Room for Rigor: Designs and Methods in Informal Science Education Evaluation

Alice C. Fu
Archana Kannan
Richard J. Shavelson
Lisa Peterson
Amy Kurpius
Abstract

We examine methodological characteristics of summative evaluations in informal science education (ISE), asking: What are the major types of designs used in summative evaluations, and what kinds of questions can they answer? What are the types of data collection methods and measures used, and how many are self-reports or direct measures? We reviewed all summative reports published on informalscience.org in the year 2012 and found an over-reliance on non-experimental evaluation designs and heavy use of self-report instruments. If a primary function of summative evaluations is to assess impact, and impact is a causal question, then these findings are problematic; the field needs to move beyond the mostly descriptive studies found in our sample. Interviews with nine leaders in ISE and ISE evaluation help explain evaluation challenges in ISE and generate ideas for advancing the field.

*Keywords:* summative evaluation, informal science education, evaluation methods, experiments, quasi-experiments, self-reports
There is growing recognition that evaluation in informal science education (ISE)\(^1\) needs strengthening. At the same time, there is recognition that the characteristic complexity of ISE poses major evaluation challenges:

… informal learning is individualized, complex and multifaceted. Impacts depend on personal, physical, social, and cultural contexts and may not become evident until sometime after the experience. In addition to these types of factors, audiences may be highly heterogeneous. For these reasons, impacts are typically more complicated to assess than in the classroom, especially since the most important learning outcomes may be non-cognitive and more difficult to measure. (Friedman, 2008, p. 12)

Further, evaluators strive to conduct studies that preserve the free-choice nature of ISE experiences, all in the context of a relatively new field that is still working to develop appropriate measures and define common language, goals, and theories (National Research Council [NRC], 2009). Coupling these challenges with constraints on time, money, and capacity, it is no wonder that it is difficult to carry out meaningful, reliable, and feasible evaluations (Falk et al., 2012; Friedman, 2008; NRC, 2009).

The evaluator’s charge is to sort through these challenges and provide useful, evidence-based information. Allen et al. (2007) present two major approaches for doing so: “reducing complexity to identify causal relationships, and embracing complexity to create a deep understanding from multiple perspectives” (p. 229). The two approaches are not mutually exclusive; in fact, both may be used in the same evaluation. However, ISE evaluations have

\(^{1}\) Informal Science Education (ISE) refers to various out-of-school science, technology, engineering, and mathematics (STEM) learning experiences for children and adults such as museum exhibits, after-school programs, field trips, community and adult programs, media, games, websites, and so on.
tended towards the latter approach. The U.S. Department of Education [ED]’s Academic Competitiveness Council reviewed evaluations of federally-funded ISE programs and concluded, “it is extremely challenging to carry out rigorous studies that can identify causality in [ISE] programs” (ED, 2007, p. 26).

The purpose of this paper is to delve into the methodological characteristics of ISE evaluations. Specifically, our goal is to extend the findings of previous studies (e.g., ED, 2007; Falk et al., 2012) and investigate the types of designs and methods used in ISE summative evaluation reports. Our study questions are:

- What are the major types of designs used in summative evaluations, and what kinds of questions can they answer?
- What are the types of data collection methods and measures used, and how many are self-reports or direct measures?

To address these questions, we draw on evidence from a sample of evaluation reports published in 2012 on the Center for Advancement of Informal Science Education’s (CAISE) website—informalscience.org—a central online resource of the ISE community.

We are interested in the methodological characteristics of summative evaluations for two reasons. First, at a time of scarce funding, there is pressure for accountability. Funders increasingly require informal science providers to demonstrate project impacts through summative evaluations (Friedman, 2008). Methodologically-rigorous summative evaluations not only provide empirical evidence of impact but can also justify the value of informal education to policymakers, taxpayers, and the broader community (Brisson et al., 2010; ED, 2007). Second, summative evaluations are increasingly seen as sources of knowledge for the field (Friedman,
We begin by discussing methodological rigor in summative evaluation, and then describe methods used to review reports and interview ISE leaders. Next, we present findings from the review of reports. We draw on insights from the interviews to elaborate findings and draw conclusions with possible next steps for addressing ISE evaluation challenges.

**Methodological Rigor in Summative Evaluation**

Summative evaluation is a form of systematic inquiry that may facilitate decision-making and knowledge-building. Summative evaluations should heed a set of overarching principles: explicitly define the questions to be addressed, build on prior literature and relevant theory, employ appropriate design and methods to provide accurate and trustworthy information, and clearly communicate information to stakeholders (e.g., Krathwohl, 2004; NRC, 2002). These principles are emphasized in various evaluation guidelines and standards (e.g., American Evaluation Association, 2004; Visitor Studies Association, 2008; Yarbrough, Shulha, Hopson, & Caruthers, 2011)). Although many aspects may comprise methodological rigor, we focus on design and measurement because we believe that they are especially critical to estimating the impact(s) of ISE interventions.²

**Evaluation Questions**

Evaluations typically begin with one or more questions that are of interest to stakeholders. The National Research Council (2002) categorized questions into three interrelated types: (a) descriptive questions, (b) causal effect questions, and (c) causal process or mechanism questions. The report included evaluation as a genre of scientific education research and

---

² Intervention refers to the ISE experience being evaluated and could be a program, exhibit, website, game, and so on.
recognized that a single evaluation may address multiple types of questions (p. 100). Evaluation designs and methods should be selected carefully for the type(s) of questions asked, not vice versa. We briefly review the three types of evaluation questions and various designs that are appropriate for each (adapted from NRC, 2002).

