use of multiple measures to develop a common database that allows for the aggregation of common sources of data from many informants (including youth, educators, and families) – and from across different sectors of the community – to create complex models that can better inform the fields of research, practice, and policy (Grack Nelson et al., 2019). Studying the relationships between STEM program quality and other sources of information that are proximal to the learning environment (including from educators, learners, and their interactions) can help to create a more holistic understanding of how factors – such as demographics, culture, family, diversity and inclusion, among others – influence the observable strengths and challenges of STEM programing and levels of participant engagement in STEM activities.

To truly advance the field, common measures and trainings – including for program quality – need to be supported by and disseminated widely across systems. This will be especially beneficial as system-building initiatives build bridges across many different sectors of communities – including schools and OST programs – and as the desire for shared measures to communicate quality and youth outcomes is increasing across communities (Allen et al., 2019, 2020, in press; Traill et al., 2015). For example, the field may benefit from connecting OST STEM learning with the school day, as is happening in many communities across the country. To support these cross-sector efforts, we have modified the DoS rubrics for school settings to further increase measurement and communication around the quality of STEM activities across school and OST settings, and we will continue to study factors that impact quality within and across different STEM learning settings. To begin to address the 'double-dip,' we have developed trainings in Reflection, Relevance, and Youth Voice, and a study is planned to examine associations between participation in training and changes in program quality. However, it is important to recognize that OST programs struggle with high rates of staff turnover (McGuiness-Carmichael, 2019). Thus, targeted professional development will be most beneficial when provided to staff members who continue to work with the program over longer periods of time so the impacts of training on persistently challenging areas of quality can be measured over longer periods of time.

Lastly, observational protocols and studies are essential for understanding the strengths and challenges of OST STEM programing in systematic ways. DoS has largely been used in practice – by system-builders, program directors, frontline staff, and evaluators – to support continuous improvement efforts. For this reason, less empirical observation research has been conducted or published by others aside from the creators of DoS. The authors invite the OST research community to contribute to the study of OST STEM program quality, using DoS, to better understand the progress and impact of the OST field's many efforts to improve quality, especially to link program quality to other factors that support youth into and through STEM pathways.

Conclusions

This study identified key trends in STEM program quality on a national scale – a characteristic 'double-dip' in programmatic strengths and challenges that persist regardless of different program-related (i.e. region, locale, time of year) and participant-related (i.e. grade, gender) factors. These data were collected using a standardized, evidence-based observation tool that supports research, practice, and policy by generating hypotheses for future studies, by communicating the quality of informal STEM learning experiences, and by informing funding decisions and professional development strategies. Continuously updating and analyzing data collected using shared measures like DoS will help researchers, practitioners, funders, and policymakers monitor the OST STEM field's progress and make decisions to move the field toward higher-quality STEM experiences. Importantly, STEM program quality is positively linked to youth outcomes in STEM and social-emotional learning (Allen et al., 2019; Fenton et al., 2019). While it is important to celebrate the strengths of OST STEM programs, the 'double-dip' underscores the urgent need to act to address the widespread and persistent challenges in OST STEM program quality across the U.S. We highlighted examples of programs that have special strengths in rubrics that are typically weak, and by using shared measures like DoS, these strengths can be communicated to other programs, schools, and other key stakeholders – such as through communities of practice, professional learning communities, ecosystems, and other system-building initiatives. Ultimately, continued efforts to ensure higher quality STEM activities in all learning contexts will allow more youth to experience the joys of scientific discovery and will enable more youth to persist and succeed in STEM.

Acknowledgements

The development and study of the Dimensions of Success was supported by funding from the National Science Foundation (Award Number 1008591). We would also like to thank the Charles Stewart Mott Foundation, the Noyce Foundation, and the STEM Next Opportunity Fund for funding studies that increased the evidence base for this tool. The views expressed in this paper are those of the authors and do not represent the views of the funding organizations or affiliated institutions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The development and study of the Dimensions of Success was supported by funding from the National Science Foundation (Award Number 1008591); Charles Stewart Mott Foundation; Noyce Foundation; STEM Next Opportunity Fund.

