Evaluation Under Pressure:
Balancing the Needs of the ISE Field with the Needs of Individual Projects

Kirsten Ellenbogen & Amy Grack Nelson, Science Museum of Minnesota

This paper builds upon our work with Karen Knutson and Kevin Crowley (University of Pittsburgh Center for Learning in Out of School Environments) on the Building Informal Science Education (BISE) project, which is supported by the DRL program of the National Science Foundation under the award number 1010924.

Great progress has been made in building the practice of evaluation in informal learning environment programs that support science, technology, engineering, and math (STEM) education. Stakeholders from federal agencies, private foundations, corporations, and community-based organizations have increasingly structured their funding guidelines to require evaluation of the impact of a project. Largely stimulated by this investment, there has been significant growth in the size and professionalism of the field of informal STEM education (ISE) evaluation. Increased resources have improved the frequency of evaluation, provided better access to evaluation results, and created some shared reference points for language and methods.

With increased resources and professionalism come increased expectations for the evaluation of individual projects to also provide field-wide evidence for the impact of ISE. But there has not been a significant growth in the use of psychometric measures that are validated to support field-wide assessments and norm-referenced tests for the field. This essay argues that expectations for the field to adopt normed assessments need to be contextualized in the practice and realities of ISE evaluation. While at the same time, evaluators of ISE projects need to build upon stepping stones such as metadata to build toward the use of shared measures that are adaptable to specific projects and sensitive to the constraints of conducting evaluation in informal STEM learning environments.

The Practice of ISE Evaluation

At the National Science Foundation (NSF), there is a persistent stream of funding for evaluation, driven by the requirement that each ISE project must document whether it meets its goals. Increasingly corporations, private foundations, and other government agencies are funding ISE and likewise, requiring evaluation of the project impacts. The growth of ISE evaluation can be documented in numerous ways. In recent years there have been an increasing number of Ph.D. and M.A. programs in evaluation, education, museums studies, and other fields that offer training related to measuring the impact of informal learning experiences. There has been a corresponding increase in tools for developing evaluation approaches and documenting outcomes. For example, evaluators now have a set of categories that define ISE learning experiences (Bell, Lewenstein, Shouse, & Feder, 2009) and a common set of categories to describe the wide range of ISE project outcomes (Friedman, et al., 2008). There are growing numbers of online databases of assessment tools that measure ISE projects (Hussar, Schwartz, Boiselle, & Noam, 2008), and evaluation reports (Crowley, Leinhardt, & Chang, 2001).

Recently, the Building Informal Science Education Project (BISE) has begun an analysis of the evaluation reports on www.informalscience.org, an online repository of evaluation and research reports designed to support knowledge sharing among professionals in ISE. A critical goal of the BISE project is to synthesize the diversity of evaluation reports posted to the site. To date, the project has produced an extensive coding framework and conducted an initial analysis of over 400 evaluation reports.

This essay does not go into depth about the BISE project findings, but the initial broad analysis indicates some consistency with earlier studies that looked across a range of ISE evaluation reports (Brody, Bangert, Dillon, 2007; Institute for Learning Innovation, 2007). For example, the evaluation data tends to be correlational rather than cause and effect. BISE also found that the 427 evaluation reports describe 19 different data collection methods. Interviews (64%) and surveys (61%) were by far most frequently used method, while observation (30%), timing and tracking (17%)

and focus groups (16%) were less common. An examination of the kinds of analysis used shows that 51% of the reports include no statistical analysis. Of the reports that included statistical analysis, the most frequently used tests were chi-square (40%), t-tests (29%), and ANOVA (20%). Fully 34% of the reports that included a mention of statistical significance did not list which test was performed.

These descriptions are useful for understanding the practice of the field, but the latest phase of the BISE project is a more in-depth analysis of evaluation reports, such as findings across media projects. There is also an analysis of common components of the ISE evaluation reports that will result in a set of guidelines for writing reports that are useful for the visitor studies field, as well as engaging to stakeholders.

