Communicating Chemistry in Informal Environments: Evaluating Chemistry Outreach Experiences

A white paper prepared for the NRC

Committee on Communicating Chemistry in Informal Settings

Vera Michalchik[1]

Introduction: Why Evaluate?

The most important consideration in evaluating chemistry outreach efforts is how to best use the evaluation to serve *project* needs. Evaluation should be about making programs more effective—at communicating ideas, changing attitudes, inspiring action, or reaching wider audiences, for example. A well-conducted evaluation typically contributes to the quality of a project by helping its leaders better define their goals, identify important milestones and indicators of success, and use evidence to support ongoing improvements. At its best, evaluation is an integral part of project design and implementation from the outset, making it truly meet the needs of the people the project is intending to serve.

Depending on when an evaluation is undertaken and its emphasis, evaluation helps project leaders:

- Better understand what interests, needs, prior knowledge, and motivations characterize the intended audience (*front-end evaluation*)
- Get early-stage audience feedback on design, development, and implementation decisions to shape the direction of the project (*formative evaluation*)
- Determine if the project goals for serving its audiences have been achieved (*summative evaluation*).

The role played by these first two types of evaluation—front-end and formative—is critical, yet also sometimes under-appreciated in the push to show final project results. Front-end and formative evaluation make it possible for project leads to know who their participants are and whether the activities, programs, presentations, or events they create work to engage those participants. Summative evaluation focuses more directly on *outcomes*, the effects a project intends to achieve. Intended outcomes can be thought of as an answer to the question: *What ideally will happen for participants based on their experience*? Outcomes are often framed in terms of what participants do, feel, or think, and are specific to a projects nature and goals. Overall, the different types of evaluation should work synergistically, targeting the important and meaningful outcomes a project is trying to achieve by helping determine if the project's design and implementation are actually effective.

The evaluation process typically starts by formulating key questions regarding a project's audience, design, implementation, or outcomes. Once relevant questions have been posed, the evaluation process continues through phases of collecting data to answer the questions, analyzing and interpreting the data, and applying findings in whatever ways are most useful. Beyond benefitting project design and improvement, evaluation is often used to provide evidence of a project's value. Therefore, findings from an evaluation can be of interest to funders, potential partners, and other stakeholders, in addition to project designers and leaders. The report resulting from an evaluation can be a vital resource for the

[1] Stanford University and SRI International

next stages of a project, including its replication or scaling, and for others looking to learn and benefit from peers' experiences.

Although program evaluation is typically conducted by trained professionals using specialized techniques, there are many ways in which anyone can use evaluative approaches to inform project design and development. Trying out an activity in a test run with a few participants is a great way to learn more about an audience, to see what changes to materials or information might be needed, and to otherwise "ground-truth" assumptions about how to meaningfully communicate chemistry with the audience.

Evaluation in Informal Settings: Understanding the Context

Projects centered on communicating chemistry in informal environments will each have their own flavor. Their themes, objectives, activities, and audiences will differ from one another and, importantly, also differ from typical chemistry learning experiences in school. Many of the standard practices for conducting educational evaluation were developed in the context of schooling. However, evaluations of informal learning experiences generally stand apart from the kinds of evaluations conducted in formal educational settings. There are some particular and distinct challenges associated with evaluating informal learning experiences.

One obvious challenge is that in many informal settings project evaluators have limited access to participants, making it difficult to query them in relation to their experience. The nature of the setting is a critical factor. The norms of afterschool or enrichment programs make children generally less amenable to taking surveys or being tested than they would be during the school day. Visitors to a museum often will agree to serve as participants in an evaluation study, but the same individuals may not want to spend the time being interviewed or responding to a questionnaire after viewing a chemistry outreach demonstration at a shopping mall. Similarly, attendees at a forum or other drop-in event might not wish to remain after the program is over to provide opinions and feedback. Broadcast or web-based media outreach efforts have a particularly difficult set of logistical challenges when it comes to collecting evaluation data. Overall, gathering pre- and post-event data from participants in informal science settings requires a great deal of planning, persistence, and sometimes even luck.

A critical characteristic of informal learning is the indeterminacy of how people freely choosing to participate in a learning experience might benefit from it. People join in informal learning activities for a wide variety of reasons: they were brought by friends, attracted by an intriguing tag line, looking for new knowledge, or, perhaps, stumbled into it by accident. People similarly draw value from informal learning experiences in different ways, according to their life histories, their social surround in the moment, and what retrospective sense they make of the experience in the future. There is not a "right answer" in relation to the "lesson," no single outcome stemming from the intervention. Rather, the very notion of a learning outcome, the basis of most educational evaluation approaches, becomes somewhat complicated outside of school.

In contrast to school, then, there are particular difficulties associated with assessing learning from informal experiences:

- Outcomes in informal settings are contingent upon the goals and agendas people set as activities unfold, creating ever-changing variations in the learning opportunities that are actually being presented in the setting.
- People participating in activities in informal settings use resources spontaneously, flexibly, collaboratively, experimentally, and differentially to suit their own ends, making the experience (or "test" conditions) non-standard across people and situations.

- Inserting any type of assessment or testing procedure into the flow of informal activities can be disruptive and demotivating, which limits the use of many types of approaches and instruments. (Michalchik & Gallagher, 2010: 212, adapted from Bell, Lewenstein, Shouse and Feder, 2009: 55-57).

A word is in order here about the difference between *evaluation* and *assessment* as terms of art in the educational research community. Evaluation is the set of techniques used to make judgments about the characteristics and qualities of a program or initiative, whereas assessment refers to the measurement of what individuals have learned as a result of participation in a given activity or program. Assessment is frequently an important component of evaluation, since measuring participant outcomes—changes in what they do, feel, or know—is the bedrock of most summative evaluations.

But given the difficulties with assessment in informal learning settings, especially, the problem of administering appropriate measures for predetermined outcomes, the lines between evaluation and assessment often get blurred. The qualities and characteristics of project implementation that are a focus of evaluation become difficult to separate from the qualities and characteristics of the audiences' participation and what people might actually learn from participation. For example, an evaluation of a hands-on wet lab for visitors to a museum will focus on the degree to which the activities are appropriately designed to attract and engage participants in meaningful scientific activity. At the same time, a key outcome for this project will be that participants are attracted to and stay engaged in the activities that are offered.

In an informal setting, an outcomes-oriented evaluation might focus on precisely the same data—that is, visitors come, they stay, and they productively engage. These types of indicators of the project's success would be the same for a public science café or forum, a media-based outreach, and many other types of informal chemistry outreach events. In a formative framework, it would be quite possible to show that improvements to the project design or implementation results in attracting more visitors, who stay longer, and even who "perform better" given the project activities. Most summative evaluations also seek ways to debrief participants, and some manage quick versions of pre- and post-experiences assessments of what participants do, feel, or think in relation to the phenomena central to the informal science experience. But short of an adequately-powered experiment with a control group—an inherently expensive and time-consuming proposition—it's usually impossible to *prove* that an experience is responsible for particular learning outcomes. Add to this the fundamental challenges with specifying and measuring particular outcomes in informal spaces, and most informal evaluation efforts rely on outcome indicators associated with people's participation and self-reported measures of attitudes.