ORCID

Rebecca K. Browne http://orcid.org/0000-0002-0598-5561
Patricia J. Allen http://orcid.org/0000-0001-8753-9938

References

Allen, P. J., Brown, Z., & Noam, G. G. (in press). STEM learning Ecosystems: Building from theory toward a common evidence base. *International Journal for Research on Extended Education*, *8*(1), 80–96. https://doi.org/10.3224/ijree.v8i1.07

Allen, P. J., Chang, R., Gorrall, B. K., Waggenspack, L., Fukuda, E., Little, T. D., & Noam, G. G. (2019). From quality to outcomes: A national study of afterschool STEM programming. *International Journal of STEM Education*, *6*(1). https://doi.org/10.1186/s40594-019-0191-2

Allen, P. J., Lewis-Warner, K., & Noam, G. G. (2020). Partnerships to transform STEM learning: A case study of a STEM learning ecosystem. *Afterschool Matters*, *31*, 30–41.

Alliance, A. (2014). *America after 3PM: Afterschool programs in demand*. Afterschool Alliance.

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. https://doi.org/10.1007/s10956-014-9504-x

Bell, C., Qi, Y., Croft, A. J., Leusner, D. W., McCaffrey, D., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. 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. 50–97). Jossey-Bass.

Blanchard Kyte, S., & Riegle-Crumb, C. (2017). Perceptions of the social relevance of science: Exploring the implications for gendered patterns in expectations of majoring in STEM fields. *Social Sciences*, *6*(1), 19. https://doi.org/10.3390/socsci6010019

Chen, T., Xu, M., Tu, J., Wang, H., & Niu, X. (2018). Relationship between omnibus and post-hoc tests: An investigation of performance of the F test in ANOVA. *Shanghai Archives of Psychiatry*, *30*(1), 5.

Chittum, J. R., Jones, B. D., Akalin, S., & Schram, ÁB. (2017). The effects of an afterschool STEM program on students' motivation and engagement. *International Journal of STEM Education*, *4*(1), 11. https://doi.org/10.1186/s40594-017-0065-4

Committee on STEM Education of The National Science & Technology Council. (2020). Charting a Course For Success: America's Strategy for STEM Education. Retrieved December 23, 2020, from https://www.whitehouse.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf.

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings re-examined. *Journal of Applied Social Psychology*, *38*(6), 1639–1655. https://doi.org/10.1111/j.1559-1816.2008.00362.x

Davis, E. A. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education*, *22*(8), 819–837. https://doi.org/10.1080/095006900412293

Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *Journal of the Learning Sciences*, *12*(1), 91–142. https://doi.org/10.1207/S15327809JLS1201_4

Fenton, M. P., Hawley, L., Frerichs, S. W., & Lodl, K. (2019). STEM professional development for youth workers: Results of a Triangulated study. *Journal of Youth Development*, *14*(4), 178–196. https://doi.org/10.5195/jyd.2019.738

Fredricks, J. A., Bohnert, A. M., & Burdette, K. (2014). Moving beyond attendance: Lessons learned from assessing engagement in afterschool contexts. *New Directions for Youth Development*, *2014*(144), 45–58. https://doi.org/10.1002/yd.20112

Friedman, A. (2008). *Framework for evaluating impacts of informal science education projects*. Retrieved from http://www.auraastronomy.org/news/EPO/eval_framework.pdf

Funk, C., & Hefferon, M. (2016). As the need for highly trained scientists grows, a look at why people choose these careers. Retrieved December 23, 2020, from https://www.pewresearch.org/fact-tank/2016/10/24/as-the-need-for-highly-trained-scientists-grows-a-look-at-why-people-choose-these-careers/

Grack Nelson, A., Goeke, M., Auster, R., Peterman, K., & Lussenhop, A. (2019). Shared measures for evaluating common outcomes of informal STEM education experiences. *New Directions for Evaluation*, *2019*(161), 59–86. https://doi.org/10.1002/ev.20353

Krishnamurthi, A., Ottinger, R., & Topol, T. (2013). STEM learning in afterschool and summer programming: An essential strategy for STEM education reform. In T. K. Peterson (Ed.), *Expanding minds and opportunities: Leveraging the power of afterschool and summer learning for students*. http://www.expandinglearning.org/expandingminds/article/stem-learning-afterschooland-summer-programming-essential-strategy-stem

Lyon, G. H., Jafri, J., & St. Louis, K. (2012). Beyond the pipeline: STEM pathways for youth development. *Afterschool Matters*, *16*, 48–57.

