Fu, A. C., Kannan, A., & Shavelson, R. J. (2019). Direct and unobtrusive measures of informal STEM education outcomes. In A. C. Fu, A. Kannan, & R. J. Shavelson (Eds.), *Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation*, *161*, 35–57.

2

Direct and Unobtrusive Measures of Informal STEM Education Outcomes

Alice C. Fu, Archana Kannan, Richard J. Shavelson

Abstract

The free-choice nature of informal STEM education (ISE) makes rigorous and contextually appropriate evaluation of outcomes challenging. Traditional measures such as surveys and interviews have been widely used in ISE evaluations, but they have limitations: They are typically self-reports that are susceptible to the reactive effects of measurement, and they tend to intrude upon the participant's learning experience. The ISE field needs measures that capture outcomes in more direct and less obtrusive ways, permitting triangulation with multiple measures on outcomes. In this chapter, we define what we mean by direct and unobtrusive measures, and we discuss the feasibility and future of using such measures in ISE evaluations by drawing on examples from the field. We include a case study in which we adapt a school-based performance assessment to embed into the informal learning experiences of participants at a STEM tinkering workshop; and we highlight successes, challenges, and implications for the future.

A central challenge confronting evaluation in informal science, technology, engineering, and mathematics (STEM) education (ISE) is developing outcome measurement methods that are both rigorous and appropriate for contexts where participants typically expect to set their own goals, exercise autonomy over their learning activities, and enjoy a

non-threatening experience (Allen & Peterman, this issue; Fu, Kannan, & Shavelson, Editors' Notes, this issue). We center this chapter on innovations that address two aspects of this challenge, namely measures that (a) *directly* capture data on learning outcomes and (b) are *unobtrusively* embedded in the learning experience.[1] We begin by discussing why these two aspects are worth our attention.

What Is Being Measured?

Any discussion of measures must attend to what is being measured. Learning outcomes in ISE are multiple, including but not limited to STEM-related interest, knowledge, skills, reasoning, engagement, attitudes, identity, and behaviors (Friedman, 2008; National Research Council [NRC], 2009). Many ISE projects typically address more than one of these; consequently, attending to what is being measured is essential. Moreover, outcomes multiply when one considers not only individual-level outcomes but also outcomes at other levels such as the group, project, and beyond (Lemke, Lecusay, Cole, & Michalchik, 2015; NRC, 2015). Finally, these outcomes are often expected to change over time, so multiple measures of the same outcome might be needed.

When we build measures of outcomes, we speak of the constructs to be measured—certain knowledge, skills, interests, attitudes, and the like. We use the terms "outcome" and "construct" interchangeably, although the outcome may be broader than the construct we end up measuring. Any single measure merely samples from a larger domain, and evaluators need to select or develop measures that adequately represent that domain. Imagine, for example, that "demonstration of critical thinking" is a desired outcome. The "critical thinking" domain comprises many aspects, and an assessment of critical thinking will include some aspects but exclude others. Precise definitions of desired outcome versus measured construct make alignments (and misalignments) apparent, thereby clarifying what is valid when interpreting scores or results.

Evaluators also need to choose study designs that answer the evaluation questions, which might involve tracking changes over time or justifiably attributing outcomes to the program being evaluated. Study design is especially important if a study seeks to assess STEM learning, wherein "learning" implies change over time; in such cases, it is important to account for participants' prior knowledge and experience. This chapter focuses on measures, not designs; that is, we focus on innovations in collecting data about informal learning constructs. Other sources have more

[1] This chapter builds from our presentation at the 2016 Visitor Studies Association conference (Fu & Kannan, 2016). This work was supported, in part, by a grant from the Gordon and Betty Moore Foundation to SK Partners, a research and consulting group with which all of the co-authors were affiliated.

extensive discussions about the relative strengths of various study designs (see, e.g., Friedman, 2008; Fu, Kannan, Shavelson, Peterson, & Kurpius, 2016; NRC, 2002; Rossi, Lipsey, & Freeman, 2004; Shadish, Cook, & Campbell, 2002).

Directness of Measures

Interviews and surveys are among the most prevalent methods for measuring outcomes in ISE evaluations. These methods enable respondents to communicate their perspectives on their own and others' learning, attitudes, engagement, behaviors, and more. In a review of thirty-six summative evaluation reports posted on www.informalscience.org, a key online resource of the informal STEM community, all but one of the reports included surveys and/or interviews (Fu et al., 2016). Perhaps more striking, 30% of these reports employed *only* surveys, interviews, or both. Interviews and surveys provide evaluators with an invaluable viewpoint, a window into the thoughts and perspectives of program participants and other respondents. However, self-report measures have limitations. For example, they are vulnerable to reactive measurement effects such as participants knowingly or unknowingly trying to please the evaluator, and over- or underestimating what they know or do. And the problem of relying too much on them was clearly stated over 50 years ago:

> Today, the dominant mass of social science research is based upon interviews and questionnaires. We lament this overdependence upon a single, fallible method. Interviews and questionnaires intrude as a foreign element into the social setting they would describe, they create as well as measure attitudes, they elicit atypical roles and responses, they are limited to those who are accessible and will cooperate, and the responses obtained are produced in part by dimensions of individual differences irrelevant to the topic at hand.
>
> *But the principal objection is that they are used alone.* [emphasis original] No research method is without bias. (Webb, Campbell, Schwartz, & Sechrest, 1965/2000, pp. 1–2)

Using multiple measurement methods achieves a more complete understanding of ISE outcomes than any single approach. To be sure, some measures of outcomes such as personal experience, identity, or self-concept, by their very definition, require self-report. Other outcomes such as knowledge and interest can be measured by approaches that are more direct than self-report, including observation, choice behavior, and hands-on tasks requiring science reasoning. Moreover, new and more direct assessment tools may become available in future as innovations in measurement unfold, as in the case of "perspective-taking" (Shavelson, Zlatkin-Troitschanskaia, & Marino, 2018).

Wary of false dichotomies, rather than speak of direct versus indirect measures, we conceptualize measurement-method "directness" as a continuum from less direct (e.g., self-report of recycling behavior) to more direct (e.g., examination of representative sample of neighborhood recycling bins). The more direct a measure is, the more closely it approximates the "criterion situation" of interest (e.g., recycling behavior). To drive the idea home, imagine that the desired outcome of a new museum exhibition is that individuals improve their knowledge about invasive (non-native) plant species in their region. A *less direct* measure of visitors' understandings about invasive species might ask individuals to rate their level of knowledge. A *more direct* measure might ask individuals to answer test questions about invasive species; and, an even more direct measure might ask individuals to walk through a garden and play a game in which they identify various native and non-native plant species.