Nonetheless, with the right resources and creative approaches, evaluators successfully study a diverse range of implementation and outcome variables in informal science settings. A recently published web document, *Principal Investigator's Guide: Managing Evaluation in Informal STEM Education Projects* (Bonney, et al., 2013) gives several examples of summative outcomes associated with informal science projects. The nature of the outcomes cited in each example is quite different. One evaluation used observations to document a dramatic increase in the percentage of participants building free-standing structures based on design changes to a construction exhibit at the Chicago Children's Museum. Evaluators for the Fusion Science Theater provided audience members a questionnaire about their science attitudes and interests before and after the show, and used informal interviews to better understand the nuance of the positive aspects of participants' experience. Evaluators for Dragonfly TV created the opportunity to ask children before an after viewing an episode to rank the importance of several aspects of inquiry (such as the importance of keeping some variables constant across trials and recording results), showing notable gains based on the viewing.

In sum, even though it's doable, conducting evaluation in informal settings can be tricky business. Not everyone engaged in chemistry outreach activities will be able to get past the challenges associated with measuring the kinds of solid summative outcomes that funders might seek. But for chemists looking to provide a public service, show the meaning of their work to a wider audience, and improve appreciation of the beauty of their field, getting to know the people the project is trying to reach and seeing how they respond to outreach efforts can be done even at a small scale.

Evaluating Informal Chemistry Outreach: Building on the Basics

Researchers studying professional chemists note that since molecular properties and processes are not available to direct perception, the practice of chemistry relies on a range of representational systems to generate meaning within the discipline (e.g., Kozma, Chin, Russell & Marx, 2000). A rich literature on chemistry education shows that the representations chemists use to build shared understanding among professional colleagues are difficult for students to conceptually connect to nanoscopic phenomena and also to macroscopic experiences in the wet lab (Gabel, 1998; Michalchik, et al., 2008; Nakhleh, 2002; Sevian & Gonsalves, 2008). Consequently, by the time they finish their last chemistry unit in school, most people develop little appreciation for the field.

While it may be true that chemistry, as taught, is hard to learn, this is only one aspect of chemistry's bad reputation. As discussed in the *Communicating Chemistry Landscape Study* (Grunwald/EDC, 2013), the challenges in communicating chemistry are manifold. Public perceptions of chemistry and the chemical industry, particularly their impact on human health and the environment, are decidedly negative—to the point where the termed "chemophobia" has been coined. Chemistry is also complex, abstract, and arguably less intuitively accessible and applicable than, say, biology or physics. In relation to other scientific disciplines, chemistry suffers from a lack of its own distinct identity or "disciplinary unity," existing in a diffuse, in-between space among other fields that it supports and facilitates (biochemistry, physical chemistry, etc.). Chemistry activities that are hands-on and inquiry-oriented are seen as messy, dangerous, and, because of cost of consumable materials, expensive. Finally, chemistry has a somewhat balkanized and even secretive culture, partly because of the influence of industry on the field and partly because of the nature of the practice of industrial chemistry research.

Why is it important to recount these challenges? Any effort to communicate chemistry to the public in a manner that is both respectful and effective must account for the prior experiences and resulting frameworks that audiences carry with them. While it may seem that the elegance and inherent interest of the content should be enough to attract audiences, chemistry outreach must first overcome a problem of affect. The findings reported in the landscape study indicate that people's feelings towards the subject are somewhat negative, and foundational to any approach to reach public audiences is to meet them where they are.

The evaluation should be grounded in the same way. A good evaluation, like a good project design, starts by taking audience into account. The full set of design principles articulated in the landscape study builds on this foundation. Moreover, the principles provide an essential touchstone for the development of ways to communicate about chemistry that attracts, engages, and shifts audience feelings about the subject. Project leaders and evaluators can use the principles to audit the project as it evolves, ensuring that:

- A **focus on audience** is maintained throughout the design and implementation of the project.

- The project also does not lose its **focus on interesting questions**, working to ensure that audiences can see how details relate to the big picture established in the project's "hook" and overview.
- The project is built along a narrative arc and additionally **tells good stories** along the way.
- Project leaders and evaluators continue to **be thoughtful about outcomes** throughout the entire life of the project.

These design principles are consistent with the in-depth criteria for assessing scientists' written communication skills developed and studied by Baram-Tsabari and Lewenstein (2013). Focusing on what scientists should know about writing for the public, the researchers constructed a schema for assessing these competencies, which included measures for the following:

> **Clarity**: Using language that is familiar and technical terms (jargon) that are well defined; ensuring ease of readability; providing appropriate type and quality of explanation.
>
> **Content**: Presenting correct information and science content; addressing the nature of science; providing connections to everyday life.
>
> **Knowledge organization**: Framing knowledge in ways that are meaningful and compelling (e.g., social progress, morality and ethics, public accountability); scaffolding explanations by providing adequate background.
>
> **Style**: Using humor, especially explicit jokes and irony language.
>
> **Analogy**: Using analogy or metaphor as dominant tropes.
>
> **Narrative**: Telling overarching and embedded stories, with beginning, middle, and end structures.
>
> **Dialogic approach**: Referring to multiple worldviews; showing respect for other worldviews; engaging in argumentation.

These criteria also, fundamentally, are about recognition and understanding of audience perspective, and should therefore be a touchstone of evaluation of chemistry outreach. The authors of the study note that the discourse underlying the professional practice of science—that is, scientists' skill in communicating with other scientists—is quite unlike and even at odds with the discourse of public communication. The authors write, "One rewards jargon, the other penalizes it; one rewards precision, the other accepts approximation; one rewards quantification, the other rewards storytelling and anecdotes" (2013: 80). Although the graduate student scientists in this study "knew" what they should do for effective communication, they only minimally employed narrative, analogy, and metaphor, and often failed to address the everyday importance and nature of science, while overemphasizing content. In their study of verbal explanations by graduate students, Sevian and Gonsalves (2008) focused on similar qualities (e.g., clear choice of language, use of mental images, connection to larger contexts) and also found a broad range of skills among their participants.

As noted in the introduction to this paper, a push to identify and measure individual learning outcomes can shortchange a primary function of evaluation: supporting the effective design and implementation of activities appropriate to the audience and the goals of the project. Premature attempts to assess project outcomes can be meaningless or, worse, can destroy the chances to continue, improve, and sustain a promising program. Between getting familiar with the prior knowledge, interests, and attitudes of the audience, on the one hand, and measuring outcomes, on the other, are myriad opportunities to reconsider, massage, and fine-tune an outreach effort. Where the rubber hits the road in communicating chemistry is where the audience and experience meet—the point at which what project developers do actually matters. To this end, the *Communicating Chemistry Landscape Study* (Grunwald/EDC, 2013) provides a comprehensive array of attributes and characteristics of experiences that can serve to organize both project design and evaluation, which will be discussed further in the next section.

