Allen, S. & Peterman, K. (2019). Evaluating informal STEM education: Issues and challenges in context. In A. C. Fu, A. Kannan, & R. J. Shavelson (Eds.), *Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation*, *161*, 17–33.

1

Evaluating Informal STEM Education: Issues and Challenges in Context

Sue Allen, Karen Peterman

Abstract

We define "informal STEM education" and explain some of the reasons its outcomes are so inherently challenging to evaluate, including the critical need for ecological validity and the fact that many informal learning experiences are low-visibility and opportunistic. We go on to highlight significant advances in the field, starting with the fundamental embracing of learning outcomes that go well beyond narrow measures of knowledge and skills, to include interest, engagement, and identity-building. Within that framework, we note the development of shared constructs and shared instruments emerging in multiple sectors of informal STEM education. We also highlight advances in unobtrusive instrumentation and powerful analytic techniques that make it possible to evaluate learners' unfolding experiences more directly than ever before. Finally, we point to underlying factors that support a growing and maturing professional community of informal STEM learning evaluators, and some of the "learning ecosystem" metaphors that frame their thinking.

This chapter provides a foundational discussion of challenges inherent in evaluation of informal science, technology, engineering, and mathematics (STEM) education, focusing on why it is so extraordinarily difficult to clearly define and usefully measure learning outcomes in these contexts. While these challenges are framed in relation to informal

STEM education, parallels exist in other contexts as well: policy, health, environmental and social justice, to name just a few. We detail some recent advances, both technical and systemic, that have allowed an explosion of creative methods and approaches. We describe ourselves as a young field that has often struggled to be "good enough" to meet the gold standards of evaluation methods. As a community, we are gradually coming of age and find ourselves on the frontlines of studying learning environments in the twenty-first century.

Defining Characteristics of Informal STEM Education

For the current purposes, we define "informal STEM education" (ISE) as "lifelong learning in science, technology, engineering, and math (STEM) that takes place across a multitude of designed settings and experiences outside of the formal classroom" (Center for Advancement of Informal Science Education [CAISE], 2017). A landmark report describes learning experiences in such settings as "guided by learner interests, voluntary, personal, ongoing, contextually relevant, collaborative, nonlinear, and open-ended" (National Research Council [NRC], 2009, p. 11).

While the word "informal" might at first glance suggest a lack of rigor, it actually refers to the nature of the learning setting rather than the STEM content or practices being learned. In fact, informal learning may lead to high levels of domain-specific expertise among those who are motivated to continue their learning, such as that of an experienced hobbyist, a citizen scientist, or a *competent outsider* who becomes expert in a topic as a result of its relevance to their personal life or community (Feinstein, 2011). Hobbyists, for example, might include master gardeners who blend informal training and their own experiences to offer gardening advice to the public at a weekly farmers market. Similarly, the world of citizen science abounds with informal learners who are trained to collect rigorous data in support of scientific research. Competent outsiders in this field include the citizens who "broke" the story of water contamination in Flint, Michigan, as well as a community of citizens in the San Francisco Bay Area who banded together to study and fight for better air quality. Learners in informal STEM settings know they are not held accountable for what they learn in any "high-stakes" way, such as grading that might affect career advancement, promotion, or future prospects, but this does not limit the quality of their learning. On the contrary, a recent meta-analysis suggests that intrinsic motivation predicts the quality of a person's performance even more strongly than the presence or absence of such external incentives (Cerasoli, Nicklin, & Ford, 2014).

The vast majority of learning (including STEM learning) happens outside the formal school day; a widely published graph on "lifelong and lifewide learning" makes this point (Learning in Informal and Formal Environments [LIFE] Center, 2005; see http://life-slc.org/about/citationdetails.html). STEM learning begins in early infancy and

extends throughout life, through social experiences with the natural and designed worlds, out-of-school programs and clubs, games and other virtual learning experiences, broadcast media and web-based explorations, and visits to places such as zoos, aquariums, and museums. We see informal STEM education as an essential and increasingly integrated complement to the formal schooling system, which may also provide deep opportunities for self-directed learning but is unavoidably influenced by unifying standards for all, age-based curricula, and high-stakes assessment practices. Projects and programs that offer some form of informal STEM education usually see their efforts as contributing to learning outcomes that accumulate for each individual over time and space in idiosyncratic and highly personalized ways.

Informal STEM Education and Fundamental Methodological Issues

If we review the key terms used to define informal STEM education in this chapter thus far, they include words such as self-directed, idiosyncratic, and highly personal. None of these terms describes ideal circumstances for research and evaluation from a traditional perspective, and yet there is a real need and passion for studying learning in these contexts.

Perhaps the most profound underlying methodological challenge is that of maintaining ecological validity. To be ecologically valid, evaluation of these experiences must not undermine precisely those characteristics that make the experiences different from formal schooling: jointly negotiated and evolving goals, low-stakes accountability, and freedom of choice. This key constraint shapes all aspects of high-quality informal STEM evaluation: identifying appropriate outcomes, choosing appropriate instruments, embedding them in the activity without disrupting it, analyzing data that often resists aggregation or comparisons, and drawing inferences and making recommendations that are fair and useful to practitioners and theorists.