Often, *more direct* measures of learning resemble traditional tests that "tend to be at odds with the engaging, continuous, and exploratory nature of these [informal and afterschool] environments" (Zapata-Rivera, 2012, p. 1; NRC, 2015). Many informal education providers and evaluators, then, are reluctant to "put people on the spot" and make them feel uncomfortable, underperforming, or worse. These negative feelings may undermine key ISE goals, such as creating positive STEM experiences, building confidence in STEM, and supporting STEM-rich identities. The most direct measure in our example—the walk through the garden—reduces some of the negative aspects of using formal tests and comes closer to observing behavior in a criterion situation.

Direct measures of knowledge and skills are rarely employed in ISE evaluations, but they can be found. In an evaluation of an astronomy exhibition, for example, evaluators sought to assess participants' proficiency with a telescope. Instead of asking participants to rate their own skills or take a written test about telescopes, evaluators timed how long it took each person to point and focus a telescope (Sneider, Eason, & Friedman, 1979).

Some evaluators have tried to make direct assessments more fun and less intimidating. For example, hand-held "clicker" devices were used to assess students' knowledge of environmental concepts in an evaluation of a residential environmental education program for school groups (Kearney, 2009). Multiple-choice questions with "colorful pictures and graphics" (p. 61) were projected onto a screen, and students used the clickers to submit their responses electronically and anonymously. Students were also informed, "It's not a test, and nobody knows what you have answered, so don't worry" (p. 100). Shifting the format of the assessment from paper-and-pencil to clickers, tablets, or other platforms may alleviate test anxiety.

In sum, we seek to measure learning outcomes *more directly*. However, traditional "more-direct" measures—especially of STEM content knowledge—often disrupt the learning experience to such a degree that

they are typically inappropriate for informal education settings. Next, we discuss this disruption to the learning experience.

Obtrusiveness of Measures

We use the term obtrusiveness to indicate how much a particular measure disrupts or intrudes upon a participant's "natural" learning experience. Again, we find it useful to conceptualize this as a continuum, with measurement methods that range from less to more obtrusive. Surveys, interviews, and tests typically fall toward the *more obtrusive* end of the continuum because participants must stop what they are doing and answer questions, whereas observations of participant actions recorded by an inconspicuous observer would fall toward the *less obtrusive* end.

Because they are relatively *less obtrusive*, observations are a staple of many ISE evaluations. For example, imagine that a desired outcome of a new zoo exhibition is that visitors engage with key elements for some minimum amount of time. A *more obtrusive* measure of engagement might stop visitors and ask them how long they spent at the exhibition and what they did while there, whereas a *less obtrusive* measure might involve trained observers who covertly track and time visitors' movements.

Indeed, timing and tracking is a classic example of an unobtrusive measure: An evaluator follows visitors and collects data on where they go, how long they spend there, and what they do (Serrell, 1998; Yalowitz & Bronnenkant, 2009). Depending on the skill of the observer, the measure is usually unobtrusive, in the sense that it does not require the visitor to stop and answer questions. Sensor-based tracking systems provide new and automated ways to track visitor movements through informal learning spaces (Mygind & Bentsen, 2017). Timing and tracking data can be used to identify and compare what exhibition elements are more and less successful in attracting and holding visitor attention or prompting other observable behaviors.

Unobtrusive Measures and Ethical Considerations

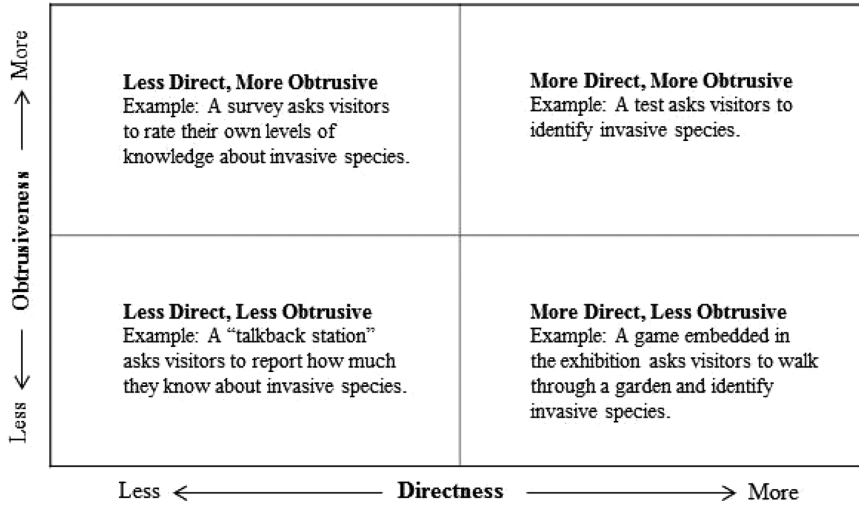
"Do no harm" is an underlying tenet of any evaluation, and the idea of capturing data unobtrusively raises ethical concerns about privacy and consent. Arguably, an unobtrusive measure is not necessarily unethical. A classic example of an unobtrusive measure is the physical erosion of tiles at a museum exhibition; variation in the wear and tear of floor tiles may indicate the relative popularity of various exhibit elements (Webb et al., 1965/2000). Today, pressure or heat sensors in exhibition areas can similarly indicate visitation rates and exhibit popularity. Such data can be collected anonymously, and the risk to participants is minimal to nonexistent. With creativity and planning, we may be able to identify new, unobtrusive data sources that "provide alternatives by which ethical criteria can be met without impinging

on important interests of the research subjects" (Webb et al., 1965/2000, p. ix). Evaluators who employ unobtrusive measures must be mindful about protecting human subjects and keeping data collection and use within the boundaries of ethical evaluation practice. See, for example, American Evaluation Association's (2018) updated *Guiding Principles for Evaluators* (principles A.6, D.2, D.3, and D.4) and Visitor Studies Association's (2008) *Evaluator Competencies for Professional Development* (competency C). Allen & Peterman (this issue) also discuss ethical issues in evaluation.