The first category, descriptive questions, includes descriptions of various kinds, such as understanding participants’ experiences, beliefs, and attitudes; characterizing programs and interventions; describing visitor demographics; and finding relationships among variables. For example, a descriptive question may ask, “What are the types of visitor behaviors at an exhibit?” Descriptive questions can be investigated using a range of designs including observational designs, ethnographies, case studies, post-only designs, and correlational studies.

The second category, causal effect questions, asks about cause-and-effect relationships: Does $x$ cause $y$? For example, stakeholders may wonder, “Does visiting a science museum (cause) increase student interest in science (effect)?” Randomized experimental designs are considered the “gold standard” for addressing cause-effect questions (Shadish, Cook, & Campbell, 2002), but when randomization is not feasible due to ethical, contextual, or logistical reasons, quasi-experimental and causal-modeling studies are used. Bear in mind that the identification of causal relationships does not rely solely on experimental studies. Causal work builds on theories, and exploratory and descriptive studies often provide the necessary empirical grounding.

The third type of question asks about causal mechanisms, the process by which $x$ causes $y$. For example, stakeholders may want to know, “How (by what mechanism) does visiting a science museum increase students’ interest in science?” Mechanism or process questions should be explicitly linked to theory and may be investigated using any of the above-discussed designs,
depending on whether the connection to theory is strong or weak. When the theoretical basis for understanding the mechanism is strong, experimental, quasi-experimental, and modeling studies can be used. When the theoretical basis is weak or poorly understood, mechanism questions become more descriptive and can be addressed using observational studies, ethnographies, case studies, design studies, and others.

**Study Designs**

We review experimental design basics and consider how various study designs measure up in addressing causal questions (Table 1).\(^3\) We do so at the risk of tedium but in the interest of establishing a shared terminology for the remainder of the paper.

A study design is a plan that guides decisions about when, how, and from whom to gather data; what data to gather; and how to analyze the data. In synthesizing the array of designs available for conducting summative evaluations, we group them into three categories based on their strength in establishing causality. The terminology we use is strictly around causality; randomized experiments are not stronger than non-experimental designs for all contexts and questions.

- Randomized experiments use *randomization and control groups*. When individuals are randomly assigned to different conditions, the control and treatment groups so formed are considered equivalent in all respects except for the treatment. Thus, any difference in the observed outcome between control and treatment groups is attributed to the treatment effect.

---

\(^3\) The labels and categories we use are subject to debate, both in ISE evaluation and in the larger field of evaluation and research methods. Lines between different types of designs can be fuzzy; different evaluators may categorize designs differently. Our goal is not to present an extensive argument for what constitutes randomized, quasi-, or non-experimental designs; rather, we use these categories to highlight strengths and weaknesses of various designs for making causal inferences.
• Quasi-experiments are designs without randomization but that use intact comparison groups and/or pre-post comparisons (Shadish et al., 2002). As a result, the comparison (control) groups are not likely to be equivalent to the treatment groups. Quasi-experiments do not control for all threats to internal validity but can still be used to make causal inferences when possible rival explanations are carefully eliminated. Stronger quasi-experiments use, for example, matched (in various ways) control groups, or both comparison groups and pre-post comparisons. Weaker designs use either non-equivalent comparison groups or pre-post comparisons.

• Non-experimental designs use neither comparison groups nor pre-post comparisons and should not be used to make causal claims. Instead, they seek to provide rich descriptions; build theories; and explore processes, mechanisms, and associations. They are suitable for descriptive questions and contexts where interventions are in early stages of development, theory is poorly understood, or variables are not easily manipulated. These include post-only designs, observational studies, case studies, document reviews, participant observation, and others (Friedman, 2008; Shadish et al., 2002).
Table 1: Study Designs

<table>
<thead>
<tr>
<th>Type of Design</th>
<th>Examples</th>
<th>Representation or Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized experiments: Designs that use control groups and randomization</td>
<td>Pre-Post with control group and random assignment</td>
<td>R O₁ C O₂ O₁ X O₂</td>
</tr>
<tr>
<td>Post only with control group and random assignment</td>
<td>R C O X O</td>
<td></td>
</tr>
<tr>
<td>Quasi-experiments (stronger): Designs that use matched control groups, or control groups and pre-post comparisons, but NO randomization</td>
<td>Designs with matched control group</td>
<td>NR MC O MX O</td>
</tr>
<tr>
<td>Pre-post with non-equivalent comparison groups</td>
<td>NR O₁ C O₂ O₁ X O₂</td>
<td></td>
</tr>
<tr>
<td>Quasi-experiments (weaker): Designs that use control groups or pre-post comparisons, but NO randomization</td>
<td>Post only, with non-equivalent comparison group</td>
<td>NR C O X O</td>
</tr>
<tr>
<td>Pre/post, no comparison group</td>
<td>O₁ X O₂</td>
<td></td>
</tr>
<tr>
<td>Post with independent pre</td>
<td>NR O₁ C X O₂</td>
<td></td>
</tr>
<tr>
<td>Non-experimental/Pre-experimental designs: Designs that do NOT use control groups, pre-post comparisons, or randomization</td>
<td>One shot post only</td>
<td>X O</td>
</tr>
<tr>
<td>Ex-post facto designs (correlational, criterion group designs)</td>
<td>The investigation starts “after the fact.” The “treatment” has already been administered; the researcher arrives afterwards and measures relationships between variables.</td>
<td></td>
</tr>
<tr>
<td>Observational and naturalistic designs</td>
<td>Ethnographies, participant- and non-participant observations, use of naturalistic methods such as interviewing participants during an ISE experience.</td>
<td></td>
</tr>
<tr>
<td>Case studies</td>
<td>Single or multiple designs and methods (qualitative and/or quantitative) to do an in-depth study of case(s) over time. Explore a new area, provide descriptions, understand a phenomenon in context.</td>
<td></td>
</tr>
<tr>
<td>Content analysis, document reviews</td>
<td>Systematic review of written, spoken, or visual information that can provide quantitative or qualitative evidence. Can involve reviews of documents, print media, television, video and audio recordings, websites, etc.</td>
<td></td>
</tr>
</tbody>
</table>

Note. This table is adapted from Campbell and Stanley (1963); Friedman (2008); Shadish et al. (2002); and Shavelson (1996). R=Random assignment to control and treatment groups, NR=no random assignment, C=control/comparison group, X=treatment group, M=matched on one or more attributes, O=measurement/observation (O₁=pre, O₂=post, where O may be one or preferably more measurements/observations, especially at pre).

