Chapter Three: Shared Measures for Evaluating Common Outcomes of Informal STEM Education Experiences

Amy Grack Nelson, Megan Goeke, Ryan Auster, Karen Peterman, and Alexander Lussenhop

Abstract

Since the late 2000s, interest in the development and use of shared measures in the informal science, technology, engineering, and mathematics (STEM) education (ISE) field has increased. The intent is to build the capacity of evaluators to measure common outcomes of ISE experiences. We begin this chapter with a definition of shared measures, a description of related technical qualities of these measures, and a discussion of benefits and concerns around the use of shared measures. We then review recent conversations and developments around shared measures, including examples of observational and survey tools to measure common ISE outcomes. Three case studies of shared measure efforts provide an in-depth look into the development of shared survey measures and highlight promising lessons for future initiatives. Building from these successes, we outline recommendations for the future of shared measures in ISE. These include adopting a multifaceted approach to enhancing measurement capacity among evaluators, supporting new kinds of collaborative work among evaluators, and increasing access to instruments and tools.

The development and use of shared measures to evaluate common outcomes has been gaining increased attention in the informal science, technology, engineering, and mathematics (STEM) education (ISE) field, as described by Allen and Peterman in Chapter One. Extending that discussion, this chapter begins with a definition of shared measures and an overview of their technical qualities. It then covers recent field-wide conversations and initiatives around shared measures for ISE evaluation with a focus on case studies of three projects. The chapter ends with visions for the future of shared measures in ISE.

What is a Shared Measure?

For the purpose of this chapter, a shared measure is defined as an instrument developed to measure a particular outcome or construct that is common across a range of programs, projects, or the ISE field writ large. Programs have intended outcomes that evaluators seek to measure. The measurable parts of the outcomes are referred to as "constructs." Although this chapter uses the terms "outcome" and "construct" interchangeably, note that the outcome may actually be broader than the construct that is measured. This chapter purposefully uses the phrase "shared measure of a common construct or outcome" as opposed to "common measure." This decision is informed by a discussion among members of the Center for Advancement of Informal Science Education (CAISE) Evaluation & Measurement Task Force regarding potential confusion and resistance among some researchers and evaluators who equate "common measure" with standardized outcomes or testing (CAISE, 2017, February 14). The term "shared measures" focuses on the creation or use of rigorous measures that can be shared, or applied, across programs that are addressing the same construct or outcome. This definition assumes that developing the instrument included a process to: (a) examine the reliability of scores or the scoring procedure, and (b) collect validity evidence—evidence that the instrument measures the

outcome or construct as intended across programs sharing it. Although similar in meaning, the focus on "measurement of a common construct" instead of a "common measure" is deliberate. This is to stress that evaluators need to shift focus from the measure to the construct and thereby attend to what any given instrument is measuring, how it relates to a program's outcomes, and the related psychometric properties of the instrument (reliability and validity evidence).

Beyond being an instrument for a singular evaluation, a shared measure can be used across different evaluations. The Collaboration for Ongoing Visitor Experience Studies (COVES) project, one of the featured case studies in this chapter, is an example of the creation of a shared measure for a common outcome of interest for many museum evaluators—overall visitor experience. Many museums already gather visitor experience data through an exit or post-visit survey. The COVES project identified this common practice and pulled together a consortium of museums to develop a shared measure that museums could use to evaluate the common construct of "visitor experience" which they collaboratively defined. The project has a centralized database for the survey data, which allows for institution-level evaluation reports as well as field-wide reports comparing museums based on this common metric.

What Technical Qualities are Important for Shared Measures?

With the increasing development and use of shared measures across the ISE field comes the need for evaluators to better understand and assess an instrument's technical qualities, in particular reliability and validity. There are some common misconceptions around what validity means and, as a consequence, many people misuse the term. For instance, the phrase "validated instrument" is frequently used to describe an instrument that has gone through some sort of validation process. However, this phrase is inaccurate and can actually perpetuate the misuse of an instrument. Validity is not a feature of an instrument, but an argument related to the interpretations or conclusions that can be drawn from data gathered by an instrument (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). The validity argument is built over a multi-step validation process by gathering various types of construct validity evidence such as content, response process, and internal structure validity (AERA et al., 2014). Throughout the process of gathering different types of validity evidence, the construct of interest may be refined and items may be revised, added, or removed.

Content validity evidence relates to how well the construct of interest is represented in the content of an instrument (AERA et al., 2014). Such evidence can be gathered by reviewing the literature and gathering feedback from experts related to the construct being measured. Experts review how the construct was defined, identify what is missing from the definition, and help to ensure the content of the items or tasks in the measure adequately cover the construct area.

Response process validity evidence relates to the cognitive processes someone uses when responding to a question or completing a task and how well that aligns with the intended cognitive processes (Furr & Bacharach, 2014). Think-aloud interviews (also called cognitive interviews) are used to gather this kind of evidence. During a think-aloud interview, someone "thinks out loud" as they read and respond to an item or task. This helps bring to light what the person is reasoning about (hopefully, the construct) and any areas of misinterpretation, confusion, or cultural bias.

Internal structure validity evidence relates to statistical analyses that examine relationships between items in an instrument (Furr & Bacharach, 2014). Internal structure validity evidence can come from psychometric analyses such as factor analysis and differential item functioning (DIF) analysis (AERA et al., 2014). Factor analysis examines the correlations between items to confirm if an instrument is measuring one construct area (or factor) or multiple areas (or factors). DIF analysis examines the fairness of items for particular subgroups of an audience (often based on gender or race/ethnicity).

The types of construct validity evidence just discussed—content, response process, and internal structure—are the types most frequently collected during the development of ISE measures. Additional types of construct validity evidence outlined in the *Standards for Educational and Psychological Testing* are more frequently used in the validation of formal science education measures, but should be considered during the development and validation of measures for ISE: evidence related to relationships with other variables and evidence based on consequences of testing (AERA et al., 2014).

Various types of validity evidence examine how scores from an instrument relate to other measures of the same construct and measures of other constructs. Convergent validity evidence is the extent to which scores are correlated between an instrument and another measure of a theoretically similar construct. This involves an examination of how well scores on a measure predict performance on a particular criterion variable; this criterion-related validity evidence can be based on a variable measured at the same time as the instrument is administered (concurrent validity evidence) or measured in the future (predictive validity evidence) (AERA et al., 2014). Discriminant validity evidence is established when there is no relationship between scores on an instrument and a measure theorized to be of an unrelated construct (Furr & Bacharach, 2014).

Validity evidence based on consequences of testing relates to the positive and negative social consequences of testing. For example, a negative consequence of standardized testing in formal education is that teachers may spend less time on subjects that are not tested (AERA et al., 2014). Although evaluations in ISE are rarely high stakes like in formal education, negative consequences are still possible and should be considered. For example, it is valuable to consider whether a measurement procedure negatively interferes with someone's ISE experience.

Even though an instrument may have various types of construct validity evidence, the validity evidence is gathered from a particular audience in a particular context or setting. An instrument may have validity evidence for use with one audience or setting, but not with others. As stressed at the Palo Alto Convening for Assessment in Informal Settings, "validity arguments are context bound: one cannot assume an assessment with strong psychometric properties in one setting has similarly strong technical properties in a different setting" (Shields, Greenwald, Bell, Crowley, & Ellenbogen, 2014, p. 12). When reviewing a shared measure for possible use in an evaluation, evaluators need to ask themselves the following questions about the validity evidence available and compare the answers to their own evaluation needs, audiences, and settings (questions adapted from CAISE, 2017b).