"Directness" and "Obtrusiveness" Considered Together

Consider "directness" and "obtrusiveness" simultaneously (Figure 2.1). For any given outcome, there are potentially different ways of measuring it, each of which can be characterized by its directness and obtrusiveness. Where a particular measure falls in this intersection of directness and obtrusiveness depends on what construct it is trying to measure and the particular context in which it is being employed. The same instrument might fall in different places, depending on its purpose and use. For example, observations can be a *more direct* measure of museum visitation patterns but a *less direct* (and

Figure 2.1. Directness and obtrusiveness quadrants. *Directness* represents how closely the measure approximates the "criterion situation" of interest (more or less directly measures the outcome). *Obtrusiveness* represents how much the measure intrudes on the informal learning experience (more or less obtrusive). Each quadrant includes an example measure of museum visitors' knowledge about invasive plant species.



probably unsuitable) measure of someone's intent or motivation; and observations may be *more* or *less obtrusive* depending on the skill of the observer, the setting, crowdedness, the type of program, awareness of participants, and so on.

Recall the earlier scenario of measuring knowledge about invasive plant species. Consider two of our example measures, the survey and the test. They differ in how *directly* they assess knowledge, but both appear in the top half of Figure 2.1 because both are relatively *obtrusive*, requiring visitors to pause what they are doing and answer questions. The third example, the garden walk, falls in the lower half of the figure and in the right-hand quadrant, as it is more direct and less obtrusive. A less direct and less obtrusive measure would be a visitor self-report that preserves the natural exhibition experience. For example, some museum exhibitions feature "talkback" stations or interactives that allow visitors to contribute their voices, opinions, stories, or experiences (e.g., Monterey Bay Aquarium, 2010). These stations range from technologies that record, display, and sometimes aggregate visitors' responses to simple "graffiti walls" where visitors add their responses to a wall for others to see (e.g., Grand & Sardo, 2017; Spicer, 2017).

More Direct and Less Obtrusive Measures

We contend that the measurement of ISE outcomes needs to include direct and unobtrusive measures. Depending on the construct, the audience, and the context, a measure that is *more direct* and *less obtrusive* may take vastly different forms. Mentioned earlier, tracking and timing studies are frequently conducted in evaluations of exhibitions, as they provide a relatively unobtrusive way to directly measure visitor movements and patterns of engagement. Fu et al. (2016) reviewed other examples: In an astronomy exhibition, students were given a raffle ticket and asked to choose which book they preferred to win—the number of astronomy books chosen (over non-astronomy books) was used as a measure of interest (Sneider et al., 1979); coupons for free admission to an art museum were distributed to study participants and then used to directly track return visit rates (Bowen, Greene, & Kisida, 2014); and, in a game about mixing colors, logs of students' choices made during gameplay revealed whether "students are trying to solve each problem in turn rather than discovering the general principle that governs the solutions to all problems" (Schwartz & Arena, 2013, p. 23). Advances in gaming, learning technologies, and data analytics have revealed opportunities to automatically and often unobtrusively capture and analyze data on learners' actions, which may reveal their knowledge, skills, and competencies (e.g., Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Tissenbaum, Kumar, & Berland, 2016; Zapata-Rivera, 2012).

Returning to Figure 2.1 and our hypothetical example, imagine a game that is designed as part of the museum exhibition. To be successful in the game, visitors must walk through a garden and correctly identify native and

invasive plants. Evaluators can track how visitors perform in this game as *more direct* evidence of knowledge about plant species. It is *less obtrusive* than a traditional test because the game is embedded into the exhibition, and visitors may choose to participate in this enjoyable activity without interrupting their museum experience. This type of measure is known as an embedded assessment, which we explore further in the next section.

Embedded Assessments

Some embedded assessments comprise a subgroup of *more direct* and *less obtrusive* measures of cognitive learning outcomes. They "allow learners to demonstrate their science competencies through tasks that are integrated seamlessly into the learning experience itself" (Becker-Klein, Peterman, & Stylinski, 2016, p. 1).[2] Many embedded assessments in ISE are also performance assessments. Performance assessments capture how learners execute tasks that are authentic to the targeted outcome. A classic example is the performance part of the test to receive a driver's license. While a written test measures driving-related knowledge, the road test directly measures driving proficiency by requiring an individual to operate a vehicle (Darling-Hammond & Adamson, 2010). In science classrooms, performance assessments often take the form of hands-on science investigations (e.g., Pine et al., 2006). A performance assessment provides a high-fidelity simulation of the real-world knowledge, skills, and competencies that we want learners to demonstrate, while standardizing to a large degree the prompt, task conditions, response format, and scoring procedures. Performance assessments have a longer history in schools, but work on performance-based and other types of embedded assessments in informal education is growing steadily. Since at least 2006, there have been presentations and discussions about the use of embedded assessments in informal education contexts (Becker-Klein et al., 2016).

At the Chicago Children's Museum, evaluators used an existing activity as an embedded assessment. At the "Skyscraper Challenge" exhibit, a computer took photographs of visitors while they built structures; visitors then selected their favorite photographs and recorded audio responses to a standard set of questions, resulting in a narrated book of their experience. This was originally designed solely as a learning activity, but evaluators recognized the assessment potential. With permission from visitors, evaluators analyzed the photographs and responses for evidence of problem-solving strategies, the use of STEM-based language and concepts, and more (Randi Korn & Associates, 2008).

[2] Embedded assessments have a longer history in formal education, where they are also characterized by their integration into instructional materials and activities (Wilson & Sloane, 2000). In classrooms, they may serve formative assessment purposes and may be deployed at critical junctures in the curriculum or in students' learning trajectories (Shavelson et al., 2008).

Another example of an embedded assessment comes from a residential environmental education program (Camargo & Shavelson, 2009). Students learned about water quality through hands-on experiences at one stream and their performance was evaluated. A few days later, students encountered a new stream during a hike and were asked to determine its water quality using familiar materials. Instructors collected systematic data on how students performed on this task, and the data from the two streams were analyzed and interpreted as evidence of student learning. This planned assessment was strategically embedded into the outdoor learning experience.