These cross-project examinations of evaluation reports give us a cumulative picture of study elements that are associated with ISE projects. This is an exciting advance of the field that allows us to present more compelling evidence about the impact of ISE projects. Despite these advances, the field is still hesitant to adopt more standard approaches to measuring and reporting that will allow evaluators to more effectively isolate variables and identify which variables lead to successful learning experiences for different audiences. Stepping back to take a broader view of ISE evaluation reveals some characteristics of the field that conflict with recommendations to create more standard measures.

One Effort, Many Disciplines

Any examination of how we measure the impact of ISE must acknowledge that there are many different disciplines involved in this effort. Fifteen years ago, deep knowledge about measuring the impact of ISE was restricted mostly to the relatively small group of evaluators and researchers who worked on projects funded by the NSF ISE program. This core has now become part of a much larger community with more diverse intellectual interests and professional training.

The community is truly interdisciplinary, bringing expertise from education, evaluation, psychology, anthropology, design, human-computer interaction, organizational theory, and the learning sciences, among other disciplines. Interdisciplinary fields have a particular challenge in finding common ground, and ISE is no exception. In fact, the rapid growth of the field compounds this problem, with more scholars "immigrating" into the field each year on top of the training programs that produce an increasing number of ISE "natives". The many disciplines within ISE evaluation do overlap in some practice, methods, and literature. Despite these similarities, great friction is caused when ISE evaluators, who come from a wide range of these disciplines, are held to the expectation of common professional practices. The friction has grown with the increased attention to measuring the impact of ISE and the increased diversity of those who measure its impact.

A professional from a psychology background may be more likely to take a standardized assessment approach to measuring the impact of ISE than someone with a degree in evaluation who is likely to use assessment tools only if the project is designed to result in changes in certain knowledge or attitudes and if using those measures does not inappropriately disrupt the normal participant experience.

Most, if not all, of the disciplines that are reflected in ISE evaluation are ones that have ethics and norms in their practice. But these norms and even the ethics vary. Professionals trained in evaluation, for example, come to the practice with a strong emphasis on meeting the needs of their stakeholders, particularly the project lead. This has many implications in the design and documentation of ISE evaluation. For example, an evaluator may need to use methods that were used previously at that institution so that the staff has a way to make internal comparisons. Or the project lead may prefer monthly updates on evaluation findings with a power point presentation as the final deliverable for the project. It can be a struggle to balance the evaluation needs of the project stakeholders with the need for the ISE field to broadly document its impact.

Varying Needs of ISE Evaluation Stakeholders

Evaluators are acknowledging the need for more valid and reliable indicators of learning in ISE projects. Yet efforts to use assessment tools that provide standard measures across projects have been limited by wide range of potential outcomes and the unique environments. Evaluators must be able to address the unique characteristics and constraints of ISE experiences in their measurements.

Professionals who develop ISE experiences need more nuanced information to help them do their work most effectively. They may be far more interested in formative evaluation that improves their project than the summative evaluation that proves its impact. At the same time that funders and other ISE stakeholders are calling for more standardized evaluation that uses the same measures across projects, social media and ISE conferences are exploding with an organized effort to decrease the role of evaluation in the development of ISE projects. These ISE professionals come from a long history of testing their own work and balancing testing with a highly creative development process. Some ISE professionals go so far as to say that evaluation is singlehandedly damaging the creative output of the field. These contentions are meant to stimulate conversations that result in changed practices and are balanced by many ISE professionals that acclaim the positive impact of evaluation in the development and documentation of an ISE experience. But these tensions indicate that evaluators may not be the only ISE professionals hesitant to adopt more standard measures. The field cannot evolve to meet this expectation if a project lead will not pay an evaluator to use standard measures.

Another stakeholder is ISE organizational leaders who need strong evidence to justify their institutions' work or perhaps the existence of the organization. In an environment of shrinking resources and calls for greater accountability, leadership often finds itself needing to justify the existence of ISE experiences such as ISE science related television programs or natural history museums. Organizational leaders need a foundation of research that goes beyond professional beliefs, personal experience, untested hypotheses, or studies that describe the learning experience but do not provide evidence that can be generalized to the field (Koster, 1999). These ISE professionals should largely be eager to include standard measures in ISE evaluation in an effort to build capacity for the field to provide strong evidence and make findings relevant to funders, agencies, board members, and other stakeholders who seek “hard evidence” of success.