Measurement Methods

Evaluators can develop or choose from a variety of qualitative and quantitative tools and techniques to measure their desired outcome(s). The choice should be well-justified; that is,
given available resources, evaluators should choose the methods that are most appropriate for the study questions. Whether qualitative or quantitative, methods should be sufficiently rigorous to measure the outcome of interest in a reliable, valid, and contextually-sensitive manner.

Methods range from self-reports to direct observations of behavior (Cronbach, 1990). Self-reports ask participants to rate or describe their own knowledge, interest, behavior, or other construct; whereas direct measures use direct observations of participant behaviors and experiences to assess outcomes. Examples of self-reports include Likert-type surveys, questionnaires, interviews, and focus groups. Examples of direct methods include tracking and timing, observations, and performance assessments. Using only two categories is a simplification, and one can imagine hybrids. Yet, there is value in identifying the source of evaluation data and recognizing when data are comprised solely (or mostly) of what people report about themselves and their experiences. Where possible, evaluations should seek multiple perspectives on the informal experience, thus facilitating the triangulation of findings.

Self-reports are relatively easy and inexpensive to develop and administer. They are also conducive to the informal context because they are relatively non-threatening (as opposed to traditional “tests,” which may make people uncomfortable). However, by requiring participants to stop and respond, these methods may interrupt or change the informal experience. Further, participants may over- or under-estimate their knowledge, interest, behavior, or other outcome; knowingly or not, they may try to please the evaluator with their response. Direct measures of outcomes are needed to supplement or validate self-reports. Ultimately, designs and measures do not by themselves make a study rigorous. In any evaluation, whether descriptive or causal, rigor denotes how well the study (design, methods, conduct) matches the purpose and justifies its responses to questions.
Impact and Causality

The purpose of summative evaluation is to assess the “outcomes or impacts” of a stable intervention (Friedman, 2008, p. 109), implying causality (Frechtling, 2010; Friedman, 2008; Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011). The NSF highlights this: “The task is not only to show that the [expected] outcomes occurred, but to make the case that the outcomes can be attributed to the intervention and not to some other factors. This is Shavelson and Towne’s second type of research question [cause-effect]” (Frechtling, 2010, p. 32).

Summative evaluations that intend to establish impact should focus on questions of causality and, if possible, causal mechanisms. Did the intervention produce any changes in outcomes, intended or unintended? What characteristics of the intervention created the impact? Of course, evaluations that address causal questions cannot ignore descriptive questions that establish a basic understanding of the intervention and its context. An ultimate goal for the field is to understand impacts and their mechanisms so that positive interventions and experiences can be replicated.

A natural question that arises is about the difference between research and evaluation. Different lines may be drawn between the two, but we believe that “good” evaluation looks like “good” research. That is, both need the best designs and methods available for answering the questions posed, although evaluation is primarily concerned with producing practical knowledge that informs decision-making and improvement, and research is primarily concerned with producing new and generalizable knowledge. The exploration of cause and effect is not strictly the purview of research, since evaluation stakeholders may reasonably wish to know what impacts can be attributed to the intervention in question.
However, reality poses complexities. Costs should be weighed against the potential utility of the evaluation information. It is reasonable to question whether a summative evaluation—and an investigation of impacts—is warranted or desirable at a given point in time. A rigorous examination of impacts requires that the intervention in question is well-established and stable, with strong foundational theories. If stakeholders are interested in simply reflecting on the intervention, gauging whether it has reached its objectives, and collecting information as the basis for discussions (but not for making claims or high-stakes decisions), a descriptive study might suffice. Similarly, if an intervention is small-scale or operates on very scarce resources, a costly evaluation of impacts may not be viable. Not all summative evaluations must address impact, and not all summative evaluations should use randomized experiments; but too few do in ISE, as we will argue in the remainder of this paper.

**Methods**

This study is descriptive. Our goal is to corroborate, with evidence, our and others’ impressions of current methodological trends and gaps in ISE summative evaluation. To address questions about the major types of designs and data collection methods used, we draw evidence from two sources: (a) a systematic review of summative evaluation reports from [www.informalscience.org](http://www.informalscience.org), and (b) interviews with ISE leaders.

**Review of Reports**

ISE has two main forms of shared literature: (a) studies published in peer-reviewed journals, and (b) non-peer-reviewed evaluation (and other) reports, also known as “grey literature” (Falk et al., 2012; NRC, 2009). In this study, we focused on the latter, given its potential reach and value to the ISE community. Conversations with practitioners, researchers, and evaluators told us that grey literature reports are accessed by members of the ISE
community; access to academic journals may also be cost-prohibitive. With few incentives for practitioners and evaluators in non-academic positions to publish in peer-reviewed journals (NRC, 2009), the grey literature is a viable and valuable option for sharing knowledge from summative evaluations. Such sharing is encouraged, and even mandated, within the community. For example, the NSF (2014) requires that all grantees in the ISE (now called Advancing Informal STEM Learning) program post their summative evaluation reports on the informalscience.org website.