Much of the current work on embedded assessments is being conducted in the area of citizen science, which engages public audiences in scientific research projects. Most citizen science projects limit volunteers to data collection activities such as observation, identification, and monitoring, although some projects are more collaborative and even co-created (Becker-Klein et al., 2016). The quality of data collected by volunteers is critical to the success of citizen science projects, and the data-validation procedures that are already in place may represent opportunities to embed assessments of volunteers' observation and data collection skills (Becker-Klein et al., 2016; Peterman, Becker-Klein, Stylinski, & Grack Nelson, 2017). Games may also be used to unobtrusively and directly assess skills. For example, students at a summer camp competed in a game that required them to inventory trees in the local area; and their time and accuracy rates were compared over time and against inventories conducted by project staff and other citizen scientists (Becker-Klein et al., 2016; Peterman & Muscella, 2007). Allen and Peterman (this issue) discuss additional examples of embedded assessments used in online or digital environments.

Assessment Development Process. Measurement expertise, creativity, and a deep understanding of context and STEM content are required to develop direct assessments that also fit seamlessly into participants' informal experiences. Since much of what is known about embedded assessments is based on research in school settings, careful study of their development and use in informal settings, including evidence of their reliability and validity, is required as we move forward.

A team of evaluators and researchers recently outlined and studied an exploratory process for developing embedded assessments for citizen science (Peterman et al., 2017). The process begins with a clear articulation of program goals and acceptable evidence or indicators of meeting those goals. It then moves to developing and refining the assessment, which is accomplished through cycles of field-testing and the collection of validity evidence including expert reviews of the assessments and students' verbalization of their thinking while engaged in the assessment (also known as "think-aloud" data; see Ericsson, 2006; Leighton, 2004; Taylor & Dionne, 2000). The process was successfully used in three different citizen science projects to develop assessments of skills related to scientific

inquiry, including paying close attention, interpreting graphs and formulating conclusions, and precision of data collection (Peterman et al., 2017). All of the assessments were developed from activities that were already in place in the projects; two entailed scoring worksheets, and the third involved scoring data collection records.

Peterman et al.'s (2017) model is broadly in line with assessment development processes used in formal education (e.g., Ayala, Yin, Shavelson & Vanides, 2002; Mislevy, Steinberg, & Almond, 2003; NRC, 2001). Assessment development in any setting is an iterative process that requires clearly defining the construct to be assessed, designing a standard task that will elicit and capture responses from learners, developing a reliable method of scoring those responses, field-testing, and collecting data to bear on the argument that the assessment reliably and validly measures what it set out to measure for a particular purpose and context.

Building and Adapting Direct and Unobtrusive Measures: A Case Study

To further the work on embedded assessments in informal settings, we report on a study of the feasibility of developing and employing direct (performance-based) and unobtrusive (embedded) assessments in a STEM tinkering workshop. "Tinkering" has gained popularity in recent years as a "playful, collaborative, and inquiry-based approach" to foster STEM learning (Shea, 2015, p. 1; Vossoughi & Bevan, 2014). STEM-rich tinkering encourages learners to explore scientific concepts and phenomena by engaging them in collaborative, open-ended making and building activities that are often centered on everyday interests and use everyday tools and materials.

We highlight some of the design choices made and lessons learned during the iterative performance-measurement development process. The workshop in this case study is located in a low-income urban setting and serves essentially all ages from preschoolers to adults. The workshop features exhibits, hands-on STEM activities, and tinkering spaces, and it offers a high degree of freedom and flexibility to participants. While workshop staff and volunteers guide activities, answer questions, and provide safety oversight, participants are mostly free to choose what to do, how to do it, and how to define their own success.

Our goal was to develop a measure that would capture outcomes in a more direct and less obtrusive way than typically done. One major area of activity in the workshop was electricity and electric circuits. Based on conversations with workshop staff, we aimed to build or adapt an assessment around this topic, an assessment that directly and systematically measured knowledge, skills, and reasoning with series circuits in an unobtrusive and contextually sensitive way.

We were familiar with existing performance assessments called *Electric Mysteries* that targeted similar constructs (Shavelson, Baxter, & Pine, 1991; Stanford Education Assessment Laboratory, n.d.; see Figure 2.2); in fact, those assessments were co-developed by a co-author (Shavelson). The original assessments had strong reliability and validity evidence from prior studies in formal school contexts (Rosenquist, Shavelson, & Ruiz-Primo, 2000; Shavelson et al., 1991), but they had not been used in informal contexts. Given the availability of and our experience with *Electric Mysteries*, we decided to use them as a starting point for adaptation to the workshop context and to collect evidence for their reliability and validity in this setting and with this population.

Assessment Adaptation

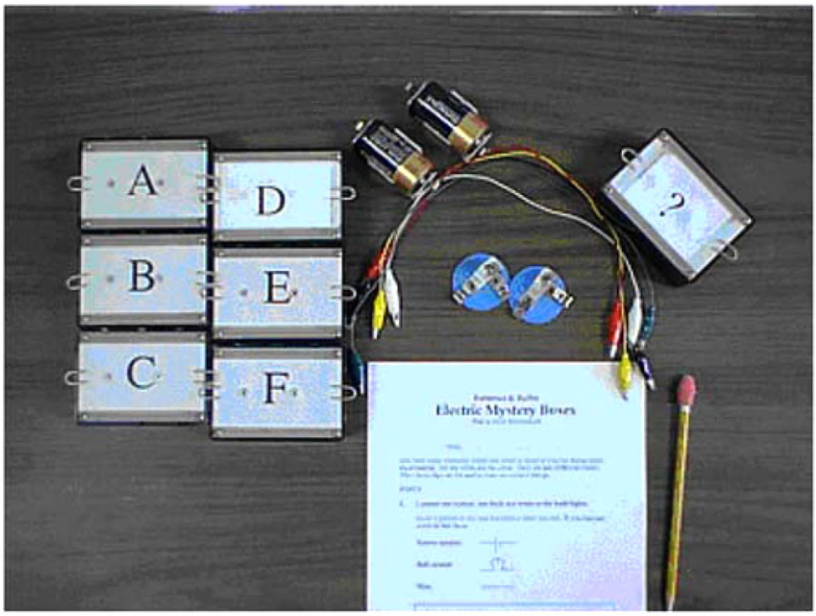
We adapted, tested, and revised early versions of the assessment over several visits to the workshop and many conversations with workshop staff, volunteers, and visitors. We also worked through materials set-up and management, task administration, methods for collecting participants' responses, timing, and other logistical matters. Here is what we learned.