Mahoney, J. L., Levine, M. D., & Hinga, B. (2010). The development of after-school program educators through university-community partnerships. *Applied Developmental Science*, *14*(2), 89–105. https://doi.org/10.1080/10888691003704717

Maltese, A. V., & Tai, R. H. (2010). Eyeballs in the fridge: Sources of early interest in science. *International Journal of Science Education*, *32*(5), 669–685. https://doi.org/10.1080/09500690902792385

McGuiness-Carmichael, P. (2019). Youth perspectives on staff turnover in afterschool programs. *Afterschool Matters*, *30*, 19–23.

Mitra, D. L. (2004). The significance of students: Can increasing "student voice" in schools lead to gains in youth development? *Teachers College Record*, *106*(4), 651–688. https://doi.org/10.1111/j.1467-9620.2004.00354.x

Mott Foundation and STEM Next. (2018). *STEM in afterschool system-building toolkit*. http://expandingstemlearning.org/

Naftzger, N., Sniegowski, S., Smith, C., & Riley, A. (2018). *Exploring the relationship between afterschool program quality and youth development outcomes: Findings from the Washington quality to youth outcomes study* (pp. 1–47). American Institutes for Research. https://raikesfoundation.org/sites/default/files/washington-quality-youth-outcomes-study.pdf

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. National Academies Press. doi:10.17226/11463.

National Research Council. (2015). *Identifying and supporting productive STEM programs in out-of-school settings*. National Academies Press. https://doi.org/10.17226/21740

National Research Council (NRC). (2009). *Learning science in informal environments: People, places, and pursuits*. Committee on learning science in informal environments. In P. Bell, B. Lewenstein, A. W. Shouse, & M. A. Feder (Eds.), *Board on science education, center for education. Division of behavioral and social sciences and education*. The National Academies Press. https://doi.org/10.17226/12190

National Summer Learning Association Report. (2018). *Summer opportunities: A research agenda*. https://www.summerlearning.org/wp-content/uploads/pdf/NSLA-Research-Agenda-Final.pdf

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

Noam, G. G., Allen, P. J., Shah, A. M., & Triggs, B. B. (2017). Innovative use of data as game changer for afterschool: The example of STEM. In H. J. Malone & T. Donahue (Eds.), *The growing out-of-school time field: Past, present, and future* (pp. 166–176). Information Age Publishing.

Noam, G. G., & Triggs, B. (2018). *The clover model: A developmental process theory of social-emotional development*. The PEAR Institute: Partnerships in Education and Resilience.

Shah, A. M., Wylie, C., Gitomer, D., & Noam, G. (2018). Improving STEM program quality in out-of-school-time: Tool development and validation. *Science Education*, *102*(2), 238–259. https://doi.org/10.1002/sce.21327

Shumow, L., & Schmidt, J. A. (2013). Academic grades and motivation in high school science classrooms among male and female students: Associations with teachers' characteristics, beliefs and practices. *Journal of Education Research*, *7*(1), 53–71.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Traill, S., Traphagen, K., & Devaney, E. (2015). Assessing the impacts of STEM learning ecosystems: Logic model template & recommendations for next steps. *Noyce Foundation*. http://stemecosystems.org/

Traphagen, K., & Traill, S. (2014). How cross-sector collaborations are advancing STEM learning. *Noyce Foundation*. http://stemecosystems.org/resource-category/key-resources/

Vincent-Ruz, P., & Schunn, C. D. (2018). The nature of science identity and its role as the driver of student choices. *International Journal of STEM Education*, *5*(48), 1–12. https://doi.org/10.1186/s40594-018-0140-5