Developing a Set of Guidelines for Rigorous Evaluations at a Natural History Museum

Anna MacPherson
American Museum of Natural History, New York City, NY, United States

Karen Hammerness
American Museum of Natural History, New York City, NY, United States

Preeti Gupta
American Museum of Natural History, New York City, NY, United States

Corresponding author: Anna MacPherson, amacpherson@amnh.org
212-769-5261
American Museum of Natural History
Central Park West and 79th St.
New York, NY
10024

Accepted for publication in the *Journal of Museum Education*

Tools, frameworks and case studies

Developing a Set of Guidelines for Rigorous Evaluations at a Natural History Museum

Anna MacPherson, Karen Hammerness and Preeti Gupta

Evaluation is often a required component for funded projects; however, the process of undertaking an evaluation does not always result in valuable information that can be used for learning, reflection, and program improvement. In light of methodological shortcomings and lack of meaningful use of evaluations, informal science education researchers have called for greater rigor in evaluations. To strengthen our own approach, we engaged in a process to review past evaluations and to determine a means to improve our work in this arena. A group of stakeholders first conducted a survey of our own evaluations at the museum. We identified features of the reports that were meaningful and led to useful insights about programs. We coded these guidelines in categories that emerged. Then, we conducted a modified Delphi process for reaching consensus about which guidelines for evaluation should be used across the museum to guide future work. This paper presents our final set of guidelines, a rubric for rating proposals, and implications for educators, researchers and evaluators in informal science learning spaces.

Key words: evaluation, natural history museum, Delphi study

The importance of high quality evaluations

At their best, evaluations of education programs at informal learning institutions are used to provide guidance and support for decision-making. At worst, however, evaluations are conducted in compliance to grant requirements and then filed away in a folder, never to be used for meaningful change. How can we meet the challenge of using evaluations well? How can an institution develop a process for feedback and learning that can lead to evidence to guide decision-making? In an era of expanding interest in using large datasets and increasingly sophisticated methods for gathering and organizing such data, how do informal learning institutions leverage this power to maximize reach and learning? This paper contends that using a standardized and rigorous set of guidelines for working with internal and external evaluators may help informal learning institutions maximize this process for learning and decision-making.

Evaluations are conducted in informal science education for multiple purposes. A *front-end evaluation* may be conducted to understand an intended audience prior to developing an intervention or project. A *formative evaluation* may be conducted to learn about areas for improvement. Finally, *summative evaluations* are conducted once the project is established or completed in order to assess its success at meeting goals.[1]Evaluations can help programs by gathering empirical data to reveal strong practices and areas for improvement and help determine next steps in relationship to institutional missions and goals. In addition to front-end evaluation to determine needs, formative evaluation for course corrections, and summative evaluations to measure impact, our institution is interested in using evaluation findings to contribute to our institution-wide research agenda. We hope that strong and consistent evaluation—using rigorous methods and producing empirical data—will position us to transform evaluation into research.

At the American Museum of Natural History in New York City (AMNH) we often wish to document the experiences of visitors to our institution and participants in our programs. We document and measure experiences through formal and informal evaluations. Like many institutions, while we are building our own internal capacity to conduct evaluations and learn from our work, we cannot conduct all evaluations ourselves. In many cases, it is useful and important to have outside perspectives and expertise. We frequently seek independent evaluators to conduct evaluations of our galleries, exhibitions and programs. Programs at the AMNH most frequently require both formative and summative evaluation. Formative evaluations are used to determine what is currently working and not working and how to make changes. We often require summative reports for stakeholders and funders. We also require an occasional front-end evaluation. A recent example is a “needs assessment” for elementary science instruction in New York City for the design of an elementary school expansion of an existing middle school program.