Educational assessment techniques that are the accepted norm in formal settings, such as pre–post testing, formal surveys, closed-response questions, and even interviews, may undermine the very nature of the brief, voluntary, and emergent learning experiences that are the hallmark of informal environments. Similarly, experimental study designs that incorporate controls may be unfeasible if the central premise of the learning experience is one of free choice and individual interpretation. It is notoriously difficult to develop "rigorous" designs and methods for contexts where participants typically expect an enjoyable, non-threatening experience. Even federal agencies that fund informal STEM education programs have historically been granted exemption from traditional evaluation requirements based on the challenges of sampling, assessment, and causal inferences (U.S. Department of Education, 2007).

Defining the character and bounds of the program, product, or "treatment" to be evaluated is a significant challenge, because learners are free

to come and go, participating in ways that evolve over time and that may be hard to characterize or quantify. "Dosages" are generally highly variable and may be vanishingly small. For example, science museums routinely describe the "holding power" of their exhibits using histograms of time-spent by visitors, which is usually measured in seconds or minutes; the histograms often show an asymmetrical distribution whose mode is the shortest-recorded time interval. Systems to quantify the duration or depth of a learning experience vary greatly across organizations. For example, museums and out-of-school time programs typically count people coming in the door; television broadcasters use probabilistic models based on household sampling; and online game developers use clickstream analysis. Few of these methods are able to count distinct individual learners, a prerequisite for any analysis of the cumulative impacts of multiple experiences over time. This forces evaluators to rely heavily on learners' self-reports when trying to characterize the "dosage" of a designed learning experience.

In the face of these challenges, some evaluators have nevertheless been pushing the boundaries of what is methodologically feasible. A recent review of reports on the CAISE web site advocated for increased rigor in the evaluation of informal science learning programs (Fu, Kannan, Shavelson, Peterson, & Kurpius, 2016). They call for use of quasi-experimental designs, including the use of deferral comparison group designs, and epidemiological approaches that use statistical models and a systems approach to understand how constellations of demographic, process, and dispositional variables result in outcomes of interest. In the context of museum studies, Allen et al. (2007) advocated for the use of counterbalanced designs and exit interviews with random assignment, in addition to a variety of culturally responsive methods such as naturalistic inquiry and case studies.

Finally, it is worth mentioning a pragmatic consideration that shapes much of the informal STEM evaluation literature: the public visibility of the experience. Informal learning experiences can be seen as lying on a continuum in terms of their salience on the community public stage. At one end of the continuum are events that might be publicized in local news media (such as science festivals, math competitions, or robotics tournaments). At the other end of the spectrum are brief, opportunistic experiences that are highly integrated in daily life (such as a family stopping in a grocery store to compare the prices of different quantities of milk, a couple walking in a park and discussing why leaves fall, or a person at home listening to a radio segment about the evening sky).

Learning experiences with low visibility are particularly challenging to evaluate. A brief, opportunistic experience in the home, such as stumbling on an interesting television program or following an intriguing YouTube link while browsing, is integrated into the timing, values, and norms of daily life; and participants may not recognize the event as a STEM learning experience worthy of reflection. Further, it is unlikely that an evaluator can be present to observe the event unless he or she invests in deep

ethnographic work (such as Ellenbogen, 2002), which is very resource-intensive and feasible only for a small number of learners. Last, it is such a brief experience that operationalizing any kind of assessment is difficult, even if one had a clear idea of the learning about to take place.

Publicly visible learning experiences are usually easier to evaluate. They are typically topic-focused and have extended duration, so there is more opportunity for a measurable impact. They take place at a specific place and time, allowing an evaluator to use their time efficiently and plan to attend. The learners are a somewhat "captive audience," expecting to be on site for a while, thereby allowing the evaluator a window of opportunity to interact with them. Finally, because they have a name and a fixed-term public presence, publicly visible learning experiences are easier for learners to reflect on in a coherent way when responding to surveys or interviews.

Most of the examples presented in this issue share this bias, drawing chiefly from high-profile events such as museum exhibits, out-of-school-time programs, citizen science initiatives, and festivals.

Learning Outcomes

Throughout the chapters in this issue, we focus on the specific challenges of *identifying and assessing the outcomes* of *intentionally designed* informal STEM learning resources, programs, and interventions. We choose this focus because it is central to evaluation, of significant interest to funders, a topic of heated debate for decades, and one that has shown significant advances over the last decade. While we recognize that the field of informal STEM research and evaluation is an international one, the chapters are largely U.S.-centric, reflecting our greater awareness of the U.S. landscape and the systemic changes in it.