Of course, the design principles in the landscape study do not suggest that experiences should be designed without direction or bearings. Being "thoughtful about outcomes" means being thoughtful about what kinds of future possibilities for participants the chemistry outreach experience provides. As such, outcomes are tightly coupled with the nature of the experience, and this view strongly suggests that, just as the design and implementation of an experience are likely to evolve over time, so are the intended outcomes. For this reason, in recent years a new type of evaluation, *developmental evaluation*, has been defined and increasingly applied (Gamble, 2008; Patton, 2011). Developmental differs from formative evaluation in that, rather than presupposing an a priori project model, developmental evaluation supports flexible responses to the ideas, successes, and failures that continue to emerge in pursuit of a general goal during project pilots and implementation.[2]

In addition to attributes and characteristics of informal chemistry experiences, the landscape study provides a broad framework for the types of goals that can guide the development of an outreach project. For any type of evaluation, this framework can serve as a helpful tool for evaluators, working with project leads, to reflect on where the project might be headed and how best to help it get there.

Table 1: Project Purpose or Goal

Dimension	Categories	Sub-categories/descriptive
Purpose/Goal	Education	• Develop interest, motivation to learn • Understand and use concepts • Explore, question, and observe • Reflect on science processes • Use science language and tools • Develop a science identity
	Communication	• Awareness of science • Enjoyment or other affective responses to science • Interest in science • Opinions with respect to science • Understanding of science

Getting Started with Evaluation: Some Basic Resources and Supports

Although no short paper can provide anywhere near enough information to serve as a "how to" for project evaluation, it is helpful for a project lead to be aware of the key steps most evaluations follow as

[2] Penuel (2010) takes such ideas further, eschewing the notion of an objectivist stance in educational evaluation, and, instead, arguing for a "dialogical epistemology." In Penuel's framework, rather as acting as arbiters, evaluators are enjoined to act as participants, responding to the diverse perspectives of stakeholders in judging the worth of a program.

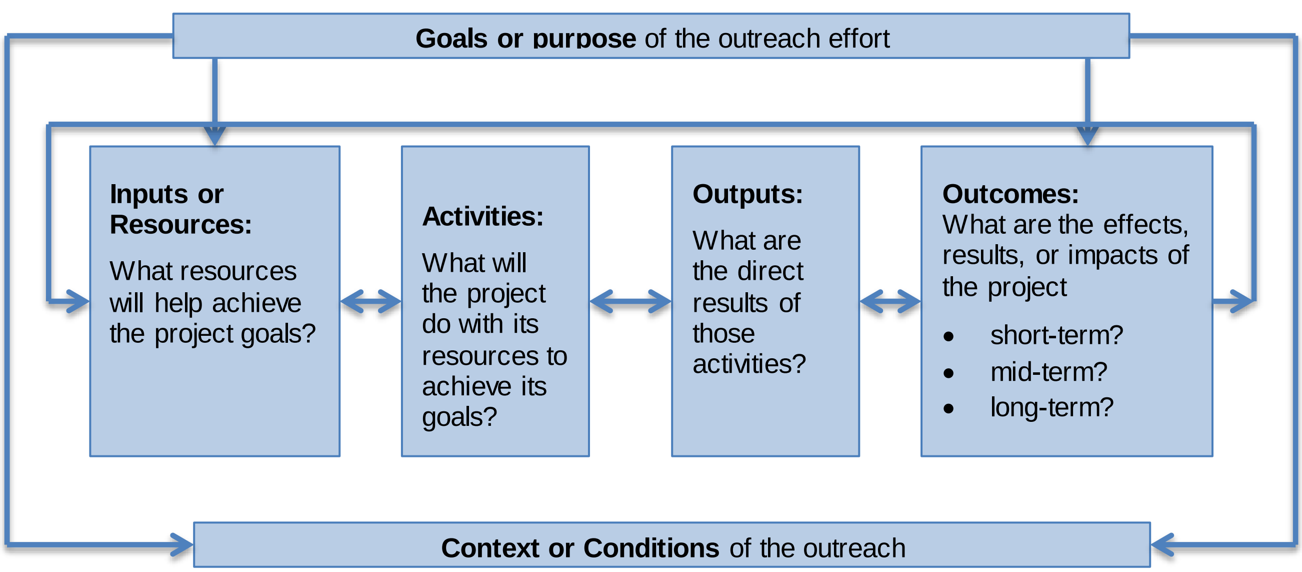
they are developed and conducted. The following elements define the typical arc of professional evaluation and can help anyone, whether an individual providing a small outreach experience or a university department partnering with a major science center, think about how they can conduct evaluation to the benefit of their efforts. Depending on the project leader's resources and capacity, these steps might look daunting. Perhaps the right attitude with which to think about these elements, especially when considering evaluation for a small initiative, is that "some data is better than no data."

Evaluation Design

Planning and background. The first step in a canonical evaluation is general planning. Planning for an evaluation must start with as much understanding of the intent of the project, its history, its context as is feasible. It also must begin with an understanding of the purposes of the evaluation itself. Every evaluation takes place under constraints of time, budget, personnel, and other issues of capacity. Scaling expectations for what can be accomplished in an evaluation is a critical and often challenging part of developing an evaluation.

Project goals and outcomes. An evaluation design needs to be built on an understanding of the project rationale. Often, this rationale is discussed in terms of a *logic model.* A logic model specifies the stages in a project to show how the project works. It defines initial resources (inputs), the nature of the experience participants will have (activities), the number, attendance for each, and other readily quantifiable metrics of those activities (outputs), and the effect or changes that will result from participation (outcomes). The process of creating a logic model helps identify and define all theses elements. A logic model should not be static, since any element is likely to change as the project develops and those changes should be reflected in the model.

Figure 1: Basic logic model



A *theory of change* also maps the relationship between activities and outcomes, but does so more explicitly in relation to the experience of the participant. For example, a theory of change for a chemistry outreach project in the media might link the fun and intriguing features of the presentation with the assumption, expectation, or, even, hope that more positive attitudes towards the discipline and the work will result. However expressed, a project model or theory needs to be based on careful and meaningful alignment between the nature of the experience for participants and the expected outcomes. If a project promotes playful exploration of materials and substances, it should not be expected to directly and necessarily lead to skill in arguing from evidence, for example.

Evaluation questions. Evaluation questions necessarily depend on project goals, the purposes of the evaluation, the phase and features of the project to be evaluated, and stakeholders' interests. Evaluation questions, more than anything, are what drive an evaluation. In the end, the results of the evaluation should be the answers to the evaluation questions. Since they structure the entire evaluation, the evaluation questions should establish both what is and what is not to be evaluated. If project leaders need to know how the qualities and characteristics of facilitation are promoting participation and engagement in a hands-on experience at a museum, for example, then facilitation should somehow be addressed in the evaluation questions. Likewise, if parental involvement is not a goal of the project, then questions about parents' participation in the hands-on activity would not be included. (However, if it turns out that parents' involvement seems to matter, and stakeholders such as the museum director or funders start to seek information about parents, then the evaluation questions could be modified.)