Blend the Assessment Task Into the Workshop Context. To seamlessly integrate the assessment with other exhibits and activities in the workshop space, we framed them as "puzzles" and not as tests. This fit well with the context because puzzles on various topics were commonly offered as part of workshop programming. Further, our assessment was largely centered on manipulating physical materials commonly used in the workshop (i.e., wires, batteries, bulbs, motors).

Throughout the piloting process, we followed basic principles of museum exhibit design (e.g., Bitgood & Patterson, 1987; Serrell, 1996), even though we were building assessments and not exhibits. This was done to attract and maintain the attention of potential participants, especially given the myriad other activities and projects that an individual could choose from at any given moment. In general, we added color as well as attractive pictures and graphics, and we minimized text in labels and instructions. To better match the "homegrown" feel in the workshop, we also shifted away from signs made on a printer to ones that we wrote and drew by hand (Figure 2.3).

Preserve Participants' Freedom of Choice. We kept the assessment voluntary and designed multiple, open-ended participation pathways. To preserve participants' freedom of choice, we made it easy for people to opt in and out of doing the puzzles. To this end, we split the original hour-long assessment (designed for classrooms) into separate, shorter tasks that each required only a few minutes to complete; the simplest task required only a few seconds from most participants (Table 2.1). We also allowed participants to work on puzzles individually or in groups and to take as much or as little time as they wished.

Figure 2.2. The original *Electric Mysteries* performance assessment (Shavelson et al., 1991; Stanford Education Assessment Laboratory, n.d.). The top photograph shows the set-up with a variety of hands-on materials, and the bottom images display excerpts of the student response booklet.



Batteries & Bulbs

Electric Mystery Boxes

End of Unit Assessment

Name _____

You have some batteries, bulbs and wires in front of you for doing some experiments. All the wires are the same. They are just different colors. They have clips on the end so you can connect things.

PART I

1. Connect one battery, one bulb and wires so the bulb lights.

Draw a picture in the box that shows what you did. If you like use symbols like these:

Battery symbol:

Bulb symbol:

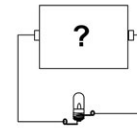
Wire:

2. Figure out what is in the mystery box labeled with a question mark "?".

The box has inside it either a battery or a wire:

or

To help figure out which one is in it, connect it in a circuit with a bulb:



Fill in the answer:

The **"?"** Box has a _____ in it.

Figure 2.3. A puzzle from the adapted *Electric Mysteries* performance assessment.

Table 2.1. Adapted *Electric Mysteries* Performance Assessment: Description of Puzzles

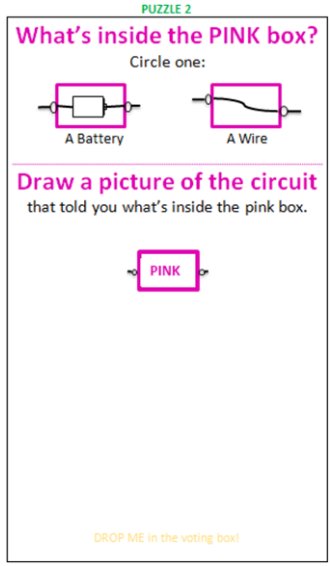
Puzzle Number and Name	*Item Number and Name*	*Description*
Puzzle 1: Light bulb puzzle	Item 1: Light bulb	Participants construct a simple circuit to light a bulb. This was a relatively straightforward and easy task, and 85% of those who attempted it performed correctly.
Puzzle 2: Pink puzzle	Item 2: Pink box	This puzzle asked participants to determine what circuit component (battery or wire) was inside a "mystery box." Just over half (53%) of those who attempted it answered correctly.
Puzzle 3: Electric puzzle mania	Item 3: Red box Item 4: Blue box Item 5: Yellow box	This puzzle consisted of three items, that is, three different mystery boxes. Participants were asked to determine the circuit component inside each box (red box contained a wire, blue box contained a bulb, and yellow box contained two batteries). Of participants who attempted the items, 28%, 19%, and 47% answered correctly on Items 3, 4, and 5, respectively; only 9% of respondents received full credit on all three items.
Puzzle 4: Motor puzzle	Item 6: Motor	This puzzle asked participants to use a motor and a fan to determine the polarity of a battery hidden inside a mystery box. Zero participants received full credit for this item, although 41% of those who attempted the puzzle received partially correct scores.

Having divided up the original assessment into separate tasks or puzzles, a challenge that we faced was how to sequence the puzzles. When we allowed participants to start with any puzzle, those who started with the most difficult puzzles sometimes appeared frustrated and reluctant to try more. When we suggested that participants try the puzzles in order from easiest to hardest (although we did not state their relative difficulty), many seemed to "warm up" and build confidence with their early successes. Yet, sequencing the puzzles may have had negative consequences for some participants. Some seemed to perceive an expectation to do all puzzles in the set; and, once they started, they may have felt less free to opt out or less successful if they chose to do so. We observed one case where a youth was pressured by an adult family member to complete all of the puzzles in the sequence; as he did so, he often looked around the room, and it seemed apparent that he wished to be doing something else. With only anecdotal evidence, we do not know whether puzzle order affected participants' performances or affective responses. It is possible that sequencing may matter more for certain types of participants; for example, younger or less knowledgeable participants may have difficulty accessing the harder puzzles and may benefit from starting with the easier puzzles.

Make the Response System Context-Friendly. In adapting the original classroom-based performance assessment, we sought a context-appropriate way to record student responses. Administering test booklets could be jarring in an informal setting. We briefly considered a computer-simulated version of the assessment in which students' responses were automatically scored (e.g., Pine, Baxter, & Shavelson, 1993); but, given the near-absence of screens in the workshop, we deliberately avoided any tablet, mobile, or computer-based systems. We piloted a number of "fun" response formats, using materials like whiteboards, markers, magnets, and stickers. However, the presence of additional materials on an already-full puzzle table required additional instructions and created confusion. Although we had initially tried to avoid a paper-and-pencil response system, we eventually chose paper ballots, which participants then deposited into voting boxes. The "ballot" was a piece of paper on which participants could "vote" for their answer and draw the circuit(s) made to solve the puzzle (Figure 2.4). Compared to the original test booklets, the ballots greatly reduced the amount of text to be read and written by participants. We observed that it was an intuitive format, as children immediately understood how to record and submit their answers. Workshop staff and volunteers also provided positive feedback and assured us that workshop visitors frequently sketched and recorded ideas on paper. In the end, paper ballots and voting boxes provided an easy, efficient, and inexpensive way to collect responses for scoring.