One perennial issue in identifying appropriate outcomes in informal learning settings is that the outcomes that are used most often for school-based settings (such as acquisition of STEM content knowledge and skills) are often too limited in how they define impact. While broader definitions of outcomes and impact are relevant to both formal and informal settings, those of us who evaluate the latter give equal priority to a wide range of outcomes that include content knowledge and skills but also attitudes, awareness, behaviors, and identity development (Friedman, 2008; NRC, 2009). Informal learning experiences are seldom primarily about imparting science knowledge and skills; more often, these experiences aim to spark curiosity, build interest, and foster intrinsic motivation as "stepping stones" to further science learning. This has direct implications for evaluation: Constructs such as interest, motivation, and curiosity are more challenging to define, operationalize, and measure (NRC, 2009). Evaluators in the visitor studies community tell a common story: Their colleagues dismiss the value of learning in informal settings, saying, "I can see the kids are having fun, but are they really learning?" to which the evaluator replies, "In school,

I can see the kids learning in a narrow sense, but are they really interested, or will they drop the subject as soon as they can?" Almost all evaluations of informal STEM education include interest as a foundational outcome, if not a key component of learning per se.

Over the past decade or so, there has been huge progress in defining a set of key constructs for the field to use as outcome categories that reach beyond content knowledge. Several major reports, both in the United States and Europe (Dorph, Cannady, & Schunn, 2016; Friedman, 2008; Hooper-Greenhill, 2004; Krishnamurthi, Ballard, & Noam, 2014; NRC, 2009) created explicit frameworks that define learning as multifaceted, with constructs such as sparking interest and building identity. There was a lot of similarity among these frameworks: They used abstractions that were narrow enough to be operationalized and assessed, yet broad enough to capture the full range of possible outcomes desired by educators and designers.

Looking Ahead

In the remainder of this chapter, we discuss how evaluators and others are responding to the challenges inherent in evaluating learning outcomes in informal STEM education. We highlight four innovations that may change the future of evaluation in this sector: (a) the development of shared instruments to assess common outcomes; (b) advances in instrumentation that support unobtrusive data collection and analysis; (c) a broader, stronger, and more cohesive informal STEM education community; and (d) growing awareness and use of the concept of "learning ecosystems."

Common Outcomes and Shared Measures

Even with a robust outcome framework, an additional question is whether the outcomes can be standardized across programs and experiences. Like the experiences themselves, the outcomes of informal STEM education are notoriously idiosyncratic, personalized, and unpredictable. Appropriate outcomes need to take into account both the nuanced nature and intent behind a designed learning resource, and the moment-by-moment forms of engagement, goals, and interpretations by the learners. Theoretically, if multiple programs address the same outcome, then a single instrument could measure that outcome across programs; we could develop shared measures of common outcomes.

Yet, for decades, many in the informal learning community largely resisted pressures to standardize their instruments, arguing that to do so would undermine ecological validity and imperil the field. There has always been a legitimate fear that shared measures will strip away all that is good about informal learning—that informal learning programmers will begin to design narrow experiences that "teach to the test." At the same time, some voices within the field, notably those of Beverly Serrell, John Falk, Alan Friedman, Gil Noam, and Rick Bonney, argued that there was

a place for shared measures, particularly if they were easy to implement; highly focused on one or two constructs, such as interest and engagement in science; and targeted toward the key goals of a great many designed programs and resources.

Some of the early movement toward shared measures came from the Program in Education, After School & Resiliency (PEAR) at Harvard University and McLean Hospital, which published a report on the need for systemic assessment in informal learning environments (Hussar, Schwartz, Boiselle, & Noam, 2008). The authors specifically noted the need for a common bank of questions to be used across sites, and they argued for creating a set of tools to be utilized by a significant number of programs across the country. They also recognized that many informal learning programs were anxious about using standard tools for evaluation; many afterschool programs included in the study, for example, had already developed their own evaluation instruments. This tension between shared and custom evaluation instruments still exists today. With funding from private foundations, the PEAR group went on to develop the "Common Instrument Suite" (Noam, Allen, Shah, & Triggs, 2017), designed to assess a small, tight set of outcomes with the minimum possible number of questions, thereby allowing programs time to also assess more nuanced and individualistic outcomes specific to their programs.

The tide was also turning in other sectors of the informal learning world. Beverly Serrell's (1998) definitive work in standardizing tracking and timing methods in museums gave the field its first opportunity to compare visitor behaviors (a long-accepted correlate or, at least, a necessary condition of learning). Serrell's work showed that data collected using a shared instrument across hundreds of different topics and settings not only could be aggregated but also could generate useful new constructs (such as "percent of diligent visitors") and provide provocative comparisons about the effectiveness and efficiency of resources spent.

In the years that followed, others pioneered the development of scales specifically for use in cross-project evaluations of informal education. Grounded in theory and developed with psychometrics to support their use across contexts, these shared measures have tremendous potential to propel evaluation and research about informal learning outcomes. Some focus on public engagement with science and, specifically, the outcomes of volunteer citizen scientists (e.g., the Developing, Validating and Implementing Situated Evaluation Instruments [DEVISE] scales; Cornell Lab of Ornithology, 2014); others target constructs such as the "activation" of science learning that can bridge formal and informal contexts (Activation Lab, 2018). Most recently, the American Association for the Advancement of Science commissioned the development of shared measures for scientists who participate in public engagement activities (Peterman, Robertson Evia, Cloyd, & Besley, in press; Robertson Evia, Peterman, Cloyd, & Besley, 2017).