The following are some sample evaluation questions that could be adapted to a range of specific projects:

- What topics might attract or intrigue the intended audience? (front end)
- What made it likely or possible for the audience to participate in the project? What barriers (e.g., lack of publicity, lack of transportation to the site) were there? (formative)
- Did participants engage in the activities as planned? What did they do differently than expected? What could be changed in project design or implementation to increase engagement? (formative)
- Did attendees at the event demonstrate increased understanding of the topic or issues addressed? Which attendees and to what degree? (summative)
- Did audience members change their attitudes or behaviors based on the outreach activities? What level of involvement made the difference? (summative)

Note that evaluation questions are *not* the questions themselves that audience members would be expected to answer in interviews or questionnaires. Often, multiple data sources contribute to answering each evaluation question, and all data collected must be interpreted to some degree. Also note that the only good evaluation question is an answerable one. For instance, if, given resources, there are no means for contacting or otherwise getting evidence about the audience members for a broadcast media presentation, then the evaluation should not pose questions about changes in audience behavior, attitudes, knowledge, or such.

Indicators. Indicators provide evidence regarding the outcomes, implementation factors, and other variables targeted by the evaluation questions. Indicators, therefore, should "point to" or otherwise help reveal whether or not a project objective has been achieved. A useful indicator must be measurable or, at least, observable, in the same way that a good evaluation question must be answerable. Indicators answering a question regarding engagement as an outcome of a public mini-demonstration of chemical reactions might include the number of people who watched, the length of time they stayed, their facial expressions and degree of attentiveness, the types of questions they asked of the presenter, and their descriptions of why they watched, what they liked, what they learned that was new, and what they might do next or differently on the basis of their experience. While many indicators will readily be quantifiable, some will not, instead providing qualitative insight as to the value, meaning, or import of the experience for participants.

Evaluation Methods

Evaluation design. Evaluation designs run a wide range of types. What type is used depends on the questions to be answered, the resources available, stakeholders' expectations, and the appropriateness of data collection and other evaluation activities to the nature of the experience and the context of the setting. In the example of a public mini-demonstration, the methods would include observations and interviews, many elements of which could be quantified, with also perhaps a brief questionnaire handed out to participants (many of whom might not care or bother to respond). Participants could provide retrospective self-report on how much, if at all, their attitudes and feelings about the topic at hand changed because of the demonstration. A more rigorous design would query the participants before and after the experience to see if in more objective terms the experience influenced attitudes. If it were important to know what content knowledge the audience learned from the experience, the most likely approach would be to provide participants a short pretest and posttest, though, as discussed, this would be quite out of the norms with many informal learning settings, especially drop-in public settings such as a festival or shopping mall.[3]

Important considerations regarding the evaluation design include whether or not the design allows for findings contrary to the hoped for outcomes. Of course, no study should simply confirm foregone conclusions and the evaluation should be structured to account for this. Similarly, any survey, interview, or other instrument used must have a basis for verifying its validity (is it measuring what it purports to measure?) and reliability (are the measurements stable and consistent?). If an instrument asks questions about liking to observe clouds, birds, and butterflies, how do we know it validly indicates *interest* in science if promoting interest is what we care about? If respondents' answers would change from one moment to the next depending on their mood or some factor other than how much they really like observing clouds, birds, and butterflies, then the reliability of the instrument is in question.

Evaluation management. Project leads working with evaluators will want to be sure that their evaluation teams possesses the credentials and qualifications needed to conduct the evaluation. Each evaluation should have some type of management plan that specifies key milestones and deliverables and includes timeline for key elements of the evaluation, including reporting. The management plan should account for budget details such as allocation of resources for research, travel, staffing and project management itself. To ensure that there are not cost overruns, cost estimates should be reasonable and appropriately allocated across evaluation sub-tasks, such as design, data collection, and analysis.

Data collection. Several types of data sources have already been mentioned: observations, interviews,

---

[3] The gold standard for providing causal evidence of learning outcomes research is a pre- and posttest experimental design with a control group and subjects randomly assigned to the two differing experiences so as to avoid bias. This type of design can shed light on the "counterfactual"—that is, what would have happened without the experimental condition—but is not often used in informal science learning research.

and questionnaires, including test-like instruments to assess attitudes or knowledge. There are many other forms of data, including artifacts participants have created in the course of their experience (e.g., drawings, constructions, photos and videos, or documents), specialized evaluation tools such as concept maps, and particular forms of behavioral observation such as tracking and timing or web analytics. Creative variations on the usual species can be quite helpful: wall space for kids to answer questions with sticky notes or teens to write some graffiti can work better than a staid survey approach.

The primary source of data for most evaluations will be the individuals who participate in the program—particularly those from the target audience who participate to the extent intended. The instruments used should be appropriate to the audience in terms of language, culture, age, background, and other potential barriers to communication. The mode of delivery (e.g., online or paper survey) should also be tailored to the audience. Any direct involvement of human subjects in data collection will require clearance in advance from an institutional review board.

It is important, in all cases, that data sources be specified, appropriate, and as free of bias as possible. A frequent concern about bias in informal science evaluations is the social desirability factor—participants, like all people, will lean towards telling a friendly interlocutor what it is they want to hear. Additionally, bias can be introduced by asking questions of a type that only some participants can answer, depending on age, social circumstances, and such. A younger child without experience talking to strange individuals may say nothing about an entertaining presentation, yet his silence should not be construed as lack of interest or appreciation compared to the exuberant sibling who carries on about what fun she had. Sampling bias is another major concern. If the only audience members an evaluator speaks with are ones who volunteer or show particular enthusiasm for the experience, these factors might represent a bias regarding who is being studied. The individuals not volunteering to speak to the evaluator might have quite a different perspective than those who do. Low response rates to a questionnaire sent after an event might also indicate a bias regarding the types of people likely to readily respond—e.g., people who loved the experience or disliked it to such degree that they wished to express the reasons why.

Data analysis. The data analysis phase of an evaluation involves, first, analysis of each individual data source (e.g., tabulating responses of parent surveys and, separately, coding behaviors in field notes of observations of children) and, second, the synthesis across data sources to answer each evaluation question. Decisions about how to analyze and synthesize data are complex; the bases for an evaluator's interpretations are often even more so. A strong evaluation, however, justifies these decisions and makes the reasons for particular representations transparent, linking the *art* of evaluation to its science. Any process for vetting findings with colleagues or experts can strengthen the quality of the findings and likelihood that the evaluation will be well received by stakeholders. The process of analysis will help make clear the study's limitations, and, at the same time, often reveal unexpected or serendipitous findings. This is also the phase in which evaluators speak to the implications of the results, deriving helpful recommendations for the project leads and, in many cases, the field more broadly.

Reporting

Progress reports. A professional evaluator typically provides both progress reports (annually, semi-annually, quarterly, or monthly, depending on need) and a final report to the project leads. The progress report typically keeps the leads apprised on any changes since the finalization of the evaluation plan, using these reports as a type of addendum to the plan, justifying the changes as appropriate. Changes might have occurred in the evaluation approach, management, data collection strategies, and analytic methods. A progress report can focus on evaluation processes, providing information about activities exclusively, or can also include preliminary data analyses and results. While it can be quite helpful to engage in discussions around early findings, which also helps keep the evaluation on track, reporting early findings can detract from the data collection and analysis effort, especially if those findings do not account for the full set of data and may be superseded when all the data is analyzed.