Offer Incentives for Completed Responses. A major drawback of the paper-and-pencil format was that it did not inspire participants to fully record their responses. Many stayed long enough to solve the puzzle and

Figure 2.4. Response "ballot" for an adapted *Electric Mysteries* puzzle.



"vote" for their answer but would hurry to leave before providing the explanatory drawing. After discussions with the workshop staff, we decided to offer small prizes such as pencils and erasers for completed ballots. The workshop staff reassured us that prizes were within the norms of their programs, and indeed, we saw an existing sign offering prizes to those who completed a different puzzle in the workshop. The prizes seemed to incentivize more students to complete their ballots, without undermining the voluntary nature of the activity; some students still declined to complete ballots, and some completed them but did not want a prize. We remain conflicted about this strategy, given the known downsides of using extrinsic rewards and incentives for participation in learning and evaluation activities (Lepper, Greene, & Nisbet, 1973; Ryan & Deci, 2000).

Assessment Psychometric Analyses

We collected responses from forty-one participants over multiple visits to the workshop. We conducted psychometric analyses to determine the "best" possible scale formed from puzzle responses, to examine scale reliability, and to do some preliminary and primitive validity analyses.

Scale Reliability. The four puzzles consisted of six items (Table 2.1) that could potentially contribute toward a final scale score. We examined the internal consistency (coefficient alpha) of the six-item scale, as well as different combinations of two or more items. The highest internal consistency (0.70) was obtained for a three-item scale of *Item 2: Pink Box*,

Item 5: Yellow Box (from *Puzzle 3*), and *Item 6: Motor*; deleting any one of these items decreased coefficient alpha. For subsequent data analyses, we constructed a total score that summed the scores on these three items, with each item weighted equally. The three-item scale made sense conceptually and seemed a reasonable representation of the entire set of items, as items from three of the four puzzles are represented in the scale. One of the excluded items was too easy (item difficulty or the percent of students answering correctly for *Item 1* was 85%); and the other two excluded items were too difficult (item difficulties of 28% and 19% for *Item 3* and *Item 4*, respectively).

Comparing Student Responses With Researcher Observations. We also compared the data collected on the paper ballots against researchers' observations. Although the observations were limited in number and depth, we used them as a check on whether students' ballots accurately represented their observed performances on the puzzles. We found a few instances when scores on ballots did not seem to fully reflect what students were observed doing. For example, some students were observed solving the puzzles correctly, but their partially completed ballots scored only partial or zero points; some of these students were observed being inattentive or rushed while filling out the ballots. In the majority of cases, however, the ballots seemed to accurately capture student proficiency at solving the puzzles, as what we saw on the ballots matched what we saw in our observations.

Score Interpretation—Validity. We attempted to examine the question of validity with the limited evidence that we had. We sorted participants into "experts" and "novices," hypothesizing that the experts would have higher total scale scores than novices. We based the sort on multiple indicators of circuits "expertise," including (a) participants' self-reports of their prior circuits experience, (b) the lead workshop educator's description of participants' prior circuits experience, (c) whether participants had previously participated in a workshop lesson on circuits, and (d) observations of students' performances on circuits-related tasks during our visits. However, all of these indicators were highly subjective and did not (separately or together) constitute a strong, independent measure of expertise. We found no statistically significant average differences between the performances of experts and novices, no matter which expertise indicator(s) was used. Future studies could compare performances on this assessment with performances on other, well-established circuits assessments, construct better measures of student "expertise," and/or employ think-aloud protocols with participants.

Moreover, we found no statistically significant differences in performance on puzzles among subgroups based on participants' gender, age, or whether they worked independently or as part of a group. Practically, these results are not very meaningful because sample sizes for some of the

subgroups were very small. Further studies with larger sample sizes are required.

As with any measure, the ballots provided only one perspective on our outcome of interest. Additional measures paint a fuller picture, and observations provided useful data that supplemented what we could have learned from the ballots alone. For example, we observed cases where one person explicitly shared a problem-solving strategy with other group members; and cases where individuals worked side-by-side on separate puzzles, occasionally narrating their actions. These observations provided a more nuanced understanding of the range of performances and interactions elicited by the assessment tasks. As a direct and in-depth measure of STEM knowledge, skills, and reasoning, the learning conversations that occur around these puzzles could be recorded and analyzed, as evaluators and researchers have done in other informal settings (e.g., Ash, 2003; Gutwill, Hido, & Sindorf, 2015; Leinhardt, Crowley, & Knutson, 2002). However, these methods are resource-intensive and can be more or less obtrusive, depending on the specific context and protocols employed.

Concluding Comment on Case Study

Performance assessments can fit naturally and unobtrusively into STEM learning experiences. Moreover, we found it possible to adapt a performance assessment for use in a dynamic ISE context—a tinkering workshop. We produced a reliable measure but the validity jury is out. What we needed and lacked were guidelines for building performance assessments for such contexts. We "tinkered" our way through the development and adaptation process, meeting many challenges and overcoming some. We are encouraged about the possibility of rigorously measuring students' hands-on performance in ISE environments; much work needs to be done.

Implications for the Future

In future evaluations of ISE programs, we urge a closer look at *more direct* and *less obtrusive* measures of outcomes (see Figure 2.1). To push forward, we need to continue studying and documenting how these measures are developed (or adapted) and implemented, and their strengths and weaknesses for different outcomes, audiences, settings, decisions, and so on. The goal is to build assessment frameworks that incorporate this knowledge and pass them along. New measures can reveal new evaluation insights, but challenges lie ahead. We share some thoughts on meeting these challenges in future.

Stand on the Shoulders of Existing Measures

There is a large body of existing measures and assessments in related fields, especially in formal STEM education. Using measurement expertise,

exhibit design principles, an understanding of the context, and ongoing feedback from participants and other stakeholders, we adapted existing classroom performance assessments to fit mostly seamlessly into an informal setting. The iterative adaptation process was critical. It is not advisable to take existing measures and use them as-is for a new context and purpose. Performance-based measures that target similar constructs may not always already exist, but as we build new measures, we should remember to scan the available literature for inspiration, possible adaptation, and clues to avoidable pitfalls.