We see these sector-specific instruments as a significant step forward for the field of informal STEM learning evaluation and research. Open conversations about whether and how to use them are widespread and ongoing; and, in most cases, the developers of these instruments are leading the way in helping others decide when and how to use them in new contexts. This makes for exciting partnerships across research, evaluation, and practice. We anticipate reaping the full benefits of the work, as many of the massive data collection efforts are shifting to analysis and publication phases. Going forward, training will be needed to prepare evaluators to make data-driven and intentional choices about existing instruments—how to determine which one(s) might be a good fit for their projects, while remembering that shared instruments are only part of a complete evaluation. These efforts are described more deeply, along with their technical qualities, in Chapter 3 (Grack Nelson, Goeke, Auster, Peterman, & Lussenhop, this issue).

Advances in Unobtrusive Instrumentation

The powerful need for ecological validity in studies of informal settings encourages evaluators to use techniques that avoid interrupting the flow and emotional tenor of participants' experiences. This tends to put informal STEM evaluators in the role of detectives, constantly searching for noninvasive ways to collect data.

One approach to this challenge is to use *embedded* assessments as much as possible, in an effort to leverage the residues or artifacts from authentic learning activities as data sources for separate analysis. This approach is particularly useful when learners leave a consistent and unambiguous set of markers of their activity. For example, when using an online game or simulation, or navigating virtual or augmented reality, every action of the learner can be digitally recorded in a continuous clickstream; the main challenge of the analysis is to interpret the intentions, understandings, and reasoning that underlie participants' actions (e.g., Owen, 2014). Another example of embedded assessment is found in online gaming environments or digital experiences that are social in nature; when participants leave a trail of comments or contribute to a community forum, evaluators can collect and analyze these communications and reflections as data for evaluation. Research on the online citizen science project Zooniverse, for example, has begun to document how online discussions facilitate learning by focusing on the sophistication of the scientific terminology used in online discussions (Luczak-Roesch et al., 2014).

Another approach to maintaining ecological validity is to collect data in a *covert* manner, outside of or on the periphery of awareness of the learners. This includes the video- or audio-recording and subsequent analysis of learning activities; this approach is especially useful in cases where learning is highly social and negotiated, such as families interacting with exhibits,

small groups of youth in programs, or conversations within a community group. In this kind of covert assessment, there are generally two categories of issues to be addressed: ethical and technical. On the ethical side, evaluators need to work with their institutional review boards (IRB) to ensure that the data, which may reveal the individual identities of participants, are collected with appropriate forms of consent or assent, used in ways consistent with their stated promises, and follow the ethical guidelines of the American Evaluation Association. On the technical side, data quality depends on the degree to which conversations and activities can be captured and recorded with minimal disruption of the activity. For this reason, covert data collection in informal learning settings has always benefited from advances in observation technology.

With regard to technology, there has been an explosion in both hardware and software in the last decade to support covert assessment methods. With the ubiquity of smartphones, tablets, go-pro cameras, and the like, video-recording has gone from being a rare practice in museums and live programs to being a standard tool for both evaluation and research. Some museums (e.g., Exploratorium, Hatfield Science Center) continue to push the boundaries of what can be observed, by installing high-quality surveillance cameras in ceilings to track visitor movement, while cordless microphones are carried by visitors or discreetly attached to exhibits. Facial and gestural recognition has advanced to the point where individuals can be individually sexed, tracked, and timed, all without leaving any identifiable information (Rowe, 2012). Radio-frequency identification tags and their descendants have made it possible to track visitors while also recording critical aspects of their exhibit interactions, such as their choices and actions, in ways that are intrinsically motivating to them and also provide valuable evaluation data (Hsi & Fait, 2005; Kanda et al., 2007).

On the analysis side, video-annotation tools (e.g., studiocode, vimeo) and qualitative data analysis software (e.g., NVivo, AtlasTI) have become less expensive and easier to use, making it possible to code and analyze learning activities with a resolution of a second or less. A range of software products support not only massive data aggregation but sequencing comparisons that parallel methods in genome sequencing (Ma, 2016). Voice recognition and semantic analysis software are tantalizingly close to the point where discourse analysis may be automated. All of these analysis tools have in common that they make it feasible to rigorously interpret a much larger and more complex dataset than was possible even a decade before. This allows evaluators to focus directly on the emergent experiences, actions, and conversations of learners in informal settings, rather than having to rely on the simpler but less ecologically valid proxies of reflective interviews and predesigned surveys.