***Final report*.** Findings are typically reported in an extended narrative that provides the evaluation questions, design and rationale, methods, analytic approaches and findings. Stakeholders highly value an executive summary that highlights findings, and the executive summary should be written in as reader-friendly a form as possible. Depending on the audience, key findings are increasingly presented in alternative formats such as Power Point slides or through infographics. The format and content should be customized to the audience.

The basis for conclusions and recommendations derived from the analysis should be readily evident to those reading the final report. The best recommendations are specific, actionable, tied to findings and data, and appropriate to the context for the project. Data collection instruments, coding schemes, and other technical tools are usually included as appendices.

Evaluation resources and support

The National Science Foundation and other governmental agencies have sponsored the development of several noteworthy guidelines or supports pertinent to informal science evaluation in recent years. They each offer a distinct perspective and can help orient project leaders and other non-evaluators to the basics. They also each provide resources for working with evaluators, guiding evaluations, or conducting some basic form of evaluation. The following is a particularly useful example:

Principal Investigator's Guide: Managing Evaluation in Informal STEM Education Projects (Bonney, Goodyear, Ellenbogen, Hellenga, Russell, Marcussen, Luke, Yalowitz, Palmquist, & Phillips, 2013)

Mentioned previously in this paper, this guide is designed for the purposes the title indicates: to help informal science education PI's work with evaluators to "integrate evaluation into all phases of project design, development, and implementation." The guide provides a basic orientation to the key elements of evaluation without assuming that the reader him or herself would actually undertake the evaluation. It is detailed enough to serve as Evaluation 101 and is an excellent starting place for a PI working with an evaluator to design a project evaluation. Published in an interactive web format (http://informalscience.org/evaluation/evaluation-resources/pi-guide#chapter-6), the guide provides: examples of evaluation used to support implementation (chapter 1), an extensive overview of basic definitions and principles (chapter 2), a how-to for finding an evaluator who is well-qualified and well-suited to the project (chapter 3), the respective roles and responsibilities of the PI and evaluator through the lifespan of the project (chapter 4), the details of developing an evaluation plan (chapter 5), and putting evaluation findings to their best use (chapter 6). The guide is particularly helpful with regard to planning an evaluation and presenting and communicating evaluation findings.

Others useful resources that can be readily found online include:

Frechtling, J. (2010). *The 2010 user-friendly handbook for project evaluation*. National Science Foundation.

Friedman, A. (Ed.). (March 12, 2008). Framework for Evaluating Impacts of Informal Science Education. Avaiable at http://informalscience.org/documents/Eval_Framework.pdf

Informalscience.org, collected evaluation reports http://informalscience.org/evaluation/browse?type=evaluations&search=goodman+amazing&search_url=http%3A%2F%2Fapi.informalscience.org%2Fsearch%2Fjson%3Fqq%3Dgoodman%2Bamazing#topSearch

Hussar, K., Schwartz, S., & Boiselle, E., & Noam, G.G. (2008). *Toward a systematic evidence-base for science in out-of-school time: The role of assessment*. Program in Education, Afterschool, and Resiliency (PEAR), Harvard University and MacLean Hospital. Available at: http://www.pearweb.org/pdfs/Noyce-Foundation-Report-ATIS.pdf

Hussar, K., Schwartz, S., & Boiselle, E., & Noam, G.G. (2013). *Assessment Tools in Informal Science (ATIS)* Program in Education, Afterschool, and Resiliency (PEAR), Harvard University and MacLean Hospital. Available at: http://www.pearweb.org/atis/

IES and NSF (2013). *Common guidelines for education research and development*. Available at: http://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf

Pattison, S., Cohn, S., & Kollmann, L. (2013). *Team-based inquiry: A practical guide for using evaluation to improve informal education experiences.* Available at: *http://nisenet.org*

Phillips, T. B., Ferguson, M., Minarchek, M., Porticella, N., and Bonney, R. (2013) Public Participation in Scientific Research: User's Guide for Evaluation of Learning Outcomes. Ithaca, NY: Cornell Lab of Ornithology.

Chemistry Outreach Evaluation Examples

The outreach efforts described in the following two cases illustrate how different the challenges for evaluating chemistry outreach can be. Anyone new to evaluation of chemistry outreach initiatives would do well to study additional examples as available in informalscience.org and elsewhere.

The Periodic Table of Videos

Starting with an initial five-week project to create a short video on each of the 118 elements, the Periodic Table of Videos (PTOV) burgeoned into a highly popular YouTube channel hosting hundreds of professionally produced videos on chemistry topics. The postings feature working chemists in their academic settings who shared with the viewers candid insights into their intellectual pursuits and professional lives. Although the raw number of hits, positive feedback, and other indicators convinced them that their work was worthwhile, Brady Haran and Martyn Poliakoff, the journalist and chemist at the helm of the project, wondered publicly "how to measure the impact of chemistry on the small screen" (2011). An outreach genre like the PTOV comes with a built-in set of web analytics regarding viewership, but it also presents a distinct set of challenges for examining the types and degrees of influence this form of outreach has.

Haran and Poliakoff describe their uncertainty regarding how to think about the quantitative analytics they derived from YouTube. While they recognized the qualitative value of the magnitude of numbers of views ("a video with 425,000 views is clearly more popular than one with 7,000 views" [2011: 181]), they note such issues as:

- One "view" could represent the same person watching the same video repeatedly or a teacher downloading the video and showing it to hundreds of students.
- Age and gender profiles rely on data that viewers might not accurately provide.
- The meaning of geographical data, based on IP addresses, is unclear—e.g., how should researchers interpret a temporary surge in viewing in a single country?
- Subscribership indicates deeper level of interest, and comparisons to the subscriber base of, say, another science channel or football club can provide a relative sense of PTOV popularity. But what does it mean when people unsubscribe?
- How much do YouTube promotions of featured videos skew the data? Do high levels of "likes" mean that PTOV is converting viewers or merely preaching to the choir? Although few in number, what do "dislikes" mean?

The comments posted by viewers watching the videos reveal some of the nuance of their experiences and give Haran and Poliakoff, as they see it, "more useful information about impact" (2011: 181). The authors attempted to interpret the comments quantitatively, asking whether the increase in the number of comments was a feature of the PTOV itself or a reflection of a general trend. They also tried to create an index of the impact of videos based on the number of comments they generated, which proved impractical. The author's additionally used a "wordle" to analyze the comments, but concluded that this approach merely confirmed the obvious: that viewers enjoyed the program.

The authors ultimately found it most "reliable" to read and interpret viewers' comments qualitatively. They identified and examined online interactions between viewers, a number of which occurred without participation from the program's producers. They also carefully studied the email and letters sent to them that described how the PTOV affected them. They categorized these comments into two primary groups:

- Adults who, for the first time, were enjoying science, despite bad experiences with the subject in school when they were growing up.
- High school students finding their interest and aptitude in science "awakened" by experiencing an approach different from merely reading chemistry texts and solving problems.