Yet evaluations in informal science education, in general, need strengthening. There have recently been multiple calls for improved rigor and capacity in informal science education evaluation.[2] The calls for stronger evaluations seem to follow, in parallel, a push for stronger research on visitor learning and the impact of educational programs offered at informal science institutions. Evaluation and research, though related, are distinct in purpose. Evaluations are intended to gather information that is of immediate value for exhibit or program improvement. Research, on the other hand, is driven by a desire to understand the nature of learning. There may not be an immediate need for the findings and the knowledge generated is intended to be more generalizable than evaluation findings. An example of a study highlighting the need for more rigorous research was a recent study of informal science institutions in the United Kingdom. This work revealed that they do not always draw upon the research base for learning; the authors of this study recommended that such institutions base their decisions more on evidence from research studies.[3] In addition, several recent efforts in the United States have focused on identifying challenges and bringing together scholars to generate solutions to persistent questions about informal science education, such as how people learn in informal spaces, how to improve equity and access, and how to connect learning in formal and informal education. For example, the National Science Foundation (NSF)-funded ITEST (Innovative Technology Experiences for Student and Teachers) convening (2010), the National Summit on Assessment of Informal and Afterschool Science Learning (2012), and the CAISE Convening on Evaluation in Informal Science Education (2013) have been efforts in the last decade to focus attention on informal learning, and their recommendations have included strengthening research and evaluation. Recently, the Gordon and Betty Moore Foundation, responding to this call, funded work by an independent consulting firm (SK partners) on evaluations in informal science education. [4] In short, there is current consensus that evaluation of informal science education is valuable but that current efforts do not yet represent state-of-the-art methods or approaches.

In relationship to the current focus on building capacity to conduct high-quality evaluations of informal science experiences, we aimed to generate a set of shared values and guidelines for evaluations conducted specifically at the AMNH. With the intention to do this work across programs and areas in the museum, we developed a plan we could undertake as a community. We identified a set of steps that would help us ground our efforts in the strengths of past work, to clarify our own values and determine where we should focus our attention.

We determined that our approach should include: (1) reviewing the literature on rigorous evaluations; (2) reviewing a set of evaluations conducted in AMNH exhibits and programs in the recent past and determining their strengths and weaknesses; (3) engaging in structured discussion around shared goals for evaluation, and, ultimately (5) outlining our values and guidelines for evaluations of student, teacher and visitor learning at the AMNH. With the purpose of developing a useable set of guidelines for evaluations of museum education programs, we defined the following research questions:

1. Which features of evaluations matter most to our institution?
2. How can we design a set of guidelines to be used across the museum to communicate our values about evaluation?

Materials and methods

Participants

To help build understanding and a common language for evaluation, we designed a process that would allow considerable input and shaping by our colleagues *across programs and departments*. Museum education should not be confined to special programs put on by the Education department. A report from the Commission on Museums for a New Century[5] (1984) calls on all museum departments to "examine their potential as learning environments and to articulate their contribution in the realm of object-based, informal, voluntary, and lifelong learning"NUMBER. With this view in mind, we sought to maximize expertise and represent a broad range of stakeholders including exhibition curators, researchers, educators, and institutional leadership. The best way to involve this range of expertise was to situate this process within our *Educational Research and Evaluation Group*, which is composed of staff members who represent diverse areas of work and manage teams with a scope of work that involves research and evaluation. We initiated this group to build community and to develop shared understanding about research in the field and how it relates to our work. Eighteen participated in discussions of existing evaluation reports connected with work at the museum and a modified Delphi process (a structured communication protocol) to reach agreement about recommendations for future evaluations.

Reviewing examples of completed evaluation reports

As a first step, to help us identify features of strong evaluations and to ground our discussion in actual reports we had pursued ourselves, the *Research and Evaluation Group* spent several months reading and discussing a series of evaluation reports. Program directors were asked to choose evaluation reports that were relatively recent, and that they felt were well designed and useful. Program directors offered a short overview of the program, exhibit or experience that had been evaluated and described the process of selecting and working with the evaluator. The evaluations used for this exercise are described in Table 1.

[insert Table 1 here]

Members of the *Research and Evaluation Group* read each evaluation report. They discussed their major observations in small groups and we recorded them in a shared document. We segmented that large shared document into a series of "thought units" and developed a coding scheme based on themes that emerged. [6] Next, members of the group (again working in small groups) coded each thought unit; through this process, the list of thought units was culled by collapsing similar ideas together. The group generated a series of "recommendations for evaluations" from the ideas recorded in the original document. The group sorted the long list of recommendations into a set of six categories. From this discussion, we collapsed the original set of six categories into four main categories: "planning an evaluation," "defining the purpose and adequately framing an evaluation," "maintaining methodological rigor," and "maximizing the value of an evaluation." The process of discussion, coding and collapsing categories generated a focused set of ideas that in turn, prepared us to generate a set of agreed-upon recommendations for rank ordering.