Examine Evidence of Reliability and Validity

Regardless of whether the measure is developed from scratch or adapted from an existing measure, evaluators must consider the reliability and validity of their measures for the current context and purpose; see Grack Nelson, Goeke, Auster, Peterman, & Lussenhop (this issue) for a detailed discussion of this topic. Not every evaluation has the resources to thoroughly examine the reliability and validity of every measure; but the higher the stakes, the sharper the focus should be on a measure's technical qualities. Keeping these concerns at the forefront is important in summative evaluations or evaluations used to make high-stakes decisions. In our study, for example, we are convinced that responses can be scored reliably, but lacking additional validity evidence, we would be uncomfortable using the puzzles as the sole measure of proficiency with circuits, especially for a summative or high-stakes evaluation. Reflecting systematically on what is known about the reliability and validity of any measure is necessary for understanding the limitations of the data we collect, the other types of data we should consider, and the types of claims we can justifiably make.

Know the Context

The development of a measure that is embedded as an unobtrusive part of participants' learning experiences relies on understanding the program and evaluation context—the goals of the program, typical program experiences, motivations of participants, community culture, and more. The obtrusiveness of a measure is in the eye of the beholder. What may seem obtrusive to one person may go completely unnoticed by another. For example, some of our study participants were students in after-school and summer camp programs; these students expected a slightly more structured experience, so recording their answers on paper seemed unobtrusive and wholly within the bounds of what they typically do in the workshop. This was less the case with participants during the workshop's more freeform drop-in hours. An unobtrusive measure is a moving target, demanding a deep and flexible understanding of audience and context.

Use Technology Responsibly

As learning and assessment technologies become more sophisticated and accessible, they may offer innovative solutions for collecting and analyzing data in efficient, reliable, direct, and unobtrusive ways. In our study, video recordings of participants working on the puzzles would have allowed for in-depth analyses of learning conversations, but these would have demanded more resources than were available. Video analysis is a resource-intensive activity, but intelligent coding and scoring systems may be on the horizon (LeCun, Bengio, & Hinton, 2015).

Breakthroughs have also been made in systems for automatically capturing and analyzing participants' actions as they work through a task; such systems can improve data reliability and provide real-time feedback to learners. Researchers from a number of different disciplines have innovated ways to assess science inquiry skills in electronic learning environments, although most of these have been used in school settings (e.g., see review by Timms et al., 2012); these include data mining of students' log files to assess their skill at designing and conducting experiments (Gobert et al., 2013). In a museum setting, visitors' interactions with a tabletop circuit activity were automatically logged, and these data were used along with video data to study collaborative behaviors (Tissenbaum, Berland, & Lyons, 2017).

For any of these technologies, however, the value of the data and the validity of the interpretations made from those data are limited by the questions asked and the theories underpinning our analyses. "We must still find the signal within the noise . . . " (Bechtol, Cosbey, & Emmons, 2017). Further, new assessment and evaluation technologies demand careful procedures around informed consent, privacy, and data security. See Allen and Peterman (this issue) for more discussion of evaluation considerations in light of new technologies.

Final Word

The evaluation questions we ask should drive the choices we make about the methods we use. These choices largely determine the bounds of our understanding, as every measure sheds light on some aspect of the program but leaves something else in the shadows. We encourage evaluators to push for improved measures, particularly those that are *more direct* and *less obtrusive*, so that we may continue to advance our understanding of the outcomes of ISE.

References

American Evaluation Association. (2018). *Guiding principles for evaluators*. Retrieved from https://www.eval.org/p/cm/ld/fid=51

Ash, D. (2003). Dialogic inquiry in life science conversations of family groups in a museum. *Journal of Research in Science Teaching*, *40*(2), 138–162.

Ayala, C. C., Yin, Y., Shavelson, R. J., & Vanides, J. (2002). *Investigating the cognitive validity of science performance assessment with think alouds: Technical aspects*. Paper

presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bechtol, E., Cosbey, A., & Emmons, C. T. (2017). An introduction to the article "Constructivist analytics: Using data to enable deeper museum experiences for more visitors—lessons from the learning sciences" by Matthew Berland. *Visitor Studies*, *20*(1), 3.

Becker-Klein, R., Peterman, K., & Stylinski, C. (2016). Embedded assessment as an essential method for understanding public engagement in citizen science. *Citizen Science: Theory and Practice*, *1*(1), 8, 1–6.

Bitgood, S., & Patterson, D. (1987). Principles of exhibit design. *Visitor Behavior*, *2*(1), 4–6.

Bowen, D. H., Greene, J. P., & Kisida, B. (2014). Learning to think critically: A visual art experiment. *Educational Researcher*, *43*(1), 37–44.

Camargo, C., & Shavelson, R. (2009). Direct measures in environmental education evaluation: Behavioral intentions versus observable actions. *Applied Environmental Education & Communication*, *8*(3), 165–173.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University. Retrieved from https://scale.stanford.edu/system/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning.pdf

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 223–242). Cambridge, UK: Cambridge University Press.

Friedman, A. J. (Ed.). (2008). *Framework for evaluating impacts of informal science education projects*. (Report from a National Science Foundation workshop.) Arlington, VA: National Science Foundation. Retrieved from http://www.informalscience.org/sites/default/files/Eval_Framework.pdf

Fu, A. C., & Kannan, A. (2016). *Beyond self-reports: Direct measures of informal learning experiences*. Presentation at the 29th annual Visitor Studies Association conference, Boston, MA.

Fu, A. C., Kannan, A., Shavelson, R. J., Peterson, L., & Kurpius, A. (2016). Room for rigor: Designs and methods in informal science education evaluation. *Visitor Studies*, *19*(1), 12–38.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521–563.

Grand, A., & Sardo, A. M. (2017). What works in the field? Evaluating informal science events. *Frontiers in Communication*, *2*(22), 1–6.

Gutwill, J. P., Hido, N., & Sindorf, L. (2015). Research to practice: Observing learning in tinkering activities. *Curator: The Museum Journal*, *58*(2), 151–168.

Kearney, A. R. (2009). *IslandWood evaluation project: Assessment of student outcomes from IslandWood's School Overnight Program* (Report prepared for IslandWood, Bainbridge Island, WA). Retrieved from http://www.peecworks.org/PEEC/PEEC_Research/S0179A96D-0179A9BD

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15.