Recommendations

These four recommendations are not offered as a to do list, but rather as a starting point for a thorough examination of how ISE evaluation might begin to integrate more standard measures. A comprehensive study will identify the full range of current practices of ISE evaluation at the same time as identifying a bridge toward the use of standard measures across projects. This approach will ensure that the creation of a new vision for ISE evaluation happens at the same time as the identification and removal of obstacles.

1. Create a Shared Message of Urgency

The ISE field needs to craft a more effective message of urgency about the need to change the practice of evaluation. Many funders, policymakers, and organizational leaders are convinced of the need to improve our ability to demonstrate the impact of the field through more standard measures. But an individual evaluator may face a greater urgency of meeting the contractual need to improve and prove the impact of a single program within the highly complex constraints of an ISE environment.

The field needs more practice-oriented messages of urgency. The messages need to respect the profession: evaluation is not the same as assessment. The characteristics that make ISE unique are also the characteristics that make ISE evaluation especially complex. Efforts to craft messages of urgency and advance the use of standard measures must include members who can represent the realities of ISE evaluation. This should include full coordination with the professional associations that represent the ISE evaluation community, including the Visitor Studies Association, the Committee of Audience Research and Evaluation, the arts and culture TIG of the American Evaluation Association (as well as the association itself), the informal strand of the National Association for Research in Science Teaching, the informal SIG of the American Education Research Association, other professional associations that represent ISE sectors, such as those that represent afterschool professionals and the newly forming professional association for ISE media professionals. As these professional associations come together to discuss messaging for their members and expectations for changes in practice, the language of this effort to change evaluation will become much more relevant and valuable to the different disciplines and contexts of ISE evaluation.

2. It's About Reporting, Not Just Measurement

The 2012 GAO report on STEM education found not just issues with methods, but also that evaluation results were not always disseminated effectively to document findings and share knowledge. In the ISE field, there is little motivation to improve the quality and reach of evaluation reports. Most professionals conducting evaluation do not receive promotion or tenure based on how widely the findings are shared. In the BISE project, we can document that evaluation methods are evolving, yet evaluation reports remain relatively unchanged. So a widespread effort to include more standard measures in ISE evaluation has to include support for improving evaluation reporting. The primary goal of an evaluation report is to address the questions and needs to the project team. Evaluators often have to write a report that balances the interests and needs of not just the project team but also the project funder, who may be interested in broader questions about the community value of the project. If an evaluation report from an individual project is going to include findings from a standard measure that can be compared to other projects, that is an additional set of requirements for the evaluator to juggle.

Guidelines such as Western Michigan University's Evaluation Report Checklist (Miron, 2004) and the forthcoming ISE Evaluation Report Guidelines from the BISE project will help ISE evaluators juggle the increasing number of stakeholders who will want to use their reports. But there are other supports that will ensure that changes in the practice of evaluation reporting happen more effectively. Some of the change must be supported as a cultural shift in expectations about transparency.

The ISE field has a long history of evaluation professionals who work as independent consultants. The consultant conducting the evaluation typically owns the instrument or other project-specific methods used in the evaluation. ISE evaluation consultants are in general altruistic and unflagging in their efforts to advance the field, including support for professional associations, publication of results, and even efforts to analyze findings across projects. But there is no expectation that all of this support for the field includes sharing instruments. This cultural norm is not unique to ISE evaluation; many evaluators who work as independent consultants do not expect to share the instruments they used in their evaluation. Any effort to integrate more standard measures into ISE evaluation will need to respect and respond to the culture of the field and build an effective case for transparency.

3. Start With Metadata And Databases

The effort to make change in the practice of ISE evaluation needs to happen on multiple fronts. Any effort to have ISE evaluation integrate more standard measures into practice should include a complementary effort to establish common metadata in the field. Particularly in a digital age, a common set of metadata for ISE evaluation can stimulate significant advances in a field. The Center for Advancement in Informal Science Education (CAISE) facilitates the Infrastructure Coordination Roundtable, a collaboration of ISE projects that are creating large-scale digital resources for the field. Participants in the Roundtable have agreed upon a set of metadata to organize their technical systems. They are using this structure to dynamically link their ISE web sites through the Informal Commons search website (http://informalcommons.org) providing one integrated access point to all of their ISE resources. A similar effort focused on creating a set of metadata for ISE evaluation would allow the field to make more comparisons across findings.