- What construct(s) or outcome(s) does the instrument measure? How are those outcomes or constructs defined?
- What audiences was the measure created for? Was the measure tested with the variety of audiences it is meant to be used with (for example, age, gender, race/ethnicity, etc.)? Is there evidence that the measure was checked for gender and cultural biases when it was tested?

- What contexts or settings was the measure created for? Was it tested in all of these contexts and settings?
- What inferences are meant to be made based on the results of the measure? What kinds of validity evidence were gathered to support these inferences?

Gathering this information and reviewing it in relation to an evaluation's purpose, audiences, and settings will help an evaluator determine if the measure: (a) fits their needs and has sufficient audience/context-relevant validity evidence, (b) fits their needs but additional validity evidence is needed, or (c) is not a good fit for their evaluation needs. The EvalFest case study, discussed later in this chapter, provides an example of how a shared measure appeared to be a good fit for evaluating a construct (engagement), but additional validity evidence was needed for its use with new audiences and in a setting that differed from the original validation process.

Reliability is another important feature evaluators need to consider when selecting a measure. Reliability refers to "the accuracy or precision of a measurement procedure. Indices of reliability give an indication of the extent to which the scores produced by a particular measurement procedure are consistent and reproducible" (Thorndike & Thorndike-Christ, 2010, p. 118). For a survey, reliability relates to the consistency of the scores or responses and can be determined using different methods. Test-retest reliability is where a survey is administered to a group of people and then administered to them again, a short time later. The correlation between the scores shows how consistent responses are (Thorndike & Thorndike-Christ, 2010). Reliability based on parallel forms is when a similar survey is created that is measuring the same construct, and people take both surveys to see how well their responses correlate between the two measures. A third option is split-half reliability where one survey is divided into two halves that are considered equivalent in content and difficulty, people take the complete survey, and the scores on the two halves of the survey are correlated (Thorndike & Thorndike-Christ, 2010). Finally, the most frequently used method to gather reliability is the calculation of coefficient alpha, also called Cronbach's alpha, after people have taken the complete survey. Coefficient alpha is a measure of internal consistency reliability calculated by considering the relationship between every possible pair of items on the survey (Furr & Bacharach, 2014). Coefficient alpha values of .70 or above are considered acceptable (DeVellis, 2012). For observational protocols, reliability relates to the consistency of the ratings between two or more observers (called inter-rater reliability). Inter-rater reliability is indexed by the correlation between the scores of two or more raters. A related index, that of agreement, is often determined by looking at the percent agreement between raters or computing the statistic Cohen's kappa where a kappa value over .60 is considered good (Weathington, Cunningham, & Pittenger, 2010). Note that there can be perfect agreement—observers assign the same score to all people—but zero reliability because the assigned scores are constant and do not vary. Both indices are important and useful in evaluating the technical quality of an instrument.

What are Some Benefits of Shared Measures?

A major benefit of shared measures is that such measures support evaluators in conducting high quality evaluation. When a shared measure is already available for an outcome of interest, evaluators may not need to develop an instrument from scratch, saving valuable time and money. Moreover, using an instrument that has already been tested and has validity evidence for the outcome or construct of interest in similar contexts to the current evaluation can increase

evaluators' and clients' confidence in the quality of their evaluation data. Finally, shared measures provide a means of comparison with similar programs. Such comparison aids in interpretation of scores and program performance.

A side benefit of shared measures is the development of common definitions of constructs and outcomes across ISE experiences. Developing an instrument to measure a particular outcome and its related measurement construct requires the clear and detailed operationalization of both. This is done by looking to the literature, grounding the definition in how the outcome is defined in projects, and gathering feedback from content area experts to ensure the construct is appropriately defined given the outcome of interest and the items are adequately covering the construct area. All of these activities also provide content validity evidence for a measure. The Collaboration in the 21st Century (C2C) project, discussed in more detail in this chapter's case studies, is an example where extensive work was done at the beginning of the project to operationalize the construct of teamwork skills within the context of STEM out-of-school time (OST) programs as definitions in the literature were lacking in detail. The common definition of a construct, such as in the case of the C2C project, can be useful for evaluators to understand what they are measuring and for practitioners to better understand an outcome area they may be interested in addressing as part of their project.

As noted, the use of shared measures across the same kind of ISE experience can also allow for aggregation and comparison of data across project or institutional evaluations. Use of evaluation results are typically limited to the project being evaluated, which is the primary purpose of an evaluation. Using shared measures can help not only with interpretation of project performance but also make the case that evaluation results can have the secondary purpose of broadening the evidence base for ISE (Ellenbogen & Grack Nelson, 2012; Hussar, Schwartz, Boiselle, & Noam, 2008; Sacco, 2014). Aggregation of data can help the field better understand the impact of ISE experiences as well as identify ways to improve projects from not just one experience but multiple experiences. The COVES and EvalFest case studies, discussed later in this chapter, provide examples of data being aggregated across ISE institutions and experiences using shared measures.

On the practical side, the use of shared measures in evaluation helps build not only the evidence base of ISE but also justification as the field attempts to secure increased financial and political support (Ellenbogen & Grack Nelson, 2012; Noam & Shah, 2013). This is hard to do with evaluation data from one project; but, when multiple projects demonstrate impact on shared measures, it strengthens the case for supporting ISE experiences (Noam & Shah, 2013).

What Are Some Concerns Around the Use of Shared Measures?

A key concern related to shared measures is their misuse if evaluators lack knowledge in educational measurement. This is a valid concern given the current draft of the American Evaluation Association's Evaluator Competencies lacks any specific reference to an understanding of validity and reliability or the skills of judging or creating quality data collection instruments (American Evaluation Association, 2017). The closest competency is, "2.8 Determines appropriate methods, including quantitative, qualitative, and mixed methods" (American Evaluation Association, 2017, p. 1). Without an understanding of educational measurement, evaluators may have difficulty judging the quality of an instrument's validation process and psychometric properties, which, in-turn, hinders their ability to decide whether or not to use an instrument for their own evaluation needs. To address concerns around potential

misuse of shared measures, professional development in measurement is required to help evaluators use shared measures appropriately and, if needed, do additional validation work before they use a measure.

There are also concerns that evaluators may use a shared measure as the "go-to" instrument because it is available even if it may not be the best match for the evaluation. This concern was raised at the Palo Alto Convening on Assessment in Informal Settings, "Some of the assessments that are usable will go viral which is great, but there might be some unintended side effects of that, e.g. providing easy access to items could lead to making them *the* de facto evaluation" (Shields et al., 2014, p. 10). A shared measure may not always be the right tool to answer an evaluation question. There are still instances where evaluators will need to develop their own measures or use other measures, in addition to a shared measure, to be able to adequately answer their evaluation questions.

Evaluators sometimes hear concerns that the use of a particular measure may result in "teaching to the test." This is a legitimate concern if a measure drives what is implemented in a project, rather than the evaluator creating or choosing a measure that is well aligned with the outcomes the project defines as important (Brody, Bangert, & Dillon, 2007). However, "teaching to the test" should not be a source of concern if "the test" reflects the project's intended outcome(s). Before searching for a shared measure, evaluators need to have a clear conceptual understanding of the outcomes being addressed by a project. Only then should they look to see if an existing measure would meet their needs. If a shared measure is used that does not align well with a project's outcomes, there is a greater risk of the measure driving the project's activities since a project may feel pressure to "teach to the test" if that is how its success is measured.