The authors conclude that, absent a proper market survey, "these comments are probably the most accurate indicators of impact and they are certainly the most rewarding to all of those involved in making the videos" (2011:182). The authors also conclude that there is a large market for well-presented chemistry information and remaining room in cyberspace for high-quality scientific outreach, despite the difficulties of measuring how much impact these efforts might have.

The Amazing Nano Brothers Juggling Show

The Amazing Nano Brothers Juggling Show (ANB) is a live theater performance that dramatizes and models the nanoscale world (atoms, molecules, nanoscale forces, and scanning probe microscopy) using juggling and story-telling in a highly entertaining manner. The Strategic Projects Group at the Museum of Science in Boston created the show and offers it through live performances as well as in a DVD set on nanotechnology. The Goodman Research Group conducted a large summative evaluation to examine the effectiveness of the show in increasing audiences' knowledge of and interest in nanoscience and nanotechnology (Schechter, Priedeman, & Goodman, 2010). The museum had previously collected comment cards from the audience that showed that they enjoyed the program. The Goodman group focused on what the audience learned and how much they engaged with the content.

The evaluation analyzed data collected in the spring of 2010 from three different groups—131 children ages 6-12, 223 teens and adults, and 10 middle school field trip teacher leaders. The evaluators surveyed children either before or after one of several ANB performances using an age appropriate-survey. They also surveyed 223 teens and adults either before or after the shows and interviewed teachers who had seen ANB as part of a field trip with their classes. The surveys took about five minutes to complete and participants were given a rub-on tattoo in appreciation. Up to one half of the families in the audience for each show participated. On the days that the surveys were administered, printed ANB performance programs, which contain a great deal of reinforcing content from the shows, were not handed out to the audience.

For the pre-survey, Goodman group evaluators approached individual audience members and asked them if they would participate in a brief survey about the show. They were given the forms and pencils, which were collected just before the show started. An invitation to participate in the post surveys, given at different shows from the pre-survey, was announced during the show's finale by a cast member. The surveys were handed out to audience members at their seats then collected at the theater exits.

Children's surveys included the following example items:

- Can scientists move individual atoms?
- Is everything made of atoms?
- Do atoms stick to each other?
- Which is smallest? Molecule, bacteria cell, atom, or grain of sand?

Questions on the teen and adult surveys were, naturally, more complex:

- Circle the SMALLER ONE in each pair:
    - atom or nanometer
    - atom or molecule
    - microscale or nanoscale
    - bacteria or virus
    - 10 million nanometers or a meter
    - 100 billion nanometers or a yardstick

- If the nucleus of an atom was the size of a basketball, approximately how large do you think the whole atom would be?
    - The size of a basketball hoop
    - The size of a car
    - The size of a football stadium
    - The size of a large city
    - The size of the United States

- Which of the following statements do you think are true? (Choose all that apply.)
    - Everything is made of atoms.
    - Atoms can be felt with special instruments but not seen.
    - Scientists can move groups of atoms, but not individual atoms.
    - Temperature affects the movement of individual atoms.

- Atoms tend to stick together.
- Gravity affects the movement of individual atoms.
- Products using nanotechnology are already in stores.
- Nanotechnology has been proven safe.
- None of these are true.

Surveys received from audience members outside the target age group were excluded. Also, since very few teens attended the performances and took the survey, only data from participants over 18 years of age were included in the analysis, and teacher's responses in the interview provided the only data regarding 13-year-old middle schoolers who experienced the performance. Interviews with teachers were conducted as follow up discussions at a date after the performance. They lasted 15-20 minutes and included questions about the teachers' perceptions of their students' knowledge and attitudes before, during, and after the ANB performance. For example, teachers were asked: "Could you please tell me what would be the 'tagline' for your class's experience at the show?" and "Was the show especially good at getting across any particular concepts or insights that students had not been exposed to or that they had had difficulty grasping before seeing the show?"

On average scores for children in the pre- and post-performance conditions were 18% higher, with significant increases in knowledge for half the content items in the survey. Adults' scores also increased notably from pre- to post-survey, with significant increases in five of the content specific items. The Goodman researchers also found that the show was both captivating and informative for audiences of all ages. Adults found the show highly educational. Teachers found that it reinforced classroom lessons and correlated well with science standards. The theatrical techniques supported learning potential by engrossing audience members. Sections of the show involved a combination of theatrical techniques that engrossed the audience and heightened their learning potential, and the juggling was successful for teaching children, teens, and adults, particularly about the structure, movement and manipulation of atoms. For the teens and adults familiar with basic nanoscience concepts, the ANB performance helped deepen their understanding.

An Outreach Typology

The examples above describe specifics for two chemistry outreach projects and the evaluation approaches taken for each. In the tables below, this section presents a selection of different scenarios of chemistry outreach—organized by scale of effort and types of setting and activity—with suggestions for the kinds of outcome indicators and approaches that might be used to answer questions about the success of the chemistry outreach project. Sometimes an indicator is a countable number (e.g., people attending a presentation, questions answered correctly on an assessment) and sometimes an indicator will require inference or interpretation (e.g., facial expressions while watching a performance, comments about why a website was attractive and valuable). Note that levels of attendance or rates of participation are generally indicative of how attractive an informal learning experience is to participants, and a great deal can be inferred from participants "voting with their feet" by choosing to engage with an activity or not.

A project will need to borrow, adapt, or develop data collection instruments that best measure the outcomes the outreach intends to promote. Two particularly relevant resources for instruments to support data collection for informal science outreach are:

- *Assessment Tools in Informal Science (ATIS)***,** a collection of researcher development instruments for measuring many types of science outcomes, available at http://www.pearweb.org/atis/

- NISE Net Team Based Inquiry Guide, a do-it-yourself approach to evaluating public science programs, developed for nanoscience but broadly applicable, with resources available at http://www.nisenet.org/catalog/tools_guides/team-based_inquiry_guide.

The ATIS collection includes science assessment instruments from a diverse array of sources organized according to age range, question format, and domains (i.e., Competence/Reasoning, Engagement/Interest, Attitude/Behavior, Content/Knowledge, Career Knowledge/Acquisition). The collection includes the Chemistry Attitude and Experience Questionnaire or CAEQ (http://pearweb.org/atis/tools/35), which is designed for first-year college students and includes some items potentially adaptable for other ages and settings, such as

Please indicate what you think about the following:

Chemists

1 unfit _ _ _ _ _ _ _ athletic

2 socially unaware _ _ _ _ _ _ _ socially aware

Chemistry research

1 harms people _ _ _ _ _ _ _ helps people

2 creates problems _ _ _ _ _ _ _ solves problems

Resources available from the NISE Net Team Based Inquiry site include instruments appropriate for a one-time museum visit or public event: a feedback survey, observation form, participant interview protocol, question planning sheet, and data reflection "cheat sheet." The feedback survey, for example, includes the following instructions and items:

Help us improve the program you just saw!

Please take a few minutes to share your opinions below.

1. What did you like most about this activity? Why is that?
2. What are some ways this activity could be improved? Why is that?
3. In your own words, what would you say this activity is about?