We purposively sampled all summative evaluation reports published in 2012 on informalscience.org. We chose this database because (a) it is the repository for NSF-funded evaluations, a major funder of ISE summative evaluations; (b) it is searchable and provides free access to reports on a variety of ISE interventions; and (c) a major goal of informalscience.org is to support knowledge-sharing (CAISE, 2015a). The website is uniquely positioned as a central hub or "one-stop shop" in ISE (CAISE, 2013).

Our review of reports is limited in breadth. Informalscience.org is one of many online ISE databases; and some of the major evaluators who share their work outside of informalscience.org are not represented in our sample. Our review is not representative of the whole field, but it provides insights about the quality of summative reports found in a central database of the ISE field.

We searched the informalscience.org database using two parameters: (a) Type of resource: summative evaluation reports, and (b) Year: 2012. Our search yielded 36 results, and we reviewed all 36 summative reports. The research team developed a coding sheet,4 which included the following types of codes:

- Bibliographic including title, year published, authors, institution, and so on;

---

4 Coding sheet is available from the authors upon request.
• Descriptive including type of intervention, type of institution, and type of evaluator; and
• Methodological including type of method(s) used, timing of data collection, sample sizes, presence or absence of comparison/control groups, and whether instruments were included in the report.

Two researchers independently coded the evaluation reports, with 97% inter-rater agreement overall. Coding disagreements were resolved by discussion between the two researchers.

Sample characteristics. We coded the evaluated interventions into three major types: exhibits, programs, and crosscutting. Each evaluation report could be coded for more than one type of intervention. 27 of 36 (75%) reports evaluated only one type of intervention: exhibits were the largest group (33%), followed by cross-cutting (25%), and programs (17%). The remaining nine evaluations (25%) combined two types of interventions.

We also coded where the evaluations were conducted. Science museums and science centers comprised the largest group (15 evaluations), followed by universities (8 evaluations), and media companies (7 evaluations).

Most—30 of 36 (83%)—evaluations were carried out by external evaluators, 4 (11%) by internal evaluators, and 2 (6%) by a collaboration of internal and external teams.

Interviews with ISE Leaders

A second data source for this study was a set of qualitative interviews. We sought insights from leaders with experience and expertise in ISE and ISE evaluation. We interviewed

---

5 We adapted this typology from the NRC (2009). Exhibits are designed learning experiences in museums, science centers, nature centers, zoos, etc., where “artifacts, media, and signage are primarily used to guide the learner’s experience,” and “the nature of the learner’s interaction with the environment is often determined by the individual” (p. 48). We use “exhibits” as shorthand for both a single element and a group of related elements (commonly called an “exhibition”). Programs include structured and pre-planned learning experiences, events, and activities such as school field-trip programs, after-school programs, teen and youth programs, adult and community programs, and professional programs. Crosscutting refers to interventions that may cut across designed settings and programs; these include educational broadcast media, mass media, websites, games, and learning technologies.
nine individuals, all of whom were highly cited in the research/evaluation literature, active in national professional networks, and/or recommended to us by colleagues. Given our focus on design and methodology, three of the interviewees had conducted evaluation or research projects using randomized experiments in informal settings (Bowen, Greene, & Kisida, 2014; Gutwill & Allen, 2010; Sneider, Eason, & Friedman, 1979). Other interview participants included “practitioners” with past or current leadership positions at informal science institutions, evaluators and researchers (both internal and external to ISE providers), and funders.

Interview protocols were pilot-tested, revised, and customized. Topics in the semi-structured interviews included but were not limited to participants’ views on evaluation “best practices” and challenges for the field. Interviews were conducted in-person or by phone with two researchers and the interviewee; each lasted 60 to 90 minutes. All interviews were audio recorded and transcribed with participants’ informed consent. After each interview, both researchers completed a summary memo that synthesized main themes and findings. We analyzed the interview transcripts and extensive summary memos from every interview for emergent themes and themes relating to evaluation design and methods.

In what follows, we first report on the results from our review of summative reports. We then interpret these results with findings from our interviews.

**Results from Review of Reports**

**Few Causal Studies**

We examined the frequencies of the three types of study designs (Table 1) in the report sample. Since evaluations could use more than one type of design, we allowed for coding multiple designs per report. Of the 36 evaluations, 17 used only non-experimental designs, 5

---

6 We conducted “mini” case studies of these three randomized experiments; case study data included these and other interviews, and reviews of related reports.

7 Interview protocols are available from the authors upon request.
used only quasi-experimental designs, and 1 used a randomized experiment. The remaining 13 evaluations used a combination of quasi-experimental and non-experimental designs. Frequencies of various categories and subcategories of study designs are presented in Table 2.

Table 2: Frequencies of Three Study Designs Found in Summative Evaluation Reports

<table>
<thead>
<tr>
<th>Type of Design</th>
<th>Design Subcategory (Frequency in Parentheses)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized experiments</td>
<td>Post-only with control group (1)</td>
<td>1</td>
</tr>
<tr>
<td>Quasi-experiments:</td>
<td>Post only with non-equivalent comparison group (6)</td>
<td>18</td>
</tr>
<tr>
<td>Designs with comparison groups and/or pre-post comparisons; no random assignment</td>
<td>Pre/post (paired), no comparison group (6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Post with independent pre (unpaired pre/post) (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Retro pre/post (4)</td>
<td></td>
</tr>
<tr>
<td>Non-experimental designs:</td>
<td>Post-only designs (14)</td>
<td>30</td>
</tr>
<tr>
<td>Designs with no comparison groups or pre-post comparisons</td>
<td>Observational designs (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both post-only and observational designs (14)</td>
<td></td>
</tr>
</tbody>
</table>

Note. N=36 reports. Total frequency exceeds 36 because some reports were coded for multiple types of designs.