Why the Heightened Interest in Shared Measures in the ISE Field?

The idea of shared measures is not new, however discussions and work around shared measures in the ISE field have been gaining traction since the late 2000s. In 2007, the authors of a paper commissioned by the National Research Council (NRC) recommended the development of an assessment system with evaluation tools for assessing science learning in informal settings, coupled with guidance for evaluators on how to use those tools (Brody et al., 2007). Two years later, the NRC's (2009) influential report, *Learning Science in Informal Environments: People, Places, and Pursuits*, recognized the lack of adequate measures for assessing outcomes of ISE experiences. The NRC outlined three criteria for developing new measures: (a) measures should address the range of ISE outcomes outlined in their six strands of informal science learning, namely science interest, knowledge, reasoning, practice, identity, and reflection on science as a way of knowing; (b) they should have validity evidence for their use; and (c) they need to align with the nature of informal learning experiences (NRC, 2009). Since then, there have been some significant advancements in the development and use of ISE shared measures, and field-wide conversations have included both the benefits and concerns around the use of shared measures as well as the field-wide needs these measures can fill. As illustrated in this section, these discussions have paved the way for more support, funding, and development of shared measures.

Of particular importance has been the development and use of shared measures in the afterschool sector. In 2008, the Noyce Foundation funded a study on the state of shared measures for evaluating afterschool STEM experiences. This work set out to answer questions about the nature of instruments used in the field, describe their psychometric properties, and assess the need for additional measures (Hussar et al., 2008). The study resulted in three recommendations

around shared measures to strengthen the evidence base for afterschool STEM: (a) development of an online database of measurement tools categorized by the National Science Foundation (NSF) impact categories (Friedman, 2008), (b) development of an item bank with questions that should be used across afterschool STEM evaluations to allow for program comparisons, and (c) creation of two shared measures (a survey and observation tool) for use in STEM afterschool programs (Hussar et al., 2008). The website Assessment Tools in Informal Science (ATIS) was developed as a result of the study's recommendations. ATIS (www.pearweb.org/atis) is a repository of measures for evaluating youth outcomes in OST settings. The site also includes information about the psychometric properties of the measures, when available. The study's recommendations also led to the development of two shared measures for the STEM afterschool field, a survey called The Common Instrument Suite and the Dimensions of Success observation tool, described in more detail later in this chapter (The PEAR Institute: Partnerships in Education and Resilience, 2017a).

In 2012, the NRC's Board on Science Education and Harvard University's Program in Education, Afterschool, and Resiliency organized the Summit on Assessment of Informal and Afterschool Science Learning. The summit focused on assessment within afterschool and summer science programs. Summit participants identified the "need to design assessments that can be used across sites, allow for aggregation of data, and capture quality across diverse program structures" (Noam & Shah, 2013, p. 36). They discussed arguments for shared measures, challenges in developing them, and considerations to keep in mind during their development. Summit participants also recommended the development of guidelines for people who are sharing their measures to ensure others can see how the construct was defined and how the instrument was developed to measure that construct (Noam & Shah, 2013).

Field-wide interest in the development and use of shared measures over the past decade are also reflected in solicitations of the NSF's Advancing Informal STEM Learning (AISL) program (previously called the Informal Science Education program). The 2009 NSF Informal Science Education solicitation was informed by the NRC's (2009) publication *Learning Science in Informal Environments*. The solicitation highlighted questions from the publication as possible areas of research for new projects, including the question: "What are appropriate cognitive and affective outcomes of informal science education initiatives (e.g., motivation, engagement, scientific identity) and can valid and reliable measurement tools be developed for those outcomes?" (NSF, 2009, p. 6). In 2010, the NSF solicitation's description of example Research Projects included a call for the development and validation of assessment tools with a particular interest in tools that "have potential utility for a group of related projects and activities in the informal science education field" (NSF, 2010, p. 5). The 2011 solicitation also highlighted instrument development in example Research Project proposals, however nothing was mentioned in the 2012 solicitation. In 2014, the solicitation specifically called out a number of funding priorities for practice-based research proposals, which included the priority area:

> Measurement of outcomes: Develops common instruments or ways to measure the outcomes of informal learning. We are interested in developing and researching new tools and frameworks to enable the sector to better understand the impacts they have on learners. (NSF, 2014, p. 6)

From 2013 to the present, the NSF AISL solicitation described example proposals that might be submitted under the Research in Service to Practice project area, including research projects that

"develop or adapt assessment instruments or scales" (NSF, 2013, 2014, 2015, 2017). The AISL program is a major funding source for ISE projects and, as a result, NSF's funding priorities have had an influence on evaluation and shared measures work happening across the ISE field.
In 2007, the NSF funded CAISE as part of their commitment to supporting the ISE field. CAISE works closely with the NSF AISL program to "build and advance the informal STEM education field by providing infrastructure, resources and connectivity for educators, researchers, evaluators, and other interested stakeholders" (CAISE, 2017a). In recent years, CAISE has provided a forum for conversations around shared measures for the field and a platform to advance shared measures work in current and future NSF-funded projects. At the 2013 CAISE convening, Building Evaluation Capacity in Informal STEM Education, evaluators identified shared measures as one of the three most urgent needs to improve evaluation within the ISE field (Ellenbogen, 2014). In continued efforts to strengthen the community of researchers and evaluators developing shared measures, CAISE and the Gordon and Betty Moore Foundation brought together six projects developing and validating measures for use in the ISE field (Shields et al., 2014). This follow-up convening allowed for discussion around the development and validation of measures in general, as well as issues specific to the ISE field. In 2014, CAISE pulled together a panel to continue the conversation around shared measures during a meeting of principal investigators of the NSF AISL program (Sacco, 2014). The panel shared their current work around shared measures in the ISE field and gathered information from the audience around their measurement needs, which were then shared with the field through a blog post (Sacco, 2014). In 2015, CAISE hosted an online forum on sharable measures. Discussion topics included concerns about sharable measures and what measures would be best for sharing results across projects (CAISE, 2015).

Internationally, there is also interest in the development and use of shared measures for studying ISE experiences. In 2011, the Wellcome Trust, a global charitable foundation, commissioned a study of informal science learning in the United Kingdom. One of the report's recommendations was to develop a forum of funders that would focus "on the development of a common set of indicators and measures that can be used in evaluating informal learning activities" (Lloyd, Neilson, King, & Dyball, 2012, p. 53). They felt funders could use these indicators and measures in a coordinated way to allow for a comparison of impacts across ISE experiences (Lloyd et al., 2012). In 2014, a number of funders and organizations came together to develop The National Forum for Public Engagement with STEM (National Co-ordinating Centre for Public Engagement, 2018a). One of the Forum's working groups is addressing challenges related to evaluation of public engagement in STEM by, "Developing a set of common questions that can be used collectively to improve the coherence of the evidence we collect" (National Co-ordinating Centre for Public Engagement, 2018b).

What Types of Shared Measures are Used in ISE Evaluations?