We reiterate our belief that all of these advances make human-subjects protections more important than ever, particularly as the U.S. Department of Health and Human Services and fifteen other Federal Departments and

Agencies have issued revisions to the Federal Policy for the Protection of Human Subjects (the "Common Rule") (Federal policy for the protection of human subjects, 2017); and as educational organizations and IRBs try to keep track of changing cultural norms in terms of what can be recorded when and by whom, what are appropriate ways to recruit and interact with youth and adults, and what forms of information are inherently private.

A Broader, Stronger, Professional Community

For many decades, the field of informal STEM learning was unfamiliar to mainstream educational practice, research, and evaluation. With the public's tendency to conflate "education" with "schools," the informal learning world has been historically sidelined from serious discussions of educational policy and practice. Evaluators, researchers, and practitioners complained of feeling like "second-class citizens," needing to justify their work as valuable and relevant. Another challenge, highlighted in a landscape study (Falk, Randol, & Dierking, 2012), was that informal STEM educators seldom self-identified as such, preferring to think of themselves as active within a particular sector (e.g., "afterschool provider," "museum professional," "film creator"). Consultant evaluators were somewhat more likely to see themselves as informal STEM evaluators, because they often worked across multiple projects, but evaluators within organizations also tended to self-identify within their sector (e.g., "museum evaluator"). The field was fragmented and lacking some of the typical characteristics of an established field: a sense of a common identity, university-based research departments, peer-reviewed journals, and a commonly recognized set of core documents.

In the last decade, there has been an unparalleled blossoming of professional connections and communal resources. The last issue on "nonformal education" published in *New Directions for Evaluation* was in 2005 (Norland & Somers, 2005). Since then, several major reports (e.g., Friedman, 2008; NRC, 2009) and national projects (e.g., National Informal STEM Education Network [NISE Net], www.nisenet.org; Building Informal Science Education [BISE], http://www.visitorstudies.org/bise; CAISE, www.informalscience.org) have pushed us forward in connecting, synthesizing, and extending what we know about learning in informal settings.

The field of informal STEM evaluation has also become more community-oriented in recent years. Several sector-specific evaluation communities now exist to allow program directors and evaluators from within the sector to join forces to build evaluation capacity. By using multisite evaluation approaches and forming communities of practice, these groups have identified shared measures, trained a range of evaluators from new to seasoned, and used the experiences and results from these processes to promote understanding of both evaluation practice and the informal learning sector being studied. Importantly, these groups have been convened by seasoned evaluators who work with interested stakeholders from

within a specific sector to identify and implement evaluation methods that are needed by those on the ground (rather than being mandated or enforced from the top-down). Ongoing communities span a range of informal learning settings and programs, including science festivals (e.g., EvalFest, www.evalfest.org); science centers (e.g., the Collaboration for Ongoing Visitor Experience Studies [COVES], http://www.understandingvisitors.org); and after school programs (e.g., National Girls Collaborative Project, www.ngcproject.org). Another example is the NISE Net project featured in Chapter 5 (Bequette, Cardiel, Cohn, Kollmann, & Lawrenz, this issue), which outgrew its original mandate as a network of museums working on nanoscale science, technology, and engineering projects and became a flexible and reflective community working to create team-based inquiry. One intriguing question is whether such groups will ever fully coalesce, or whether they will be kept separate by deep and structural differences in their programs.

For those outside of these communities, opportunities exist to take advantage of the collective wisdom of the field. About a decade ago, the PEAR group at Harvard University created Assessment Tools in Informal Science, one of the first online repositories of evaluation instruments for informal learning contexts (http://www.pearweb.org/atis; The PEAR Institute: Partnerships in Education and Resilience, 2009–2017). This repository was a key step forward for evaluators in the field who were eager to explore tools and access information about their validity and reliability but simply did not know where to look. The repository was designed with filters (such as age of target audience and content domain) to help a user narrow the list of possibly relevant instruments. The description of each instrument includes psychometric evidence and user reviews to help evaluators make informed decisions about whether and how an instrument suits their needs. In the years that followed, various learning resource networks funded by the National Science Foundation (NSF), such as CAISE, the STEM Learning and Research Center (STELAR), and the Community for Advancing Discovery Research in Education (CADRE), followed suit to provide databases of instruments and/or reports that have been used to evaluate NSF Division of Research on Learning in Formal and Informal Settings projects. In some cases, this work has also been published as professional development materials on topics such as selecting and working with an evaluator, selecting appropriate tools and instruments, and conducting culturally responsive evaluation (Bonney, Ellenbogen, Goodyear, & Hellenga, 2011); see also http://www.informalscience.org/evaluation for a full list of resources.

CAISE warrants its own mention. CAISE is a 10-year investment by NSF that began as an online clearinghouse for ISE projects and their evaluations, and it has become an internationally recognized site for all things related to informal STEM education. In addition to a database of over 1,000 evaluations, the site includes more than 2,000 research publications, and a

collaboratively designed "Knowledge Base" that shares current evidence on informal STEM impacts on a range of audiences in a variety of settings. The site also offers ongoing forums on a variety of topics, most of which intersect with evaluation in some way, and some of which focus explicitly on evaluation. The CAISE community currently numbers well over 4,000 members from 50 countries. An ongoing challenge for the field will be the sustainability of such a community hub, given that informal STEM education is still comprised of actors that default to seeing themselves as belonging to different professions and subsectors.