Very few summative evaluations can address questions about causal effects (Table 2). Only 1 of 36 reports (3%) used a randomized experimental design to explore the causal links between the intervention and its intended outcomes. Although 50% (N=18) of the evaluations used quasi-experimental designs, a majority of them were only strong enough for descriptive purposes and lacked the necessary rigor to address causal questions. Non-experimental designs were the most commonly found (N=30, 83%). Results suggest that the kinds of questions addressed in ISE summative evaluations are (or should be) mostly descriptive; the designs used rarely allow for causal inferences.

**Different Kinds of Comparisons**

More than half the studies in our sample used some form of comparison. When randomized experiments are not feasible, quasi-experimental designs can use comparison groups and pre-post comparisons, although comparisons alone do not establish causal impact.
Regardless of design, evaluators may also employ comparisons to internal or external benchmarks.

**Quasi-experimental designs.** Fifty percent of evaluations (18 of 36) used quasi-experimental designs (Table 2). Among them, we identified four types of comparisons: post-only with non-equivalent comparison groups (N=6), paired pre-post comparisons (N=6), unpaired pre-post comparisons (N=2), and post with retrospective pre (N=4) (Table 2). None employed both comparison groups and pre-post comparisons.

**Comparisons to benchmarks.** Non-experimental designs lack direct comparisons such as control groups and pre-post comparisons. Yet, it may be possible to provide some evidence of impact using comparisons to benchmarks or standards (Friedman, 2008). Internal benchmarks include data from within the institution such as projected goals; prior front-end, formative, and summative evaluations; and marketing data. External benchmarks refer to established standards from the research and evaluation literature: “Such benchmarks can provide at least some sense of the degree to which a project has been effective as an aid to learning” (Friedman, 2008, p. 32). Of the 30 evaluations in our sample that used non-experimental designs, 8 used comparisons to external benchmarks, and 1 used internal benchmarks.

Notably, 5 of the 8 external benchmarks were the tracking and timing standards established by Serrell (1998). These standards describe visitors’ use of exhibitions, based on data from 104 interpretive exhibitions of various types and sizes; see Serrell (1998, 2010) for metrics including Sweep Rate Index and % Diligent Visitors.

**Reliance on Self-Reports**

We categorized measurement methods into eleven types, as shown in Figure 1. A majority of evaluations (75%) used more than one type of method, so we allowed for coding
multiple methods per report (Table 3). Interviews (N=27 evaluations) and surveys (N=24) were most commonly seen (Figure 1). Observations were next most frequent (N=11), followed by tracking and timing (N=8), and document or website reviews (N=7) (Figure 1). Other methods such as mapping and sorting, performance assessments, tests, and focus groups were used in only a small number of evaluations.

Figure 1. Frequency of data collection methods used in summative evaluation reports. The x-axis represents the number of reports in which a particular method was found. Total frequencies exceed 36 because some reports were coded for multiple types of methods. The y-axis lists the 11 methods for which we coded. The “Other” category includes web analytics and secondary data analysis.

Figure 2 presents the types of measurement methods used in each of the 36 reports; the use of self-report surveys and interviews was nearly ubiquitous. Ninety-seven percent (N=35) used surveys, interviews, or both. When only one type of method was used, it was almost always (8 of 9 evaluations) a survey or an interview. With only one exception, other measurement methods were always used in conjunction with surveys or interviews.
Figure 2. Types of methods used in each summative report. Each row represents one evaluation report (N=36). Each column represents a different method of data collection. Shaded cells denote the particular methods (columns) coded in a given report (row); for example, report 10 used survey and interview, but no other data collection methods.

Across 36 reports, we counted 112 data collection instruments total. We were interested in how many were self-report versus direct measures. Self-reports included surveys, interviews, focus groups, and self-reported mappingsorting. Direct measures included tracking and timing, observations, performance assessments, and tests. As expected, a majority (N=77, 69%) of the instruments were self-reports; 26 (23%) instruments were direct measures.

---

8 Instruments may be used to capture outcomes such as interest, engagement, attitudes, knowledge, behavior, and skills. Unfortunately, a detailed analysis of outcomes matched to measures was beyond the scope of this project.

9 We adopted these basic categories to distinguish between data that represent participants’ own reports (self-reports) and data that come from other sources (direct measures, other). This approach streamlined our coding but...
Discussion: Next Steps in ISE Summative Evaluation

Our review of ISE summative evaluation reports found heavy reliance on non-experimental designs. Such evaluations address mostly descriptive questions; their data do not support most cause-effect and causal mechanism inferences. We also found heavy use of self-report measures, many of which lacked reliability and validity evidence. Regardless of design and methods, studies often lacked an explicit chain of logic linking questions, design, measurement, data analysis, results, discussion, conclusions, and recommendations. Taken together, these limitations can be problematic for the field, especially when it comes to making a case for the impact and value of ISE to policymakers, funders, and other decision-makers. In what follows, we synthesize our interview findings, current literature, and our own knowledge and experience to address some of the challenges in ISE evaluation. These recommendations should be interpreted cautiously as informed, but untested, conjectures.

What We Know: Invest, Plan, Balance

Some of the most critical steps for advancing the field are commonly known and were reiterated by many interviewees. We briefly review some of these arguments: invest resources in rigorous evaluations; plan ahead for evaluation and integrate it into projects; and balance methodological rigor against sensitivity to the informal context.