ISE evaluators use a variety of methods to gather evaluation data including interviews, observations, surveys, focus groups, artifact reviews, embedded assessment activities, and more. This chapter focuses on the most common types of shared measures currently available for ISE evaluations: observation tools and surveys. Observation tools are often used for measuring ISE outcomes such as engagement, skills, behaviors, and social aspects of learning (Friedman, 2008; NRC, 2009). Surveys have been used to measure a range of ISE outcomes including awareness, knowledge, understanding, skills, interest, attitudes, engagement, identity, motivation, self-

efficacy, and behavior (Friedman, 2008; NRC, 2009). While a shared measure may meet the needs of an evaluation, overdependence on any single measure is problematic; Chapter Two (Fu, Kannan, & Shavelson, this issue) discusses the need for multiple measures, particularly in the case of surveys.

Observation Tools

Observation tools have been developed to measure a variety of common outcomes of ISE experiences. Below are some examples of currently available tools.

The most widely used and oldest shared measure in the ISE field is "timing and tracking" of visitor interactions with exhibits and exhibitions. Timing and tracking has been used to study visitor behavior since the early 20th century, primarily to understand the paths individuals take through an exhibition, where they stop while moving through an exhibition, the amount of time someone spends at an individual exhibit and in an entire exhibition, and visitor interactions with exhibits (Yalowitz & Bronnenkant, 2009). In the late 1990s, Serrell (1998) aggregated findings across timing and tracking studies of 110 exhibitions. Out of this work came two common metrics for analyzing timing and tracking data related to where people stop in an exhibition and how much time they spend there: (a) The Sweep Rate Index—the square footage of an exhibition divided by the median time visitors spend in an exhibition, and (b) the Percentage of Diligent Visitors—the percentage of visitors who stop at more than half of the available exhibit elements within an exhibition (Serrell, 1998; Yalowitz & Bronnenkant, 2009). Using these metrics, evaluators can easily compare the use and success of a given exhibition to exhibitions of different sizes, audiences, and settings (Serrell, 1998).

A variety of observation tools have been developed to measure more specific aspects of the experiences people have with exhibits and exhibitions. The Adult Child Interaction Inventory (ACII) is an observation/interview tool that measures adult-child interactions in exhibitions where the role of an adult caregiver is considered key to a child's STEM learning (Beaumont, 2010). The freely-available instrument includes an observation protocol based on six research-defined adult interaction roles, as well as an interview instrument to record the adult's interpretations of their own actions (Boston Children's Museum, n.d.). The instrument was tested in children's museums and science museums for use with an adult-child pair, focused on children ages 3-5 and their adult caregiver. Content validity evidence was gathered for the instrument through reviews by caregivers, cultural experts, and museum staff (Beaumont, 2010). Reliability information was gathered during the testing of the observation tool through inter-rater agreement, where observers were able to establish at least 80% agreement before using the instrument (Beaumont, 2010).

The Museum Exhibit Skills Inventory (MESI) (Braswell, 2016) is a 14-item observation tool that documents children's display of "the 6Cs" (communication, collaboration, creativity, content knowledge, critical thinking, and confidence) while interacting with a hands-on exhibit. The researcher collected construct validity evidence by examining internal structure and correlations between the MESI subscales and an observation measure with similar construct areas (convergent validity evidence). Inter-rater agreement was acceptable, with Cohen's kappas above .60 for each item (Braswell, 2016). Reliability was also computed for each of the measure's three subscales. The coefficient alphas for the subscales ranged from .81 to .91 (Braswell, 2016).

Observation tools have also been developed to measure outcomes of specific types of ISE programs. The Dimensions of Success (DoS) instrument, developed by The Partnerships in Education and Resilience (PEAR) Institute (2017a), is an observation tool that measures the quality of STEM learning experiences in youth afterschool and summer programs. The measure is a rubric for rating 12 dimensions of quality that are categorized into four domains: features of the learning environment, activity engagement, STEM knowledge and practices, and youth development in STEM (Shah, Wylie, Gitomer, & Noam, 2018). Content validity evidence was gathered from expert reviews and feedback from observers as they pilot-tested the instrument to further clarify the constructs (Shah et al., 2018). In order to use and gain access to the DoS instrument, individuals must go through an extensive certification training process to become comfortable with rating the 12 dimensions and ensure they can achieve adequate reliability of observations (The PEAR Institute: Partnerships in Education and Resilience, 2017a). After participating in DoS trainings, inter-rater agreement for the 12 dimensions had Cohen's kappas ranging from .73. to .94 and percent agreement ranging from 95 to 100 (Shah et al., 2018).

Survey Tools

There are also a number of national projects that have developed survey tools to measure common outcomes of ISE experiences. Some of the largest shared survey projects currently underway in the ISE field are described here and in the case study examples.

The Developing, Validating, and Implementing Situated Evaluation Instruments (DEVISE) project developed a suite of scales to measure common outcomes of citizen science programs, which are programs where the public participates in authentic scientific research. DEVISE created and tested measures of the following constructs: interest in science, nature relatedness, self-efficacy for science, self-efficacy for environmental action, motivation for science, motivation for environmental action, skills of science inquiry, and environmental stewardship (Phillips, Porticella, Faulkner, & Bonney, 2018). The process of developing the tools began with a review of evaluation activities occurring in citizen science projects, identification of common outcomes across projects, and a review of existing instruments used to measure those outcomes (Shields et al., 2014). Construct validity evidence was gathered for these measures through expert review (content validity evidence), think-aloud interviews with citizen science participants (response process validity evidence), and confirmatory factor analysis (internal structure validity evidence) (Phillips et al., 2018). The reliabilities for the suite of DEVISE scales had coefficient alphas that ranged from .75 to .93 (Phillips et al., 2018).

The PEAR Institute (2017a) developed the Common Instrument Suite, a customizable 10-item survey, to measure how STEM afterschool and summer programs can impact a variety of STEM-related youth outcomes including interest, identity, career interest, career knowledge, enjoyment, and participation in STEM activities. The survey is meant for youth in grades 4 and above. There is a cost to use the instrument, which depends on how the survey is customized and the number of youth completing the survey (The PEAR Institute: Partnerships in Education and Resilience, 2017a). The data is collected in a centralized database so The PEAR Institute can provide programs with individualized reports as well as compare findings to other programs across the country. Although publications refer to the collection of validity evidence, the validation process and resulting validity evidence are not described. The instrument was found to have acceptable reliability when tested, with coefficient alphas above .85 (The PEAR Institute: Partnerships in Education and Resilience, 2017b).

The Learning Activation Lab developed and tested a wide range of scales to measure various dimensions of science and STEM learning activation in youth ages 10-14. In their terms, learning activation is "a state composed of disposition, practices, and knowledge that enables success in proximal science, technology, engineering, art, and mathematics learning experiences" (Learning Activation Lab, 2018a). Their shared measures for youth include scales to measure the four dimensions of science learning activation (science fascination, science values, science competency beliefs, and scientific sensemaking), scales to measure the four dimensions of STEM learning activation (STEM fascination, STEM values, STEM competency beliefs, and STEM innovation stance), a survey to measure emerging STEM activation in younger youth, a survey of engagement in science learning opportunities, and a survey of youth's preferences related to participating in science-related activities (Learning Activation Lab, 2018b). Construct validity evidence was gathered for these scales by developing them based on literature reviews and adapting other published scales (content validity evidence), think-aloud interviews with youth (response process validity evidence), and confirmatory factor analysis (internal structure validity evidence) (Learning Activation Lab, 2018b; Moore, Bathgate, Chung, & Cannady, 2011). Reliabilities for the scales were computed using coefficient alpha and ranged from .79 to .90 (Learning Activation Lab, 2018b).