Consensus building

We decided that to have the most useful set of guidelines, we would need to come to consensus about which guidelines were *most* important to us. We began the next phase of the work by rephrasing each "thought unit" as a recommendation for evaluation work at the museum. To enable us to continue to engage in a collaborative process, we sought out a means for honing down our recommendations while still using some kind of consensus model. After reviewing several options, we decided to use Delphi panel methodology, in which a group of experts on a subject ranks ideas in a set as important or not important and provides written justification. This ranking process is repeated until agreement about the most important ideas is reached. [7] We chose this method because it would be most likely to allow us to leverage our collective expertise and draw out multiple perspectives from our various stakeholders while minimizing the bias of influential individuals that can occur from in-person panels. This process has been used in science education in the past [8] and precedent suggests that little new information is gained from Delphi panels that exceed thirty participants. This panel involved 18 participants.

We designed the Delphi Survey in Qualtrics and distributed anonymous links to all members of the group. The survey listed all categories and all guidelines and participants were asked to rate, on a scale of 1-5, from "Not at all important" to "Very Important," and to respond to the following: "In your opinion, how important is this to the AMNH approach to evaluation?" In addition, we asked participants to provide an open-text "justification" for their rating.

A cut score of 4 ("Important" or "Very Important") was used to reduce the list of guidelines. In Round 2 of the process, participants were shown other group members' anonymous "justifications" for rating guidelines from Round 1, and again asked to rate the guidelines on a scale of 1-5. We used a cut score of 4 to generate the final set of guidelines.

Results

Common themes in our top-ranked guidelines

An average rating above 4.0 meant that, on average, panelists felt a recommendation was between "important" and "very important" for an evaluation conducted at AMNH. Thus, guidelines with a rating >4.00 were retained following the first round of the modified Delphi process. Rankings after Round 2 are listed in Table 2.

[insert Table 2 here]

Examining the written justifications that participants provided revealed that participants tended to rank recommendations highly that were related to three central themes: (1) strengthening the connection between goals, design and final report (2) providing critical feedback, in addition to highlighting existing strengths and (3) describing participants as less monolithic, and taking into account demographic information of visitors and participants that affect their experiences at the museum.

Requiring a "chain of logic" that links program goals, evaluation questions, design and final report

Several of the highest rated guidelines were about tying an evaluation to program improvement through a chain of logic (e.g. 1, 2, 4, 5, 7, 8 in table 2). These recommendations focused on establishing a chain of reasoning that links the program, the design of the evaluation, and the way in which findings will be used. For example, the top-rated guideline was "Evaluators should know the goals of the program/exhibit/course." Further down the list was, "Questions should be clearly articulated." Finally, "Evaluators should provide a chain of logic driving the design and methods. Logic chain should link initial questions, design and methodology, data analysis and conclusions/recommendations" also emerged as a highly rated guideline. As one participant in the Delphi study explained, "In general, I rated items about having a logic model or chain of reasoning highly—to be able to use the findings, the evaluation needs to lay out the logic model and closely connect to the design."

The importance of critical feedback

Another theme that emerged in the highly rated recommendations was that evaluations should offer *critical feedback* to programs. For example, participants rated "Evaluations should identify things that are not working in addition to things that are working" very highly. Also, "Evaluations should move past the idea that "the museum is a great place to learn" to what are the specific assets and resources and how are they used" emerged as an important guideline. One participant remarked in her justification, that she "felt that making sure evaluations are not wholly 'positive' is an important step for us and pretty important across the board." Another participant echoed this sentiment, saying, "I think the statements about what is not working is crucial and sometimes overlooked."

Not treating participants and visitors as a monolithic group and recognizing the importance of detailed demographic data

Participants also rated highly statements that emphasized that program participants or museum visitors are not a monolithic group, and a high-quality evaluation will separate findings by group or, at the very least, *acknowledge that there are different profiles of participants* or visitors. Three top-rated guidelines addressed this: (1) Evaluations should be culturally sensitive (2) Evaluations should take into account that there are different kinds of visitors and (3) Evaluations should acknowledge the different "profiles" of visitors. One justification for the importance of such a recommendation stated:

> I think that acknowledging that there are different profiles of visitors, teachers, students will strengthen evaluation reports because we will know what is working for whom . . .