Invest resources. It is commonly felt that experimental and strong quasi-experimental designs are resource-intensive; they are costly in money, time, and personnel. One interviewee described the challenge of using experimental designs for exhibit evaluations, because “it is costly to make different versions” of the exhibit; that is, it requires considerable resources to

---

10 The remaining 8% were categorized as "other": these were document reviews and secondary data analyses.
design reasonable alternative conditions against which the treatment can be compared (J. Gutwill, interview, July 25, 2013). Another interviewee described the general preference for investing in programming over evaluation: “If you give them a choice between a more elaborate evaluation or more elaborate exhibit, many funders will pick the more elaborate exhibit every time” (A. Friedman, interview, August 26, 2013). Without sufficient resources, conducting rigorous evaluations can be challenging.

**Plan ahead.** Planning ahead may be critical to effectively leveraging evaluation resources. Instead of separating programming from evaluation, evaluation can be considered integral and interrelated with program planning and implementation (Frechtling, 2010). Funders need to see methodologically-rigorous evaluation as a built-in means of improving practice, informing decisions, and possibly advancing the knowledge of the field. The two grantmakers we interviewed agreed on the need for evaluations that can reliably assess program impact and inform funders’ decision-making (R. Ottinger, interview, October 7, 2013; C. Tang, interview, September 19, 2013).

**Balance rigor and contextual sensitivity.** Informal environments, by definition, provide learners with free-choice experiences, and it is essential that evaluations do not undermine this. Activities associated with experimental designs, such as assigning individuals to treatment and control groups, and controlling experiences to isolate effects, may run counter to expectations about informal learning. When some choice is lost, evaluators can maximize other free-choice aspects of the experience. For example, participants in the Star Games evaluation were assigned to control and treatment groups but encouraged to spend as little or as much time as they desired with the exhibits (A. Friedman, interview, August 26, 2013). Outcome measures should be conducive to the setting; for example, museum visitors may find formal assessments
jarring. The Director of Education Programs at Monterey Bay Aquarium summarizes, “We want to try and make the evaluation activities match with the kinds of experiences that they’re having” (R. Bell, interview, May 16, 2013). A rigorous and authentic evaluation requires careful trade-offs between methodological rigor and contextual appropriateness.

**What We Seek: Rigorous Designs**

The U.S. Department of Education (ED, 2007) and the NSF (Frechtling, 2010; Friedman, 2008) view a chief purpose of summative evaluation as measuring impact or effectiveness—a causal question. Yet, in our analysis of summative reports, we found a gap in the types of designs used—mostly non-experimental, some quasi-experimental, and nearly no experimental designs—and, thus, a gap in the types of questions that can be justifiably addressed.

In ISE evaluations, quasi-experiments are often employed over randomized experiments because they may be easier to implement in free-choice environments. Whether quasi-experiments adequately address causal questions depends on how well they minimize threats to internal validity. For example, the Exploratorium’s Director of Visitor Research and Evaluation described a quasi-experimental design that they commonly employ—different versions of an exhibit are tested on the floor from one week to the next—and illuminated the potential for selection effects:

...whoever walked into the museum, they are going to use one condition and then next week whoever walked into the museum, they are going to use the other condition. I didn’t get to randomly assign those people so it’s hard to be sure that they are similar groups. If you happen to get some bus tour of physicists one week, you are in trouble. (J. Gutwill, interview, July 25, 2013)
Without random assignment, existing differences in treatment and comparison groups could account for differences in outcomes. A quasi-experiment that uses both comparison groups and pre-post comparisons is a stronger design for a causal-effect question than a design that uses only one or the other (Table 1). Unfortunately, the quasi-experiments in our sample of reports were of the latter (weaker) kind, lacking the rigor necessary to eliminate alternative explanations and establish causal links.

Causal questions are generally best addressed by experimental designs (ED, 2007), which may be difficult but not impossible to carry out in informal contexts. For example, when multiple versions of an intervention can be easily switched out, such as with some exhibits or labels, one could randomly reveal a different version to each approaching visitor (Allen et al., 2007). In general, it is prudent to focus experimental studies on just a few carefully-chosen outcomes and possible causes (Allen et al., 2007).

Random assignment is a chief challenge of conducting experimental evaluations. We highlight two strategies to increase its feasibility and palatability.

**Look for conditions conducive to lottery.** At the Crystal Bridges Museum of American Art, a large-scale experimental design was used to investigate the effects of school field trips on students' critical thinking skills, historical empathy, tolerance, and interest (Bowen et al., 2014; Greene, Kisida, & Bowen, 2014). The opening of a new art museum in the region, coupled with the offer of free school tours, created an overwhelming demand: in its first two semesters, the museum received applications from 525 school groups representing over 35,000 students (Greene et al., 2014). One study author explained that this presented a natural opportunity to employ random assignment:
They literally could not handle that many kids. It just was logistically impossible. There had to be some way to ration who was going to get these visits. The typical way that these things would have been rationed is...first-come, first-served. But instead, when you have something that has to be rationed and you want to do a study, you can do a lottery. It's fundamentally fair...the other thing that a lottery allows you to do is to have a control group. The people that didn’t win the lottery, they are the perfect comparison group to compare the treatment group to, because the only thing that makes them different is that they lost the lottery. (B. Kisida, interview, October 8, 2013)

The Crystal Bridges team used a stratified randomization procedure, first creating pairs of school groups that were matched on demographic characteristics and then randomly assigning groups within each pair to treatment or control conditions.