Shared Measures Case Studies

Three projects featured here are developing shared survey measures for common outcomes of different types of ISE experiences. For each project, the case study describes the purpose of the project, the construct being measured and how it was decided upon, the development and validation process of the project's instrument(s), the benefits to the ISE field, and how the project relates to future directions in ISE evaluation. These projects are models of how shared measures can help push the ISE field forward. For example, all three cases use the meaningful involvement of evaluation stakeholders to ground construct definitions in the ISE experiences being evaluated, hence helping to increase the utility of the measure and the resulting evaluation results. The C2C project in particular contributed to the ISE field detailed definitions of a construct area tailored to STEM OST contexts, something that had been previously lacking (Grack Nelson, 2017). The COVES and EvalFest case studies are examples of community-created structures and systems to use, collect, and aggregate data across similar ISE institutions and experiences; these projects use shared measures in an effort to not only advance the evaluation practice of the collaborating institutions but also develop and test ways to produce cross-project findings to advance knowledge for the ISE field.

While these projects are models of shared measures in the ISE field, they also have limitations. All three projects focus on self-report surveys. Self-report surveys are beneficial because they tend to be easy to administer, are lower in cost than most other data collection methods, and allow an evaluator to collect a lot of data in a short period of time. However, as discussed by Fu et al. (this issue), self-report surveys also have limitations that need to be considered. People may not be the best judge of their own skills, knowledge, or behavior and may unknowingly respond inaccurately to a survey question. There are also concerns with people providing socially desirable responses. However, the fact that evaluations of ISE experiences are typically not viewed as "high-stakes" (e.g. people are not graded as they are in a formal education setting) may decrease incentives for people to exaggerate or fake their responses. To help address some of the concerns with self-report surveys, it is best to use surveys in

conjunction with other measures, such as observational data, to triangulate findings and provide a more holistic understanding of a project's impact. An additional limitation of the shared measures described in these case studies is that they all lack validity evidence in relation to other measures, which as mentioned earlier in this chapter is common for ISE measures but something the ISE field needs to consider in the development and validation of shared measures. For example, observations of how youth communicate in teams might be compared to survey data for the C2C project, observations of how people engage with a science festival booth activity might be compared with survey responses for the EvalFest project, or data from tracking visitors' behaviors in a museum might be compared to survey data for the COVES project.

Case Study 1: The Collaboration in the 21st Century (C2C) Project

The NSF-funded C2C project addresses the lack of evaluation tools for STEM OST programs to assess teamwork skills—a 21st century skill vital for preparing youth to enter the STEM workforce. The project developed and validated a shared measure to assess team communication skills in 6th – 12th grade STEM OST programs (Grack Nelson, 2017).

The instrument is a freely-available self-report survey that measures youths' perceived team communication skill level and their comfort, ease, and likelihood of using the skill. An imaginary teamwork scenario provides the framing for responding to the survey's 28 items. The instrument was developed using a four-phase development and validation process to ensure the survey gathered reliable data and had adequate validity evidence for use with the diverse youth population that participate in a wide range of STEM OST programs.

Phase 1 focused on identifying and then operationalizing the teamwork skill area most important to STEM OST programs. This process was informed by an extensive literature review and in-depth interviews with 34 STEM OST program providers across the country. The skill area of "team communication" was most commonly addressed across the interviewed programs (85% of programs), thus becoming the focal construct of the shared measure. The definition of team communication skills was then operationalized by going back to the interviews and literature. The resulting definition included three construct areas: (a) closed-loop communication, (b) information exchange, and (c) listening.

Phase 2 involved gathering content validity evidence. The three construct areas and items were reviewed by 11 STEM OST program providers as well as three experts from the fields of youth development, teamwork science, and measurement. The feedback from providers was important to help ensure the survey's scenario and items were relevant to the way teams were used in STEM OST programs and the team communication skills addressed in these programs. The advisors' feedback was important for gathering evidence related to the construct areas of interest and the content (scenario and items) being used to measure that construct. After each set of reviews, there was refinement of the construct and instrument.

Phase 3 included think-aloud interviews and a pilot test of the survey with youth and programs across the country. During Phase 3, the instrument was revised through an iterative process of conducting think-aloud interviews, making revisions, and carrying out more interviews. A total of 30 youth from 11 STEM OST programs participated in the interviews. The interviews also helped to identify potential areas where verbal ability may be an issue and the survey text was revised as needed. The purpose of the pilot test was to begin to look at the reliability of the responses, conduct exploratory factor analysis to identify the model structure to test through confirmatory factor analysis in Phase 4, and identify items to remove to decrease the

length of the survey. A total of 378 youth from 19 STEM OST programs participated in the pilot test.

Phase 4 was a national field test of the instrument with 959 youth from 40 STEM OST programs across the country. Through confirmatory factor analysis, a five-factor model of team communication skills was found to be a good fit (internal structure validity evidence), two factors for closed-loop communication, two factors for information exchange, and a listening factor. Responses for each of the five factors were reliable with coefficient alphas ranging from .70 to .79. Additional internal structure validity evidence was collected using DIF analysis to look at item fairness for different groups of youth. Comparisons were made between girls and boys, white youth and African-American youth, and white youth and Asian-American youth. None of the items were found to have DIF across the three groups of comparisons.

This project resulted in an instrument that evaluators can be confident will gather reliable data, has adequate validity evidence for use with STEM OST programs, is grounded in what actually occurs in these programs, and provides data that a wide range of programs will find useful. However, this instrument is only one measure of the range of teamwork skills STEM OST providers cover in their programs and is not a direct measure of skill, but a self-report of youths' perception of their skills. Future work is needed to develop an observational measure to assess these skills more directly and triangulate survey findings. There is also the need for ISE evaluators and researchers to develop measures relevant to other teamwork skills, as well as other 21st century skills addressed in ISE experiences. Future research can address this need, helping to increase the capacity of evaluators to more rigorously define and measure important outcomes related to 21st century skills development in STEM OST programs.

Case Study 2: The EvalFest Project

EvalFest (www.evalfest.org) is a NSF-funded community-created multisite evaluation project that was designed to improve and learn from common evaluation methods used across science festivals in the United States. Science festivals are science and technology celebrations that take place over multiple days or weeks within a community, region, or entire state. A main component of science festivals is the opportunity to engage the public with science through interactions with scientists, engineers, and STEM organizations, and a wide variety of activities such as exposition style events, lectures, lab tours, and more. At the beginning of the project, all U.S.-based science festivals were invited to participate in EvalFest; approximately 75% joined (25 festivals), representing a large portion of the festival community at that time. Since then, the festival community has grown and the project now includes 45% of U.S.-based festivals.

The EvalFest project was developed in response to recommendations from Hussar et al. (2008) for more systemic assessment in the ISE field and the development of shared tools to use across program evaluations in different sectors of ISE. The project applied these recommendations by identifying shared items that would be used to evaluate all partner science festivals, testing new and potential shared measures that could be adopted as needed, and ensuring that shared measures also resulted in shared data for and across the EvalFest community.