Another participant clarified, "I think context, audience, and any other details to describe the environment is useful for making sense of the data as well as looking at how this could be transferred to a new situation."

Guidelines that were not rated highly

Recommendations that focused on methodology (e.g. "evaluations should not rely on self-reported learning gains" or "evaluations should include use mixed-methods approaches") tended to be ranked lower, often accompanied by justifications along the lines of, "this recommendation may be relevant in some cases, but it is not general enough to be included in a set of institutional guidelines." One participant even acknowledged that the requirement of "mixed methods" had been a recommendation that she assumed she would always advocate for. However, the process of evaluating recommendations and providing justifications made her see that perhaps recommendations about methodology might not be as generalizable across programs. She commented, "I could see ways in which having mixed methods might not be necessary under all circumstances, however, so even though I really appreciate that approach I didn't rate it as highly."

Products of the work

The primary product of this collaborative process was a final set of *guidelines* (included in Appendix A) for designing and conducting evaluations at the AMNH and a *rubric* (included in Appendix B) for assessing proposals and final reports which we can use when receiving materials in response to our requests for proposals, as well as when developing internal evaluations. The guidelines elaborate the museum's vision for rigorous evaluations, a description of our context, and our central commitments in terms of the evaluation process. The rubric for proposals includes main categories of framing, design, and value, budget, organization and relevant experience.

Discussion

We found that engaging in a process to develop shared vision of what "good" and "strong" evaluations look like, as well as developing a common understanding of rigor and methods, helped us learn about how to improve our work. Having the evaluation guidelines clarified the way we write requests for proposals and has helped us better communicate with each other about the facets of our evaluations. Members of the group are currently applying the guidelines and rubric in evaluation efforts at the museum. For example, in the first formal evaluation following the development of the guidelines, group members recruited and worked with an external evaluator to evaluate a mobile technology application that visitors could use during a day at the museum. During that process we realized we needed a rubric to rate proposals from evaluators; the rubric we developed for this purpose is in the Appendix. Using the rubric, we were able to select an evaluation firm that prepared a proposal with a clear chain of logic, from evaluation questions to proposed methods. The evaluation itself resulted in findings that we were able to apply immediately to enhance the application as well as findings that have continued to guide us through new technology initiatives (e.g. ongoing work on "emerging media") Internally, we have conducted two front-end "prototyping" evaluations for exhibition hall renovations and used the guidelines to shape the methods and analysis. We have used the guidelines and rubric as a common document throughout the evaluation process to ensure that we are remaining committed to our vision and, ideally, receive valuable and helpful information from each evaluation.

Conducting rigorous evaluations that will be useful for program development is a challenge for all informal learning institutions. We recommend that informal science institutions and cultural institutions use a systematic approach developing a vision and guidelines. We acknowledge that most museums do not have internal evaluators and researchers and thus may not have the resources to conduct an extensive Delphi process. However, simply reading previous evaluation reports (or, in the case of museums new to evaluation, reading reports from other institutions) can stimulate group members to reflect on how previous evaluations have been conducted, whether the findings were useful and were able to be applied, or whether the report has simply remained in a file folder somewhere. Drafting the documents (guidelines and rubric) provided us a way to negotiate our shared values. Our guidelines ensure that evaluations are actively and regularly used; that our evaluations gather data that is meaningful and helpful and in line with our mission. Furthermore, we hope that this attention to rigor in evaluations will place us in a strong position with initial findings so that we can dive in, when appropriate and possible, to conduct *research* on teaching and learning at our institution. Though an immediate goal is often evaluating the reach and impact of a program and providing recommendations for program improvements, we always aim to use research more broadly to improve our work and visitors' experiences.

Acknowledgements

We would like to acknowledge the contributions of all of the members of the Research and Evaluation working group at the American Museum of Natural History.

Anna MacPherson is the Sr. Manager of Educational Research and Evaluation at the American Museum of Natural History. She earned her doctorate in science education at Stanford University in 2015. Before doctoral work, she taught high school science in New York City.