Many other examples of the types of instrumentation used by individual science outreach projects, including ones for chemistry, can be found at informalscience.org and various dedicated project websites. Noteworthy is the degree to which projects typically adapt and develop tools to best match their own project goals and implementation features. For example, for a formative evaluation of Penn State's Marcellus Matters: EASE (Engaging Adults in Science and Energy) project, the evaluation team used the technique of prompting participants with a set of "reaction words" to solicit opinions about the qualities and characteristics of the eight weekly sessions on various topics pertinent to Marcellus gas drilling and related science concepts and community issues. These "reaction words" prompted participants to speak to, for each session, what it was that they found notably important, confusing, unnecessary, interesting, boring, relevant, familiar, unfamiliar, oversimplified, and valuable.

Following is the typology chemistry outreach efforts with possible indicators related to select categories of outcome goals and approaches to measurement.

Individual or small outreach efforts

A. Public presentations, demonstrations, and drop-in events	
Outcome indicator	**Measurement approach / instrumentation**
Engagement, interest, enjoyment	Length of time present, level of attentiveness, facial expression, questioning and other forms of participation, responses in brief surveys or interviews regarding why they participated, what they liked, etc.
Understanding of content	Questions asked, verbal responses or hands raised in response to questions, responses on brief surveys regarding what they learned.
Intentions towards future involvement or activity	Information seeking (e.g., taking brochures, filling out interest cards, liking a facebook page), verbal responses, responses on brief surveys regarding what they might do differently.
Impact on behavior and attitudes	Follow-on interviews or surveys regarding discussions with others (e.g., dinner conversations), posts (e.g., twitter), information seeking, signing up for programs.

B. Websites, videos, broadcasts and other media-based resources	
Outcome indicator	**Measurement approach and instrumentation**

Engagement, interest, enjoyment	Data analytics, posts and comments, responses on linked surveys or online forums regarding why they participated, what they liked, etc.
Understanding of content	Data analytics, posts and comments, responses on linked surveys regarding what they learned.
Intentions towards future involvement or activity	Participant information seeking, registering on sites, liking pages or posts, reposting online, responses on brief surveys regarding what they might do differently.
Impact on behavior and attitudes	Follow-on interviews or surveys regarding post-experience activities, reposts (e.g., twitter), information seeking, registering for sites.

C. Involvement with an afterschool program, museum-based program, or ongoing public forum

Outcome indicator	Measurement approach and instrumentation
Engagement, interest, enjoyment	Participant level of involvement (e.g., choosing "chem club" over "outdoor time," in an afterschool setting) active participation in questioning and other scientific practices, responses in brief surveys or interviews regarding why they participate, what they like, etc.
Understanding of content	Observations of participants engaging in questioning and other scientific practices with inquiry orientation, verbal responses to questions, responses on brief surveys at end of program regarding what they learned.
Intentions towards future involvement or activity	Information seeking, verbal descriptions of plans or ambitions, responses on brief surveys regarding what they might do differently based on participation.
Impact on behavior and attitudes	Follow-on interviews or surveys regarding behaviors and attitudes specific to the outreach goals. Appropriate attitudinal assessments from ATIS or other sources.

Broader, systematic outreach efforts

A. Public programming and performances

Outcome indicator	Measurement approach and instrumentation

Engagement, interest, enjoyment	Responses in brief surveys or interviews regarding why they participated, what they liked, etc. Appropriate science interest assessments from ATIS.
Understanding of content	Content knowledge assessments such as the one used for the *The Amazing Nano Brothers Juggling Show*.
Intentions towards future involvement or activity	Participant information seeking, verbal descriptions of plans or ambitions, responses on brief surveys regarding what they might do differently based on participation.
Impact on behavior and attitudes	Follow-on interviews or surveys regarding behaviors and attitudes specific to the outreach goals. Appropriate attitudinal assessments from ATIS or other sources.

B. Websites, videos, broadcasts and other media-based resources

Outcome indicator	**Measurement approach and instrumentation**
Engagement, interest, enjoyment	Data analytics, posts and comments, responses on linked surveys or online forums regarding why they participated, what they liked, etc.
Understanding of content	Data analytics, posts and comments, responses on linked surveys regarding what they learned.
Intentions towards future involvement or activity	Participant information seeking, registering on sites, liking pages or posts, reposting online, responses on brief surveys regarding what they might do differently.
Impact on behavior and attitudes	Follow-on interviews or surveys regarding post-experience activities, reposts (e.g., twitter), information seeking, registering for sites.

C. Ongoing programming in afterschool, museums, or public settings

Outcome indicator	**Measurement approach and instrumentation**
Engagement, interest, enjoyment	Responses in surveys or interviews regarding why they participated, what they liked, etc. Appropriate science interest assessments from ATIS.
Understanding of content	Content knowledge assessments carefully aligned with the experiences and objectives of the programming.

Intentions towards future involvement or activity	Responses in surveys or interviews regarding choice of activities, courses, or careers. Appropriate science attitude assessments from ATIS or other sources.
Impact on behavior and attitudes	Follow-on interviews or surveys regarding behaviors and attitudes specific to the outreach goals. Administration of Appropriate attitudinal and behavioral assessments from ATIS or other sources.

Concluding comments

All programs invariably need to adapt their evaluation to their own goals, activities, resources, schedules, and sensibilities about the importance of certain types of evidence in making judgments about program quality. Efforts to provide plug-and-play evaluation tools are likely to leave a project leader frustrated at the lack of relevant data, since careful alignment between goals, plans, implementation, and measurement is essential for meaningful program evaluation.

This chapter has offered some general principles and guidelines, with pointers to the specific resources that project leads can use in developing their own data and interpretation for purposes of improving programming. It has also emphasized the need for principled approaches to conducting evaluation, working with professional evaluators when resources allow, and applying a full set of considerations to the questions about determining project success that should accompany chemistry outreach endeavors.

References

Baram-Tsabari, A. & Lewenstein, B. (2013). An Instrument for Assessing Scientists' Written Skills in Public Communication of Science, *Science Communication, 35*(1), 56-XX.

Bonney, R., et al. (2012). Principal Investigator's Guide: Managing Evaluation in Informal STEM Education Projects. (http://informalscience.org/evaluation/evaluation-resources/pi-guide

Frechtling, J. (2010). *The 2010 User-Friendly Handbook for Project Evaluation*.. Rockville MD: Westat, Inc. http://www.westat.com/Westat/pdf/news/UFHB.pdf

Gabel, D. (1998). The complexity of chemistry and implications for teaching. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 233-249). Great Britain: Klewer Academic Press.

Gamble, J. (2008). *A developmental evaluation primer*. Montréal: The J.W. McConnell Family Foundation.

Haran, B. & Poliakoff, M. (2013). How to measure the impact of chemistry on the small screen. *Nature Chemistry, 3*, 180-183.

Kozma, R. B., Chin, E., Russell, J., & Marx, N. (2000). The roles of representations and tools in the chemistry laboratory and their implications for chemistry instruction. *The Journal of the Learning Sciences, 9*(2), 105-143.

Nakhleh, M. B. (2002). *Some thoughts about molecular-level representations in conceptual problem solving.* Presented at Problem Solving in Chemistry: An Online CONFCHEM Conference on Chemistry. Available at http://www.chem.vt.edu/confchem/2002/b/nakhleh.html

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York, NY: The Guilford Press.