Shared metrics for the project are identified by either the project's leadership team or through mini-grants that partners lead to serve as incubator sites for testing new methods themselves. The leadership team has tested three shared metrics to date—a set of core questions for all science festivals to include in their attendee surveys, the Learning Activation Lab's

Engagement Survey for use with science festival audiences (see description below), and a festival follow-up survey. The leadership team's process for identifying shared metrics included the following steps: identify a common construct or data collection method based on the interests of EvalFest partners, develop the instrument by building on existing items or scales when available, gather response process validity evidence to ensure the items function as expected in the context of science festival activities and with science festival audiences, and when relevant gather internal structure validity evidence.

For example, in an attempt to move beyond simple satisfaction ratings and toward evaluation of a construct associated with the broader literature on science learning, EvalFest leaders identified off-the-shelf shared measures that had the potential to provide meaningful evaluation data to festival directors. Each was presented to the EvalFest community and the construct of engagement, as measured by the Learning Activation Lab's Engagement Survey, was selected. The Learning Activation Lab defined the construct of engagement as, "one's focus, participation, and persistence on a task" (Chung, Cannady, Schunn, Dorph, & Bathgate, 2016, p. 1). The Engagement Survey was designed for and tested with middle school students who had just completed a science activity in a formal or informal learning environment. The Learning Activation Lab gathered a wide variety of construct validity evidence for the Engagement Survey including content, response process, and internal structure validity evidence (Chung et al., 2016).

Initially, it was unclear whether or how the Engagement Survey would work with the broader range of children and adults who attend science festivals and specifically for the free-choice kind of activities at exposition style events (called expos). For this reason, additional construct validity evidence was needed. Response process validity evidence was gathered through think-aloud interviews during expos (after people participated in individual booths as well as completed the entire expo experience), with youth (ages 10 and above) and adults. The think-alouds gathered data on the items included on the Engagement Survey, as well as the survey's rating options. The results from this study indicated that the items and the rating options were interpreted correctly by children and adults when they were asked to focus their responses on a specific festival booth. However, some of the survey items did not function as expected when attendees were asked to reflect on their expo experience as a whole. This was not surprising since the Engagement Survey was originally developed to be administered immediately after engagement with an individual science activity. Thus, response process validity evidence supported use of the Engagement Survey with adults and children at booth type experiences, but not for use with an entire expo experience. A subset of partners then used the Engagement Survey to collect data with adult and child audiences interacting with festival booths at expos. These data were used to gather internal structure validity evidence and calculate reliability. Internal structure validity evidence came from confirmatory factor analysis that confirmed a one factor structure for the scale and the reliabilities of scores when using the survey with hands-on booths and demonstration booths was coefficient alpha .76 and .71, respectively.

The EvalFest project was designed to create not just shared metrics but also shared data for the community. Data collection is centralized through an online survey platform, and data are stored in a shared database. This allows for both individualized evaluation reporting as well as comparisons across festivals in the EvalFest community. Cross-festival comparisons provide the ISE field with additional understanding about science festivals as an informal learning mechanism. The aggregated data are also used to share trends and compare and contrast results with the broader international science festival community.

Multisite evaluations, such as EvalFest, that are created for a particular community, by the community, and led by members of the community hold tremendous potential for the future development and support of shared measures. This community-based shared measure process can strengthen the quality and rigor of evaluation practice, increase buy-in to use the shared measures, and support the growth of sectors within ISE and beyond. The EvalFest project is a model for other future shared measure projects, especially in relation to the procedures used to share data across partners and how the use of shared measures across multiple similar ISE experiences can lead to new field-wide understandings.

Case Study 3: The Collaboration for Ongoing Visitor Experience Studies (COVES) Project

The COVES project (www.understandingvisitors.org) is a data initiative through which science museums across the country systematically collect feedback on visitor experiences. Isolated visitor study initiatives were common when COVES was conceptualized; in fact, several museums were conducting seasonal and ongoing visitor studies with similar goals of understanding visitor experiences in their space. However, when trying to compare data across sites, subtle differences in the phrasing of questions and larger differences in methodology prevented coming to any shared understandings across the field. To address this impasse, funding was sought from the Institute of Museum and Library Services (IMLS) (Museum of Science's Research and Evaluation Department, 2014). The initial funding has grown and resulted in a standardized instrument, systematic structure for ongoing data collection, and sustainable model of continuation beyond grant funding.

From the onset, the COVES team valued shared ownership of all participating institutions in driving a deeper understanding of their visiting audiences. The shared measure needed to support an individual institution's own questions and provide insight into necessary improvements, as well as inform field-wide learning. Thus, the COVES team approached the development of the shared measure as being *by* museum professionals *for* museum professionals. The COVES team collaboratively defined the common construct the shared measure was meant to address: visitor experience. The team defined "visitor experience" based on four overarching areas: who visits, why they visit, what they see and do while they visit, and how they feel about their experiences. To create the shared measure, the COVES team collected surveys developed and employed by various institutions conducting their own studies, identified areas of shared interest, and discussed appropriate phrasing when question content was similar but wording varied. The survey was developed so the questions aligned with the four areas of the visitor experience construct. The resulting survey is meant to be administered to adult visitors at the end of their museum experience before exiting the building.

To establish content validity evidence, three separate groups of experts were invited to review the instrument and provide feedback. The first set of reviewers were individuals from the institutions who would be using the instrument through their participation in COVES. The second set of reviewers was composed of three measurement experts. The third set of reviewers were Visitor Studies Association professionals who contributed to instrument review at the Association's annual conference during a roundtable workshop and poster session.

To gather response process validity evidence, the COVES team conducted think-aloud interviews for individual questions and monitored survey responses to identify concerns. When adding new questions to the instrument, the team gathers additional response process validity evidence through additional think-alouds to test the specific question for phrasing and format.

While a key part of the COVES project is the development of a shared measure, the truly innovative aspects of the project are the infrastructure developed to support shared use and meaning making as well as the flexibility built into the instrument. Structurally, the COVES survey and associated data live in the same online survey software platform. Individual institutions collect their own data that feed into this centralized database. The data are then analyzed by a centralized COVES team. By consolidating the data, the ideal of using shared measures to inform field-wide learning is possible. The embedded flexibility of the survey to include institution-specific questions beyond the core COVES survey questions ensures that the resulting data also meet the needs of each participating institution. Examples of this include adding marketing or awareness questions specific to an institution or customizing the list of possible experiences for each museum to include temporary exhibitions based on run-date. Although these examples highlight instances in which comparison across sites is limited by modifying questions, they increase the utility of the data for each participating institution and still allow for comparison across the bulk of the survey.

The COVES project has transitioned from a grant-funded collaboration to a sustained field-wide initiative through which participating institutions contribute financially to support ongoing work. This significant endeavor pushes the future direction of evaluation in the ISE field for several reasons. First and foremost, this helps solidify the notion that understanding audiences through data is a necessary step in informing decision-making, and hopefully encourages additional use of evaluation as capacity expands throughout the ISE field. Second, with this ongoing support, the COVES team has developed a dashboard tool for sharing data in an ongoing fashion, rather than relying on longer, often costlier reporting mechanisms, helping to promote data use. Third, the COVES team will begin publishing an annual field-wide report, which can both contextualize other museums' understandings of visitor data, as well as highlight emerging trends in visitation across the field. As the project continues to grow, the COVES team envisions a larger and more thoughtful collaboration of museums using data to understand visitors and their experiences, and to support evaluation capacity building, particularly at smaller ISE institutions.

What are Future Directions for Shared Measures?