There has been an expansion of journals embracing serious studies of learning in informal settings. *Visitor Studies* has grown from humble roots to become a high-quality, peer-reviewed journal. *Science Education* and the *Journal of the Learning Sciences* have dedicated sections for learning in informal settings. The *International Journal of Science Education* was developed to include Part B, a separate and quarterly publication that focuses specifically on communication and public engagement. The online journal *Connected Science Learning* is a new collaborative between the Association of Science-Technology Centers (ASTC) and the National Science Teachers Association (NSTA) that publishes studies at the intersection of formal and informal science learning for the betterment of both.

In terms of conferences and professional associations, there has been a significant maturation of the Visitor Studies Association, the U.S.-based but international association for those who study learning in informal settings, to include broader membership, working groups and subcommittees, and published professional competencies. The American Educational Research Association (AERA) has an active and growing special interest group focused on Informal Learning Environments Research (ILER-SIG), and the National Association of Research in Science Teaching (NARST) has, similarly, a strand dedicated to science learning in informal contexts. Research-practice partnerships have become far more common (e.g., Sobel & Jipson, 2016), contributing to the utility of studies conducted and the forums for sharing their findings.

The coming-of-age of informal STEM education would not have been possible without highly strategic investments by some key funders. NSF, for example, provided funds for various resource networks (CAISE, STELAR, and CADRE) that offer invaluable technical support and community resources for formal and informal STEM education evaluators, researchers, and practitioners. The Noyce Foundation (and its new embodiment as STEM Next) and the Charles Stewart Mott Foundation provided strategic support for development of the "Common Instrument Suite" for afterschool programs with STEM components, as well as the data-system based on it (Noam et al., 2017). The Wellcome Trust and the Gordon and Betty Moore Foundation invested funds in the Science Learning Plus program that pushed the boundaries of research-practice collaborations as well as international (U.S.–U.K.) projects.

Learning Ecosystems

A final theme that is rapidly gaining interest across the landscape of STEM education evaluation and research is the concept of "learning ecosystems" that include not only schools but also museums, libraries, institutions of higher education, businesses, and informal learning programs. The concept is a variation of the basic model of Bronfenbrenner (1977, 1986) that puts a child at the center of a series of concentric rings of influence, from the most immediate contacts (such as parents and siblings), through intermediate-level influences (such as schools, museums, and libraries), to the largest social context in which the child is ultimately embedded (religious norms, cultural practices, etc.) The ecosystem metaphor acknowledges the full complexity of a learning system and emphasizes the multidirectional influences among players.

The concept of "learning ecosystems" has been embraced by those who work in, or study, informal learning settings because it expands education beyond the purview of schools, allowing for many intersecting contributions to a child's development. Also, by putting the child at the center, the model fits naturally with concepts of learning as voluntary and self-directed. A recent report on effective out-of-school programs (NRC, 2015) uses this framing, and it concludes by advising policy-makers to acknowledge and support the full range of contributors to a child's learning ecosystem. Of course, while the focus of most studies is learning by children, the model can be applied to individuals of any age, given the foundational assumption of lifelong learning that characterizes informal STEM education.

The development of STEM learning ecosystems has also been selected as a strategic investment opportunity by the STEM Funders' Network, a discreet but powerful coalition of over twenty private foundations with interests in supporting STEM education (http://stemecosystems.org/). To date, sixty-eight communities have been supported as they strengthen their existing STEM learning ecosystem by establishing a more connected network of learning opportunities in- and out-of-school. These communities have been invited to participate in a community of practice and share lessons learned as they attempt to strengthen the links among the educational organizations in their regions, cities, or states.

The implications of an ecosystem perspective for evaluation are actually quite radical. Historically, despite the foundational logic model of informal STEM learning as being lots of small experiences that accumulate over a lifetime, each individual object or program was funded and evaluated on its own. An ecosystem perspective calls for the development of methods that focus on individual learning trajectories that may be hard to predict, let alone track and assess. Short of studying individual learners longitudinally as they move across settings throughout their daily lives (a valuable but extremely expensive option), how will we measure what they learn? Can evaluators design methods that characterize the impact of a STEM resource

or program, not just as a stand-alone offering, but in terms of its capacity to support and connect to other experiences, resources, and programs in the ecosystem? Can evaluators of different initiatives find ways to assess collective impacts by pooling their resources? A lot of new work will be needed here. At present, even tracking an individual's depth of experience with any one STEM resource is difficult, let alone cumulative impacts from different aspects of different experiences in different settings.