Metadata would allow synthetic explorations of ISE evaluations beyond the life of any one project. The GAO report on STEM education shares that program officials were not consistently able to report outcome measures such as number of participants for their funded projects. A set of metadata agreed upon by the ISE evaluation field would be a significant step toward addressing not just outputs like participant numbers, but also shared understandings, and measures of outcomes like skills. Metadata would create common language to document impact for an ISE project that has for example, participants that range from age 7 to 70, without trying to reduce the complexity of the audience or their learning experiences. Metadata would be a proving ground for understanding how measures could be developed that are shared across projects while still being attentive to the specific context and needs of a project.

Field-wide metadata is the first step in creating shared databases (Schneider, 2004) that could build ISE evaluation into a community that is driven by and attentive to strong field-wide evidence. There are few examples of metastudies that allow us to compare outcomes across projects. Metadata and common databases for the field would transform our ability to build evaluation evidence that transcends individual projects.

4. Focus on Validated Measures of Interest and Identity

Metadata is one way to make changes in ISE evaluation that will build field-wide evidence, and it is a strategic approach that could occur alongside efforts to use more standard measures. Efforts to integrate more standard measures into ISE evaluation have been strategically framed to make change more likely. For example, there have already been strong recommendations (Hussar, Schwartz, Boiselle, and Noam, 2008) to agree upon a small number of assessment questions that would be used across the entire ISE field, rather than changing all measurements at once. This strategy would allow for some comparisons, even though these measurements would not have the psychometric qualities of a cohesive instrument.

Selecting a small number of assessment questions to be used across the field could support the development of more coordinated data collection across projects, while still accounting for project specific goals and outcomes. Leaders in the effort to get ISE evaluation to adopt more standardized measures need to be wiling to emphasize that not all outcomes need to be measured with a common set of instruments.

If ISE evaluation is going to hone in using standard measures for a small number of outcomes, the change should happen around outcomes that are highly significant to ISE. Interest and identity are two outcomes that may not be unique to informal learning environments, but they are considered hallmark outcomes of ISE experiences (Bell, Lewenstein, Shouse, & Feder, 2009). At the same time, interest and to a greater extent, identity are considered difficult to measure. There is a lot to be gained if evaluators can show reliable and compelling evidence of interest and identity development across ISE projects.

References

Bell, P., Lewenstein, B., Shouse, A. W., & Feder M. A. (Eds.). (2009). *Learning science in informal environments: People, places, and pursuits.* Washington DC: The National Academies Press.

Brody, M., Bangert, A., & Dillon, J. (2007). *Assessing Science Learning in Informal Settings*. Commissioned paper for the National Research Council. [On-line]. (Available at: http://www7.nationalacademies.org/bose/Learning_Science_in_Informal_Environments_Commissioned_Papers.html)

Crowley, K., Leinhardt, G., & Chang, C.F. (2001). Emerging research communities and the World Wide Web: Analysis of a Web-based resource for the field of museum learning. Computers and Education, 36 (1), 1-14.

Friedman, A. (Ed.). (March 12, 2008). *Framework for Evaluating Impacts of Informal Science Education Projects* [On-line]. (Available at: http://insci.org/resources/Eval_Framework.pdf)

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.* Boston, MA: Program in Education, Afterschool & Resiliency.

Institute for Learning Innovation (2007). *Evaluation of Learning in Informal Learning Environments.* Commissioned paper for the National Research Council. [On-line]. (Available: http://www7.nationalacademies.org/bose/Learning_Science_in_Informal_Environments_Commissioned_Papers.html)

Koster, E. H. (1999). In search of relevance: Science centers as innovators in the evolution of museums. *Daedalus, Journal of the American Academy of Arts and Sciences 128*(3): 277– 296.

Miron, G. (2004). *Evaluation report checklist.* [On-line]. (Available at: http://www.wmich.edu/evalctr/archive_checklists/reports.xls

Schneider, B. (2004.) Building a scientific community: The need for replication. *Teachers College Record, 106* (7), p. 1471-1483.