Leinhardt, G., Crowley, K., & Knutson, K. (Eds.). (2002). *Learning conversations in museums*. Mahwah, NJ: Lawrence Erlbaum.

Lemke, J., Lecusay, R., Cole, M., & Michalchik, V. (2015). *Documenting and assessing learning in informal and media-rich environments*. (The John D. and Catherine T. MacArthur Foundation Reports on Digital Media and Learning.) Cambridge, MA: MIT Press. Retrieved from https://mitpress.mit.edu/books/documenting-and-assessing-learning-informal-and-media-rich-environments

Lepper, M. R., Greene, D., & Nisbet, R. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of "overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*(1), 129–137.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3–62.

Monterey Bay Aquarium. (2010). *Hot Pink Flamingos: Stories of hope in a changing sea. Exhibit press kit*. Retrieved from http://storage.montereybayaquarium.org/storage/pressroom/presskit/pdf/hot%20pink%20flamingos%20press%20kit.pdf

Mygind, L., & Bentsen, P. (2017). Reviewing automated sensor-based visitor tracking studies: Beyond traditional observational methods. *Visitor Studies*, *20*(2), 202–217.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.

National Research Council. (2002). *Scientific research in education*. Washington, DC: The National Academies Press.

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

National Research Council. (2015). *Identifying and supporting productive STEM programs in out-of-school settings*. Washington, DC: The National Academies Press.

Peterman, K., Becker-Klein, R., Stylinski, C., & Grack Nelson, A. (2017). Exploring embedded assessment to document scientific inquiry skills within citizen science. In C. Herodotou, M. Sharples, & E. Scanlon (Eds.), *Citizen inquiry: A fusion of citizen science and inquiry learning* (pp. 63–82). New York: Routledge.

Peterman, K., & Muscella, D. (2007). *Games as embedded assessments*. A panel presentation at the annual meeting of American Evaluation Association, Baltimore, MD.

Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., … & Foley, B. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. *Journal of Research in Science Teaching*, *43*(5), 467–484.

Pine, J., Baxter, G., & Shavelson, R. J. (1993). Assessments for hands-on elementary science curricula. *MSTA Journal*, *39*(2), 3–5.

Randi Korn & Associates, Inc. (2008). *Summative evaluation of the Skyline exhibition*. Prepared for the Chicago Children's Museum. Retrieved from http://www.exhibitfiles.org/dfile2/ReviewFinding/461/original/2008_RKA_CCM_Skyline3_summ_dist.pdf

Rosenquist, A., Shavelson, R. J., & Ruiz-Primo, M. A. (2000). *On the "exchangeability" of hands-on and computer simulation science performance assessments*. Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, UCLA.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Schwartz, D. L., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age*. Cambridge, MA: The MIT Press. Retrieved from https://mitpress.mit.edu/books/measuring-what-matters-most

Serrell, B. (1996). *Exhibit labels: An interpretive approach*. Walnut Creek, CA: AltaMira Press.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, *4*(4), 347–362.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., … Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education*, *21*(4), 295–314.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Marino, J. P. (2018). International performance assessment of learning in higher education (iPAL): Research and development. In O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, & C. Kuhn (Eds.), *Assessment of learning outcomes in higher education: Cross-national comparisons and perspectives* (pp. 193–214). New York: Springer.

Shea, M. (2015). *Tinkering in STEM education*. Retrieved from https://www.exploratorium.edu/sites/default/files/pdfs/connectedcollection_tinkering.pdf

Sneider, C. I., Eason, L. P., & Friedman, A. J. (1979). Summative evaluation of a participatory science exhibit. *Science Education*, *63*(1), 25–36.

Spicer, S. (2017). The nuts and bolts of evaluating science communication activities. *Seminars in Cell & Developmental Biology*, *70*, 17–25.

Stanford Education Assessment Laboratory. (n.d.). *Batteries and bulbs*. Retrieved from https://educationassessmentlab.weebly.com/batteries-and-bulbs.html

Taylor, K. L., & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, *92*(3), 413–425.

Timms, M., Clements, D. H., Gobert, J., Ketelhut, D. J., Lester, J., Reese, D. D., & Wiebe, E. (2012). *New Measurement Paradigms*. Report prepared by the New Measurement Paradigms working group of Community for Advancing Discovery Research in Education (CADRE). Retrieved from http://cadrek12.org/sites/default/files/NMP%20Report%20041412_0.pdf

Tissenbaum, M., Berland, M., & Lyons, L. (2017). DCLM framework: Understanding collaboration in open-ended tabletop learning environments. *International Journal of Computer-Supported Collaborative Learning*, *12*(1), 35–64.

Tissenbaum, M., Kumar, V., & Berland, M. (2016). Modeling visitor behavior in a game-based engineering museum exhibit with hidden Markov models. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 517–522). Raleigh, NC: International Educational Data Mining Society.

Visitor Studies Association. (2008). *Evaluator competencies for professional development*. Retrieved from https://www.visitorstudies.org/evaluator-competencies

Vossoughi, S., & Bevan, B. (2014). *Making and tinkering: A review of the literature*. (Paper commissioned by the National Research Council Committee on Successful Out-of-School STEM Learning.) Retrieved from http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_089888.pdf

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (2000). *Unobtrusive measures* (rev. ed.). Thousand Oaks, CA: Sage. (Original work published 1965)

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181–208.

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

Zapata-Rivera, D. (2012). *Embedded assessment of informal and afterschool science learning*. (Paper commissioned for Summit on Assessment of Informal and Afterschool Science Learning, convened by Board on Science Education, National Academy of Sciences.) Retrieved from http://sites.nationalacademies.org/DBASSE/BOSE/DBASSE_080110

ALICE C. FU, PhD, is an independent researcher and consultant focusing on informal STEM and environmental education; her research interests include STEM assessment and evaluation, the interface between formal and informal learning environments, and research-practice connections.

ARCHANA KANNAN, MA, MS, is a PhD student in Curriculum and Teacher Education at Stanford University; her research interests include informal science and environmental education, assessment and evaluation in ISE, and professional training for informal educators.

RICHARD J. SHAVELSON, PhD, is I. James Quillen Dean (Emeritus) and Margaret Jacks Professor of Education (Emeritus) at Stanford University specializing in measurement, statistics, and evaluation, especially in STEM education.