Another implication of the ecosystem perspective is that culturally responsive evaluation will be critical. This is partly because the learning outcomes result not just from any one intervention but from the paths, interactions, values, and reflections of the learner in their larger community. Understanding the system will take many perspectives, including those of the educators in various organizations (teachers, afterschool providers, librarians, etc.) and, perhaps most importantly, the perspective of the learner. Trust and mutual understanding will be critical to an evaluator who needs to track and make sense of the personal and idiosyncratic learning trajectory of an individual within a larger community structure. In Chapter 4, Garibay and Teasdale (this issue) make the important point that cultural responsiveness is about more than respect and ethics; it is an issue of validity. That is particularly true of evaluation within a learning ecosystem.

Finally, we note that communications technology has created an unexpected side effect in relation to evaluation within a learning ecosystem. The unprecedented rise of mobile technologies (laptops, cellular phones, tablets, etc.) means that it is no longer enough to specify the nature of a designed STEM resource by its original medium (film, radio, museum exhibit, live program, etc.). Instead, a learner's experience is some confounding of the original design with the medium in which they actually experience it (recorded, live-streamed, asynchronous, online, podcasted, YouTubed, tweeted, etc.). This distinction was identified by the designers of the Online Project Monitoring System (OPMS), whose mandate was to meticulously characterize and quantify the learning outcomes reported by all NSF-funded informal STEM education projects over a seven-year period (Silverstein & Goodyear, personal communication, 2017). Their typology breakthrough came when they realized they had to ask awardees to describe not only the "deliverable" they had created (video segment, audio segment, theater production, newsletter, etc.) but the "delivery methods" for each one (CD-ROM, project website, non-project website, wiki site, blog, etc.) In short, we believe *the increasing use of mobile and cross-platform technologies reflects a manifestation of the STEM ecosystem model, as learners move from place to place and experience to experience.* While this explosion has almost certainly advanced informal STEM education by increasing access to learning resources as well as the power to comment on and repurpose them, it also makes evaluation more complex and ethical issues more central.

Ironically, from being a fringe area of STEM education, informal learning is becoming one of the most aligned with the cutting edge of learning in

the twenty-first century. Its familiar slogans of "supporting lifelong learning" and "anywhere, anytime, anyone" have become increasingly relevant metaphors in an era based on mobile media and gradual blurring of the very concept of a learning "setting." A recent NRC report (2016) argued that the very definition of science literacy needs to be updated to reflect this shift, and informal STEM education evaluators will be on the frontlines in that effort.

Conclusion

Finding creative and meaningful ways to engage public audiences is a hallmark of informal STEM education. Educators often use informal contexts to move beyond traditional learning methods, engaging the public with STEM through the creation of innovative, memorable, and meaningful experiences. Likewise, authentic evaluation of these experiences must also go beyond traditional approaches and innovate methods that ensure the data collected are meaningful to participants, informal educators, and other stakeholders. We hope the field will strive to protect its creative and nimble identity, leveraging shared measures when it makes sense to do so, while at the same time allowing the goals of informal learning programs to dictate the measures selected, rather than the other way around. Rigor and ecological validity can both be achieved, especially by leveraging some of the less obtrusive data collection and analysis tools now available.

Though the examples in this and other chapters are rooted in informal learning contexts, we believe that the larger issues have broad appeal and application for the broader evaluation field. The discussions of methodological issues and innovations, challenges in articulating evaluation outcomes, strengthening of professional communities, and learning in the twenty-first century have the potential to inform the future of evaluation, not only for those of us who are lucky enough to work in informal learning contexts but for other evaluation communities as well.

The past decade has opened up new avenues for the field to explore. This work is messy. It is full of challenges and pitfalls but also full of potential. We are privileged to have the chance to work in a context that is equally committed to authenticity and striving toward rigor.

References

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

Allen, S., Gutwill, J., Perry, D. L., Garibay, C., Ellenbogen, K. M., Heimlich, J. E., … Klein, C. (2007). Research in museums: Coping with complexity. In J. H. Falk, L. D. Dierking, & S. Foutz (Eds.), *In principle, in practice: Museums as learning institutions* (pp. 229–245). Lanham, MD: AltaMira Press.

Bonney, R., Ellenbogen, K., Goodyear, L., & Hellenga, R. (Eds.). (2011). *Principal Investigator's guide: Managing evaluation in informal STEM education*

projects. Washington, DC: Center for Advancement of Informal Science Education & Association of Science-Technology Centers. Retrieved from http://www.informalscience.org/sites/default/files/caisevsapi_guide.pdf

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513–531.

Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, *22*(6), 723–742.

Center for Advancement of Informal Science Education. (2017). *What is informal science?* Retrieved from http://www.informalscience.org/what-informal-science

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980–1008.

Cornell Lab of Ornithology. (2014). *Evaluation instruments*. Ithaca, NY: Cornell University. Retrieved from http://www.birds.cornell.edu/citscitoolkit/evaluation/instruments

Dorph, R., Cannady, M. A., & Schunn, C. D. (2016). How science learning activation enables success for youth in science learning experiences. *Electronic Journal of Science Education*, *20*(8), 49–85.