Future developments rely on the capacity of the field to build shared measures that are innovative, reliable, and have adequate validity evidence. The future in measurement demands that attention be paid to building evaluation capacity. Thus, this section outlines a multi-faceted approach to enhancing measurement capacity among evaluators; a broader perspective on evaluation capacity building that focuses on non-evaluators is discussed in Chapter Five (Bequette, Cardiel, Cohn, Kollmann, & Lawrenz, this issue). This chapter's proposed approach features the creation of an ISE measurement resource center, coupled with measurement-related professional development for evaluators and strategic funding and structures to support coordinated cross-project data gathering and analyses. These advances will both support and drive new kinds of collaborative work among ISE professionals and evaluators. They will also require a change in mindset of sharing measures to help ensure they are easily accessible for use by all ISE evaluators to help push the field forward.

Enhanced Measurement Capacity

The ISE field needs an online resource center where evaluators can find measures appropriate for evaluating ISE experiences for a diversity of audiences. A few online repositories currently exist, but they do not cover the full range of ISE experiences and audiences; and, at times, the information provided with measures can be extremely limited.

The ATIS website (http://www.pearweb.org/atis) is focused on one sector of the ISE field, namely youth afterschool programs. For each featured instrument, the site includes information about the general domain being measured (but not the specific construct), the audience the measure was tested with, and some reliability and validity evidence. However, the site lacks consistency in how reliability and validity information is presented and oftentimes this information is extremely limited, not clearly described, or missing. For example, validity information is often described as "present and acceptable" or "established" without a description of the validity evidence so that potential users can judge for themselves. There are even a few instances where reliability is described under validity and vice versa, which could add to an evaluator's misunderstanding of how to judge an instrument's properties. There are some instances where the site links to an article with more detailed information about the instrument but the article may not be easily accessible to ISE evaluators that lack online journal access.

The STEM Learning and Resource Center's website (http://stelar.edc.org/resources/instruments) has a database of instruments that have been used by NSF-funded Innovative Technology Experiences for Students and Teachers (ITEST) projects, which includes ISE projects for youth; but, again, it represents only a limited part of the ISE field. The site shares measures and links to articles that may describe reliability and validity information but does not pull out and provide this information in an easy way for evaluators; and, like the ATIS site, articles may not be easily accessible to evaluators.

The ISE field will greatly benefit from a "one-stop shop" for evaluators' measurement needs, similar to CAISE's catalog of ISE evaluation reports on the website www.informalscience.org. An important part of an online resource would be not only instruments and detailed instructions on how to use them, but also information about the construct being measured, the instrument's validation process, and resulting validity evidence. Evaluators need enough information to judge an instrument's quality and appropriateness for their evaluation needs. The resource center could also consider providing means and incentives for users to add additional validity evidence gathered when using a measure with new audiences and in new contexts. This would be a complex and costly endeavor to coordinate and ensure quality of submissions of additional validity evidence but would be a valuable way to continue to build the utility of shared measures for the ISE field. Additionally, an online resource center should provide detailed guidance on what to consider when choosing an instrument, how to judge the validation process and psychometric properties of an instrument, and how to gather additional validity evidence.

The online resource center must be supported by targeted professional development. As the field moves toward the use of more shared measures, measurement-related professional development is needed to build evaluators' capacity to be able to critically examine the validity evidence available for a measure and its related psychometric properties, appropriately use these measures, and gather additional validity evidence as needed. This will require professional development experiences where evaluators not only learn about measurement; but have real-world opportunities to practice the skills of reviewing instruments, judging their quality, and

using them for their own evaluations. Without professional development, the ISE and evaluation fields are doing a disservice to evaluators and not adequately addressing the concerns people have around the misuse of shared measures.

One of the main benefits of shared measures is the ability to aggregate data across evaluations using the same measure to gain a deeper understanding of the outcomes or impacts of ISE experiences. This kind of aggregation requires formal structures with consistent, dedicated, and customized resources and staffing, going beyond the online resource center and professional development already described. Incentives and other means to encourage buy-in are also necessary to ensure evaluators contribute to a larger dataset, especially since evaluation is often focused and funded on a project-by-project basis (Ellenbogen & Grack Nelson, 2012). Some projects, like EvalFest and COVES, have developed systems to gather data across common projects as part of their shared measures work, which can serve as models for the ISE field. These projects have created communities around using shared measures, provided support for the aggregation of data, and shared aggregate results back to the evaluators and program staff who shared their data. Lessons learned from these projects will move the ISE field in new directions around the use of shared measures to provide cross-project evidence of program impacts.

New Kinds of Collaborative Work among Evaluators

Advancing the development of shared measures in ISE will require more, and possibly new kinds of, collaboration among ISE evaluators in the future. Evaluators typically work in isolation, focused on what it is they are evaluating. Creating shared measures requires work beyond the individual evaluation, and it demands various types of expertise that may not reside with a single evaluator. For this reason, evaluators, measurement experts, educational researchers, and ISE experts and practitioners need to work together to define constructs, develop items and protocols, and test measures (Noam & Shah, 2013). This collaborative work is starting to happen in the ISE field and, as discussed at the Palo Alto Convening on Assessment in Informal Settings, needs to continue into the future:

> The community of people doing assessment in informal science education is coming together. As an interdisciplinary community, we should leverage diverse expertise and figure out who has what skills... We need to learn how to work together, share knowledge, and help each other tinker with and further develop our items, instruments, and study designs. (Shields et al., 2014, p. 12)

Interdisciplinary collaboration will strengthen the measures that are created. “Strong instruments will not emerge from a process where we work in silos” (Crowley, 2014).

Increasing Accessibility of Shared Measures

Even though there are a number of projects producing shared measures for the ISE field, there are barriers to accessing some measures. Some evaluators consider their instruments proprietary and are reluctant to share with others. Others publish instruments and psychometric information in peer-reviewed journals that are often inaccessible to ISE evaluators. Additionally, some projects charge a fee to access an instrument, which means evaluators may choose not to

use an instrument if they cannot see it before paying or the price is prohibitive. Fees may be needed to develop structures related to the use of a measure, such as a shared online platform, database, and reporting features, similar to the COVES project described in this chapter. However, the future of shared measures will require changing mindsets around access to individual tools if the ISE field wants to further build the capacity of evaluators and enhance the rigor of ISE evaluations.

Concluding Thoughts

Shared measures are making headway in ISE evaluation and will only be a more integral part of ISE evaluators' practice in the future. As illustrated in this chapter, there have been a wide range of conversations and calls to action in the ISE field around shared measures. As a result, the number of ISE shared measure projects have been increasing and will continue to do so. However, having more measures available is not enough. The ISE field needs to push even further forward. Evaluators need professional development so they can critically review instruments for their quality and appropriateness for the audience, setting, and outcome being measured. Instrument developers need to ensure that they are providing sufficient and easily understandable information with their instruments so that evaluators can judge an instrument's psychometric properties and how well the instrument meets their evaluation needs. Structures and systems will need to be developed to aggregate data from shared measures and use those findings to generate new knowledge about ISE experiences and outcomes. Building evaluators' capacity to critically examine and use shared measures, as well as potentially change their practice to think about how shared measures can inform larger field-wide questions beyond a singular evaluation, may require a change in thinking about the necessary competencies of evaluators and their role in advancing field-wide knowledge about ISE experiences. Such a change will not only advance ISE evaluators' practice, but also provide unique opportunities for field-wide learning that may not otherwise be possible without the use of shared measures.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Evaluation Association. (2017). *AEA Evaluator Competencies 8/28/17 final*. Washington, DC: Author.