The Crystal Bridges program exemplifies a classic situation where “demand for service outstrips supply” (Shadish et al., 2002, p. 269). The degree of its popularity might appear exceptional, but oversubscribed exhibitions, shows, and programs, commonly use a first-come, first-served ticketing system. Without detracting from the informal experience, such a system could be replaced with a lottery that incorporates random assignment for evaluation purposes.

Defer instead of deny. A common reason against random assignment is unwillingness to deny the intervention to potential participants. When the intervention can be administered repeatedly or in stages, random assignment can determine who receives the intervention first (treatment group) and who receives it later (control group). In the Crystal Bridges study, control group students were guaranteed tour spots for the following semester: “...the experience wasn’t
taken away from them, it was just deferred” (B. Kisida, interview, October 8, 2013). Examples of other informal experiences that may naturally be administered repeatedly or in stages include multiple entry times for popular exhibitions and shows, staggered release of educational technologies, or waiting lists for checking out science kits and curriculum materials. All participants eventually enjoy the (assumed) benefits of the experience, although some must wait their turn.

**What We Seek: Better Measures**

In our review of reports, the majority of measures were self-reports, often presented without validity and reliability evidence. This echoed the experience reported by many of our interviewees; several emphasized the need for better measures in ISE evaluation.11 There is an immediate need for ISE outcome measures that are valid, reliable, direct, and unobtrusive or embedded into the informal experience.

**Direct.** Self-reports provide an important means of documenting participants’ experiences from their own perspectives, but multiple perspectives offer a more complete picture of the experience being evaluated. Many interviewees questioned the quality of evaluations that rely too heavily on self-reports. In particular, if evaluation data are to inform decisions and future actions, self-reports alone are insufficient:

Summative evaluations that simply use self-reports of learners or visitors or people who receive the treatment are not all that reliable. So if in fact we are doing a summative evaluation for the purpose of knowing whether to do more exhibits or programs like this, or whether to change directions, or whether to recommend this strategy to someone else,

---

11 These included but were not limited to S. Allen, A. Friedman, J. Gutwill, R. Ottinger, and B. Serrell.
then I think it is important to have really good data, by which I mean more rigorous than just surveys and self-reports. (A. Friedman, interview, August 26, 2013)

Friedman was a trustee of the Noyce Foundation, and the foundation’s Executive Director elaborated on how they began looking beyond self-report data from their programs:

…that’s one piece of critical friend feedback that our board gave to us: just having self-report survey data is not sufficient…you need to have a check on that. [In one evaluation] they had both observational data and…self-report, and the gap was significant. So that led us to say, hmm, we can’t just accept what the kids are saying. We need to know what the instructors are saying and their own observations, as well as the observations of third-party observers. So that’s a big lesson. (R. Ottinger, interview, October 7, 2013)

Ottinger acknowledged that collecting third-party observational data can be expensive and beyond the capacity of some providers. This recognition spurred the foundation to fund efforts to find feasible solutions, such as testing whether they can forgo third-party observers and instead use program staff to reliably observe and rate their own programs (Shah, Robertson, Carvalho, & Noam, 2014).

Direct measures help corroborate and sometimes contradict self-reports (e.g., Wojcik, Hovasapian, Graham, Motyl, & Ditto, 2015). The most common direct measure in our sample was tracking and timing, which Serrell described as a “reality check” on exit interviews:
If people tell you in an exit interview, “I read everything,” and then you look at your tracking and timing data, and they didn’t stop at half of the exhibit elements that were there, how could you read everything if you didn’t stop? The first thing that has to happen is that they have to stop and pay attention. No attention, no engagement, no learning, no effect. (B. Serrell, interview, August 29, 2013)

Interviewees described other direct, behavioral measures. For example, in a summative evaluation of an astronomy exhibition, evaluators developed “Raffle Choice” to measure students’ interest in astronomy (Sneider, Eason, & Friedman, 1979). Students received a raffle ticket for a book and selected one choice from four options (two astronomy-related and two non-astronomy books); to reduce bias, students were left alone to make their choices. The number of astronomy books chosen was used as an indicator of student interest in astronomy. Friedman explained why they developed this measure:

I know we considered just asking people how interested they are in astronomy. So that would be a self-report. And then we decided we didn’t believe what they’d say. Because the people who we’d just shown an astronomy exhibit would think we wanted them to say they were interested in it. So there is [the] phenomenon of trying to please the interviewer, maybe to get the interview over as soon as possible, or just because you appreciated their interest in what you thought and so you want to be nice….So we thought we were getting a less biased sample. (A. Friedman, interview, August 26, 2013)
Another example was in the Crystal Bridges evaluation (Kisida, Greene, & Bowen, 2014). The study authors corroborated survey data by showing students in the treatment group were more likely to return to the museum:

…You have to take survey responses with a grain of salt, because people will lie to you and people will tell you what they [think you] want to hear….so we also gave everybody a coupon for a [free] return visit to the museum. And we coded those coupons. And then when the museum collected them, we were able to measure whether the treatment group or the control group came back at higher rates. Any time you can get a behavioral measure, it’s easier to believe than a self-reported measure. (B. Kisida, interview, October 8, 2013)

**Unobtrusive.** One challenge of ISE evaluation is meeting expectations of a free-choice experience (discussed earlier). Rather than disrupt the experience, unobtrusive measures can be embedded as a seamless part of informal learning.