Penuel, W. R. (in press). A dialogical epistemology for educational evaluation. In W. R. Penuel & K. O'Connor (Eds.), *Learning research as a human science. National Society for the Study of Education Yearbook, 109 (1)*.

Schechter, R., Priedeman, M., & Goodman, I. (2010). *The Amazing Nano Brothers Juggling Show: Outcomes evaluation*. Available at: http://www.grginc.com/documents/GRGMOSReportFINAL10-15.pdf

Additional Resources

Guiding Principles For Evaluators. American Evaluation Association. 2004. Fairhaven MA: AEA. *http://www.eval.org/publications/guidingprinciples.asp*

Public Statement on Cultural Competence in Evaluation. American Evaluation Association. 2011. Fairhaven MA: AEA. www.eval.org

Evaluation Checklists. Western Michigan University. 2011. http://www.wmich.edu/evalctr/checklists/

Program Evaluation Standards: A Guide for Evaluators and Evaluation Users. Yarborough, Donald B., Lyn M. Shulha, Rodney K. Hopson, and Flora A. Caruthers. 2011. The Joint Committee on Standards for Educational Evaluation. Third Edition. Thousand Oaks CA: Sage Publications, Inc.

Glossary

Adapted from the IDA Science and Technology Policy Institute document *Touchstones of Evaluation Quality*

Bias: Unknown or unacknowledged error

Case study: A rich description and analysis typically using multiple modes of qualitative data collection. Case studies can also incorporate quantitative data

Changes in evaluation goals: Revisions to an underlying evaluation plan as a result of tracking time, resources, and performance schedules, and designed to improve efficiency and effectiveness of the evaluation

Coding: An analytical process in which data, in both quantitative form (such as questionnaires results) or qualitative (such as interview transcripts) are categorized to facilitate analysis

Comparison group: A sample or group of participants with characteristics (e.g., socioeconomic, demographic, geographic, etc.) similar to the experimental group, exposed to all of the services or conditions as the experimental group, except for the variable being tested

Control group: A sample or group of participants with characteristics (e.g., socioeconomic, demographic, geographic, etc.) similar to the experimental group, not exposed to any services or conditions but whose outcome is observed

Counterfactual: Refers to what would have happened in the absence of an intervention, in order to make causal inferences about the effect of the program

Cultural context: Attention to social, institutional, and cultural factors that may require special attention

Evaluation design: A plan for conducting an evaluation that specifies: (1) the set of evaluation questions; (2) the data that will be collected and analyses that will be undertaken to answer the evaluation questions; (3) the estimated costs and time schedule for the evaluation study; and (4) how the evaluation information will be disseminated and used

Evaluation goals: Goals or purposes of the evaluation in accordance with organizational values and goals

Evaluation questions: Specific questions that address an evaluation's overall goal, reflecting an awareness of differing points of view about the project

Experimental design: A design characterized by random assignment and generally the strongest study design for determining with a high degree of confidence whether the intervention alone caused observed outcomes

Experimental group: A sample or group of participants exposed to certain services or conditions for purposes of studying program outcomes

Formative evaluation: Systematic study conducted regularly or built into the activities to assess whether a project, activity, or grantee is reaching stated goals in order to guide ongoing improvements

Function: What an evaluation is expected to accomplish

Implementation evaluation: An evaluation of actual program activities, typically to find out what actually happened or is happening to the program, and in some cases to improve the program implementation

Intended uses: Specific ways in which information generated from the evaluation will be applied

Instruments: Tools used to collect data, commonly including survey questionnaires, interview protocols, and focus groups, among others

Integrity/honesty standard: Evaluators display honesty and integrity in their own behavior, and attempt to ensure the honesty and integrity of the entire evaluation process

Justification of data collection: Rationale for expending time and resources to collect information within the evaluation plan

Limitations of study design: Evaluation assumptions, theories, methods, measures, or analytic techniques that could significantly affect or limit the interpretation of findings

Logic model: A graphical representation or tool that summarizes key elements of a program, including resources and other inputs, activities, outputs (products and services delivered), and intermediate outcomes and end outcomes (short-term and longer-term results) that the program hopes to achieve

Management plan: Statement that defines the approach to a project; how a project will be conducted, monitored, and controlled, how the interests of key stakeholders will be addressed, and steps that will be taken to assure that the intended product will be delivered on time and within budget

Measurable indicators: Quantitative or qualitative factors or variables that provide a simple and reliable means to measure achievement, to reflect changes connected to an intervention, or to help assess the effectiveness of an intervention

Outcomes: Desired effect of an education effort (e.g., improved understanding of a STEM concept, changes in instructional practices, or an increased number of students pursuing STEM degrees or careers)

Outputs: Direct products of education efforts, usually measured in terms of the volume of work accomplished or the number of classes taught, people served, or educational materials distributed; outputs are not ends in themselves, rather they are products that lead to desired outcomes

Phase of the evaluation: Refers to design, data collection, analysis, and reporting stages in the evaluation process

Pre-post design: An assessment design that compares outcomes before and after the program treatment

Professional and ethical principles: Standards evaluators follow to conduct sound evaluations, addressing such concepts as systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare

Program: A set of resources and activities directed toward one or more common goals, typically under the direction of a single manager or management team

Program effects: Effects of programs on participating populations

Project context: Consideration of geographic location, social, historical, and cultural conditions, political climate, and how long the project has been in operation

Quasi-experimental design: A design characterized by minimal control over the allocation of treatments or other factors being studied; this design often uses matched comparison groups or measurements at successive time intervals (time series analysis)

Randomized experiment: An experiment that randomly assigns units to different conditions

Reliability: Consistency of a measurement technique or instrument, often used to describe a test

Response rates: Refers, for example in survey research, to the number of people who answered the survey divided by the number of people in the sample. It is usually expressed in the form of a percentage.

Responsibilities for general and public welfare standard: Evaluators articulate and take into account the diversity of general and public interests and values that may be related to the evaluation

Sampling methods: That part of statistical practice concerned with the selection of a subset of individuals from within a population to yield some knowledge about the whole population, especially for the purposes of making predictions based on statistical inference

Stakeholders: Individuals, groups or organizations who represent those populations served by the program and who have an investment in the program, its outcomes and its evaluation. Stakeholders should be involved in the project from the framing of evaluation questions through disseminating results

Summative evaluation: Systematic study of a program at its conclusion; conducted to determine the success or effectiveness of a program, the impact on participants, cost effectiveness, and whether the program should be repeated or replicated

Systematic inquiry standard: Independent and data-based inquiries

Theory of change: A conceptual tool that delineates the intentional processes linking project activities and described outcomes; a well-articulated theory of change specifies short-term, intermediate, and long-term impacts

Unbiased: Without bias or prejudice; impartial; not affected by any extraneous factors, conflated variables, or selectivity which influence its distribution

Unit of analysis: The major entity that is being analyzed in an evaluation

Validation plan: A set of plans or activities to test the validity of a proposition, inference or conclusion

Valid: Best available approximation to the truth of a given proposition, inference, or conclusion; types of validity include construct, external, and face validity.