Beaumont, L. (2010). *Developing the Adult Child Interaction Inventory: A methodological study*. Boston, MA: Evergreene Research and Evaluation.

Boston Children's Museum. (n.d.). *The Adult Child Interaction Inventory (ACII) and resource DVD*. Boston, MA: Author.

Braswell, G. S. (2016). Creation and validation of an observational tool to assess children's domain-general skills at museum exhibits. *Visitor Studies, 19*(2), 211-224.

Brody, M., Bangert, A., & Dillon, J. (2007). *Assessing learning in informal science contexts*. Commissioned paper for the National Research Council, Science Learning In Informal Environments Committee.

Center for Advancement of Informal Science Education. (2015). *Summary of sharable measures forum*. Washington, DC: Author.

Center for Advancement of Informal Science Education. (2017a). *About CAISE*. Retrieved from http://www.informalscience.org/about-caise

Center for Advancement of Informal Science Education. (2017b). *Evaluation tools and instruments*. Retrieved from http://informalscience.org/evaluation/evaluation-tools-instruments

Center for Advancement of Informal Science Education. (2017, February 14). *Meeting of the CAISE Evaluation and Measurement Task Force*.

Chung, J., Cannady, M. A., Schunn, C., Dorph, R., & Bathgate, M. (2016). *Measures technical brief: Engagement in science learning activities*. Retrieved from http://activationlab.org/wp-content/uploads/2018/03/Engagement-Report-3.2-20160803.pdf

Crowley, K. (2014, January 20). Updates from the field: Meeting on assessment in informal science education. [Blog post]. Retrieved from http://www.informalscience.org/news-views/updates-field-meeting-assessment-informal-science-education

DeVellis, R. F. (2012). *Scale development: Theory and applications*. Thousand Oaks, CA: Sage.

Ellenbogen, K. (2014). *Summary of the CAISE convening on building capacity for evaluation in informal science, technology, engineering and math (STEM) education*. Washington, DC: Center for Advancement of Informal Science Education. Retrieved from http://www.informalscience.org/sites/default/files/ECB_Convening_Summary_10-10-14.pdf

Ellenbogen, K., & Grack Nelson, A. (2012). *Evaluation under pressure: Balancing the needs of the ISE field with the needs of individual projects*. Retrieved from http://sites.nationalacademies.org/DBASSE/BOSE/DBASSE_080110

Friedman, A. J. (Ed.). (2008). *Framework for evaluating impacts of informal science eduation projects: Report from a National Science Foundation workshop*. Washington, DC: The National Science Foundation.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Grack Nelson, A. (2017). *Development and validation of a survey to measure perceived team communication skills in middle and high school STEM out-of-school time programs*. (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Hussar, K., Schwartz, S., Boiselle, E., & Noam, G. (2008). *Toward a systematic evidence-base for science in out-of-school time: The role of assessment*. Cambridge, MA: Program in Education, Afterschool & Resiliency.

Learning Activation Lab. (2018a). *Activation*. Retrieved from http://activationlab.org/activation/

Learning Activation Lab. (2018b). *Tools: Measures and data collection instruments*. Retrieved from http://activationlab.org/tools/

Lloyd, R., Neilson, R., King, S., & Dyball, M. (2012). *Review of informal science learning*. London, England: Wellcome Trust.

Moore, D. W., Bathgate, M. E., Chung, J., & Cannady, M. A. (2011). *Technical report: Measuring activation and engagement. Activation Lab, Enables Success Study*. Pittsburgh, PA & Berkeley, CA: Learning Activation Lab.

Museum of Science's Research and Evaluation Department. (2014). *Creating a Collaboration for Ongoing Visitor Experience Studies (C-COVES)*. Boston, MA: Museum of Science.

National Co-ordinating Centre for Public Engagement. (2018a). *About the National Forum*. Retrieved from https://www.publicengagement.ac.uk/nccpe-projects-and-services/nccpe-projects/national-forum-public-engagement-stem

National Co-ordinating Centre for Public Engagement. (2018b). *Evaluating public engagement with STEM*. Retrieved from https://www.publicengagement.ac.uk/nccpe-projects-and-services/nccpe-projects/national-forum-public-engagement-stem/evaluating-public-engagement-with-stem

National Research Council. (2009). *Learning science in informal environments: People, places and pursuits*. Washington, DC: The National Academy Press.

National Science Foundation. (2009). *Informal science education (ISE) program solicitation (NSF 09-553)*. Washington, DC: Author.

National Science Foundation. (2010). *Informal science education (ISE) program solicitation (NSF 10-565)*. Washington, DC: Author.

National Science Foundation. (2013). *Advancing informal STEM learning (AISL) program solicitation (NSF 13-608)*. Washington, DC: Author.

National Science Foundation. (2014). *Advancing informal STEM learning (AISL) program solicitation (NSF 14-555)*. Washington, DC: Author.

National Science Foundation. (2015). *Advancing informal STEM learning (AISL) program solicitation (NSF 15-593)*. Washington, DC: Author.

National Science Foundation. (2017). *Advancing informal STEM learning (AISL) program solicitation (NSF 17-573)*. Washington, DC: Author.

Noam, G., & Shah, A. M. (2013). *Game-changers and the assessment predicament in afterschool science*. Cambridge, MA: Program in Education, Afterschool, and Resiliency.

Phillips, T., Porticella, N., Faulkner, H., & Bonney, R. (2018). *Common measures for individual learning outcomes: Technical brief series*. Ithaca, NY: Cornell Lab of Ornithology.

Sacco, K. (2014, October 7). Measuring learning across ISE projects. [Blog post]. Retrieved from http://www.informalscience.org/news-views/measuring-learning-across-ise-projects

Serrell, B. (1998). *Paying attention: Visitors and museum exhibitions*. Washington, DC: American Association of Museums.

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.

Shields, S., Greenwald, E., Bell, J., Crowley, K., & Ellenbogen, K. (2014). *The Palo Alto convening on assessment in informal settings: Synthesis report*. Washington, DC: Center for Advancement of Informal Science Education.

The PEAR Institute: Partnerships in Education and Resilience. (2017a). *A guide to PEAR's STEM tools: Dimensions of Success & Common Instrument Suite*. Cambridge, MA: Harvard University.

The PEAR Institute: Partnerships in Education and Resilience. (2017b). *Common Instrument Suite*. Retrieved from https://www.thepearinstitute.org/common-instrument-suite

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson Education, Inc.

Weathington, B. L., Cunningham, C. J. L., & Pittenger, D. J. (2010). *Research methods for the behavioral and social sciences*. Hoboken, NJ: John Wiley & Sons, Inc.

Yalowitz, S. S., & Bronnenkant, K. (2009). Timing and tracking: Unlocking visitor behavior. *Visitor Studies, 12*(1), 47-64.

AMY GRACK NELSON, Ph.D., is Evaluation and Research Manager at the Science Museum of Minnesota.
MEGAN GOEKE is an Evaluation and Research Associate at the Science Museum of Minnesota.
RYAN AUSTER is a Senior Research and Evaluation Associate at the Museum of Science, Boston.
KAREN PETERMAN, Ph.D., is the founder of Karen Peterman Consulting, Co., which specializes in the evaluation of STEM education projects and research on evaluation methods for informal learning environments.
ALEXANDER LUSSENHOP is a Research and Evaluation Associate at the Museum of Science, Boston.