Karen Hammerness is the Director of Educational Research and Evaluation at the Museum of Natural History. She has authored multiple articles and books on the topic of teacher learning. She earned her doctorate in educational psychology from Stanford University in 1999.

Preeti Gupta is the Director of Youth Learning and Research at the American Museum of Natural History. She earned her doctorate in Urban Education at the City University of New York in 2009. Before coming to the museum, she was the Senior Vice President for Education and Public Programs at the New York Hall of Science in Queens.

Notes

1. Allen et al., *Framework for evaluating impacts*
2. For example, Fu et al., "Room for Rigor"
3. Falk et al., "UK Science Education Community"
4. Fu et al., "Framework for Evaluation, " Fu et al., "Room for Rigor"
5. Commission on Museums for a New Century, *Museums for a New Century*
6. Miles, Huberman, and Saldana, J., *Qualitative Data Analysis*
7. Clayton, "Delphi" Dalkey and Helmer, "Delphi Method"
8. Kloser, "Core Science Teaching Practices," Osborne et al., "Ideas About Science," Seakins and Dillon, "A Modified Delphi Approach, "Delbecq, Van de Ven, and Gustafson, *Group Techniques*

Bibliography

1. Clayton, Mark J. "Delphi: A Technique to Harness Expert Opinion for Critical Decision-making Tasks in Education." *Educational Psychology 17*, no. 4 (1997): 373-386
2. Commission on Museums for a New Century, *Museums for a New Century* (Washington, D.C.; American Association of Museums, 1984)
3. Dalkey, Norman, and Olaf Helmer. "An Experimental Application of the Delphi Method to the Use of Experts. *Management Science* 9, no. 3 (1963): 458-467.
4. Delbecq, Andre L., Andrew H. Van de Ven and David H. Gustafson. *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*. Glenview, IL: Scott, Foresman, 1975.
5. Education Development Center, Inc. Convening Report: Defining an afterschool research agenda. ITEST, 2010.

6. Falk, John, Jonathan Osborne, Lynn Dierking, Emily Dawson, Matthew Wenger, and Billy Wong. Analysing the UK Science Education Community: The Contribution of Informal Providers. *London: Wellcome Trust* (2012).
7. Allen, Sue, Patricia B. Campbell, Lynn D. Dierking, Barbara N. Flagg, Alan J. Friedman, Cecilia Garibay, and David A. Ucko. "Framework for Evaluating Impacts of Informal Science Education Projects." In *Report from a National Science Foundation Workshop. The National Science Foundation, Division of Research on Learning in Formal and Informal Settings*. 2008.
8. Fu, Alice C., Archana Kannan, Richard J. Shavelson, Lisa Peterson, L and Amy Kurpius. "Room for Rigor: Designs and Methods in Informal Science Education Evaluation." *Visitor Studies* 19, no. 1 (2016): 12-38.
9. Fu, Alice C., Lisa Peterson, Archana Kannan, Richard J. Shavelson, and Amy Kurpius. "A Framework for Summative Evaluation in Informal Science Education." *Visitor Studies 18*, no. 1 (2015): 17-38.
10. Kloser, Matthew. "Identifying a Core Set of Science Teaching Practices: A Delphi Expert Panel Approach." *Journal of Research in Science Teaching 51*, no. 9 (2014): 1185–1217. https://doi.org/10.1002/tea.21171
11. Miles, Matthew B., Huberman, A. Michael Huberman and Johnny Saldana. *Qualitative Data Analysis: A Methods Sourcebook 3rd Edition*. Sage, 2014.
12. Osborne, Jonathan, Sue Collins, Mary Ratcliffe, Robin Millar and Rick Duschl. "What 'Ideas About Science' Should be Taught in School Science? A Delphi study of the Expert Community." *Journal of Research in Science Teaching 40, no.* 7 (2003): 692–720. https://doi.org/10.1002/tea.10105
13. Seakins, Amy and Justin Dillon. "Exploring Research Themes in Public Engagement Within a Natural History Museum: A Modified Delphi Approach." *International Journal of Science Education, Part B* 3, no. 1 (2013): 52-76.