Ellenbogen, K. M. (2002). Museums in family life: An ethnographic case study. In G. Leinhardt, K. Crowley, & K. Knutson (Eds.), *Learning conversations in museums* (pp. 92–112). New York: Routledge.

Falk, J. H., Randol, S., & Dierking, L. D. (2012). Mapping the informal science education landscape: An exploratory study. *Public Understanding of Science*, *21*(7), 865–874.

Federal policy for the protection of human subjects (2017, January 19). *2018 Requirements, 82. Fed. Reg. 7149 (final rule)*. Retrieved from https://www.gpo.gov/fdsys/pkg/FR-2018-06-19/pdf/2018-13187.pdf

Feinstein, N. W. (2011). Salvaging science literacy. *Science Education*, *95*(1), 168–185.

Friedman, A. J. (Ed.). (2008). *Framework for evaluating impacts of informal science education projects*. Arlington, VA: National Science Foundation.

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.

Hooper-Greenhill, E. (2004). Measuring learning outcomes in museums, archives and libraries: The Learning Impact Research Project (LIRP). *International Journal of Heritage Studies*, *10*(2), 151–174.

Hsi, S., & Fait, H. (2005). RFID enhances visitors' museum experience at the Exploratorium. *Communications of the ACM*, *48*(9), 60–65.

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*. (A study prepared for the Noyce Foundation). Program in Education Afterschool & Resiliency. Harvard University and McLean Hospital. Retrieved from http://www.nc4il.org/images/stem-in-libraries/evaluation/Toward-Systematic-EvidenceBase-Science.pdf

Kanda, T., Shiomi, M., Perrin, L., Nomura, T., Ishiguro, H., & Hagita, N. (2007). Analysis of people trajectories with ubiquitous sensors in a science museum. *Proceedings of the 2007 IEEE International Conference on Robotics and Automation* (pp. 4846–4853). Roma, Italy: IEEE.

Krishnamurthi, A., Ballard, M., & Noam, G. G. (2014). *Examining the impact of afterschool STEM programs*. (A paper commissioned by the Noyce Foundation). Afterschool Alliance. Retrieved from http://www.afterschoolalliance.org/ExaminingtheImpactofAfterschoolSTEMPrograms.pdf

LIFE Center. (2005). *The LIFE Center's lifelong and lifewide diagram*. Retrieved from http://life-slc.org/about/citationdetails.html

Luczak-Roesch, M., Tinati, R., Simperl, E., Van Kleek, M., Shadbolt, N., & Simpson, R. J. (2014, June). Why won't aliens talk to us? Content and community dynamics

in online citizen science. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp. 315–324). Ann Arbor, MI: ICWSM.
Ma, J. (2016, July). *Using sequence analysis to understand visitor behavior*. Paper presented at the 29th annual meeting of the Visitor Studies Association, Boston, MA.
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.
National Research Council. (2016). *Science literacy: Concepts, contexts, and consequences*. Washington, DC: The National Academies Press.
Noam, G. G., Allen, P. J., Shah, A. M., & Triggs, B. (2017). Innovative use of data as game changer for OST programs. In H. J. Malone & T. Donahue (Eds.), *The growing out-of-school time field: Past, present, and future* (pp. 161–176). Charlotte, NC: Information Age Publishing.
Norland, E., & Somers, C. (Eds.). (2005). *Evaluating Nonformal Education Programs and Settings: New Directions for Evaluation*, *108*, 1–83.
Owen, V. E. (2014). *Capturing in-game learner trajectories with ADAGE (assessment data aggregator for game environments): A cross-method analysis* (Doctoral dissertation). University of Wisconsin-Madison, Madison, WI.
Peterman, K., Robertson Evia, J., Cloyd, E., & Besley, J. (in press). Assessing public engagement outcomes by the use of an outcome expectations scale for scientists. *Science Communication*.
Robertson Evia, J., Peterman, K., Cloyd, E., & Besley, J. (2017). Validating a scale that measures scientists' self-efficacy for public engagement with science. *International Journal of Science Education, Part B*, *8*(1), 40–52.
Rowe, S. (2012, August). *Cyberlab: Data collection for large-scale, long-term multimodal analyses*. Paper presented at the 6th International Conference on Multimodality, London, UK.
Serrell, B. (1998). *Paying attention: Visitors and museum exhibitions*. Washington, DC: American Association of Museums.
Sobel, D. M., & Jipson, J. (Eds.). (2016). *Cognitive development in museum settings: Relating research to practice*. New York: Psychology Press.
The PEAR Institute: Partnerships in Education and Resilience. (2009). *Assessment Tools in Informal Science (ATIS)*. Retrieved from http://www.pearweb.org/atis
U.S. Department of Education. (2007). *Report of the Academic Competitiveness Council*. Washington, DC: Author. Retrieved from https://eric.ed.gov/?id=ED496649

SUE ALLEN, PhD, is the director of Allen & Associates, a research and evaluation consulting firm focused on informal STEM education, and a senior research scientist at the Maine Mathematics and Science Alliance.

KAREN PETERMAN, PhD, 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.