Camargo and Shavelson (2009) developed a performance assessment for students in a residential environmental education program. Students first receive a series of lessons and hands-on experiences at a designated stream to learn about water quality. While on a hike a few days later, students encounter a new stream, and the instructor asks them to likewise determine its water quality using a familiar set of materials. Instructors are trained to systematically observe, record, and analyze student actions, as well as review students’ notes and reports; these data are interpreted as direct evidence of student learning.
At the Chicago Children’s Museum’s Skyline exhibition, visitors to the Skyscraper Challenge exhibit try to build tall and stable structures within a designated period of time (Randi Korn & Associates, 2008). While they build, a computer takes photographs of their progress. Visitors then select six of their favorite photographs and audio-record responses to a standardized set of questions (one question per photograph). Visitors can view their narrated “books” online using a unique code. Summative evaluators recognized this “photo-narrative” experience as an opportunity for an unobtrusive measure, although it was not designed as such. With visitors’ permission, evaluators analyzed the narrated books for descriptions of the building process, problem-solving strategies, self-reported learning, and use of STEM-based language and concepts (Randi Korn & Associates, 2008).

Digital technologies offer innovative opportunities to integrate assessments with learning. Schwartz and Arena (2013) argue for educational assessments that focus on “choice” as the outcome: “…interactive assessments can evaluate students in a context of choosing whether, what, how, and when to learn” (p. 3). Assessments can capture “specific constellations of choices relevant to learning” (Schwartz & Arena, 2013, p. 18). For example, a game about mixing colors automatically logs critical choices made by students while they play; these logs are analyzed for evidence of different learning processes, such as whether “students are trying to solve each problem in turn rather than discovering the general principle that governs the solutions to all problems” (Schwartz & Arena, 2013, p. 23). This and other types of embedded assessments in digital experiences are at the cutting edge of efforts to develop unobtrusive research and evaluation measures in both formal and informal education (e.g., Schwartz & Arena, 2013; Shute & Ventura, 2013; Zapata-Rivera, 2012). As pioneers resolve technical and practical challenges,
including cost-efficiency and scalability, these “stealth” measures may become an increasingly viable means of evaluating ISE outcomes.

Common. Interest in developing common or standardized measures is evident in various ISE-related sectors, including citizen science (Cornell Lab of Ornithology, 2015), out-of-school time STEM (Program in Education, Afterschool & Resiliency, 2014, 2015), science museums (Museum of Science, Boston, 2014), and learning more broadly (Learning Activation Lab, 2015). The use of shared measures was one of three dominant themes that emerged from a 2013 CAISE convening on building evaluation capacity in informal STEM (Ellenbogen, 2014); CAISE (2015b) recently hosted an online “sharable measures” forum for those interested in discussing issues and exchanging resources related to the topic.

Participants in our interviews, including practitioners, funders, and evaluators, similarly expressed interest in measures that facilitate comparisons across projects and over time. For example, grantmakers lamented that there are “a lot of one-offs” (C. Tang, interview, September 19, 2013) and “too much individual program evaluation” (R. Ottinger, interview, October 7, 2013). An evaluator observed that “fragmentation” and “lack of comparability” “cripples our ability to make claims as a field” (S. Allen, interview, October 24, 2013). This same person highlighted the responsibility of evaluators to bring their experience from a range of projects, relate what is “typical” or not, and enable appropriate comparisons (S. Allen, interview, July 10, 2013). A museum leader characterized evaluators as “broker[s] of knowledge” who can draw out “patterns” and “see things across the landscape” (E. Babcock, interview, June 20, 2013). Without good comparisons, it is hard to look across studies, see the big picture, or make meaning from evaluation results.
Serrell’s (1998, 2010) tracking and timing work is a classic example of a shared measure. Serrell relayed an anecdote from her evaluation of the Darkened Waters exhibition (Serrell, 1991), which provided the impetus for her standard-setting work:

When I showed and discussed the tracking and timing data [from Darkened Waters]…the people who were stakeholders…said, ‘Is that all?’ They were disappointed in the amount of time that people spent and the number of people that looked at different exhibits….I had to say, ‘Well, gee, I think that, actually, Darkened Waters people spent a lot more time and looked at a lot more stuff than other exhibits.’” (B. Serrell, interview, August 29, 2013)

To substantiate her impression, Serrell gathered tracking and timing data from other exhibitions and found that Darkened Waters was exceptional among its peers (B. Serrell, interview, August 29, 2013). Subsequent tracking and timing studies have used Serrell’s (1998) benchmarks as a convenient and cost-effective way to gauge results. One evaluator called these metrics “powerful” and elaborated: “If I hadn’t had Beverly Serrell’s study, then what would I have said? ‘Well, we got this fraction, it looks pretty good to me, but I don’t really know.’ So I think having benchmarks helps a lot to see how you’re doing compared with the rest of the field” (S. Allen, interview, July 10, 2013).

One challenge with common measures is the heterogeneous nature of ISE. ISE encompasses a spectrum of learning experiences for a range of audiences with a diversity of possible outcomes. One grantmaker cautioned that this diversity is a strength of ISE, and it may not make sense to “force fit” common measures across these varied experiences: “It shouldn’t be
the lowest common denominator” (C. Tang, interview, September 19, 2013). Tang further emphasized the need to think deeply about the constructs being measured and their definitions, and cultivate shared understanding of these complexities.

Common measures may facilitate informative comparisons and estimates of field-wide impact, but one-size-fits-all measures are problematic. The compromise may lie in developing measures for use within particular “buckets” or similar types of ISE interventions and outcomes; the technical soundness of each measure depends on well-defined constructs and evidence of reliability and validity in specific contexts, such as with particular types of learning experiences, target audiences, and so on.

As several contributors to the CAISE (2015b) forum emphasized, the success of any efforts to share measures and data depends on taking small steps first. There remains a field-wide need for common measures of basic demographic and descriptive variables such as participants’ age categories, social or family groupings, program types, dosage, and so on.

What We Seek: Technical Evaluation Expertise

A considerable amount of technical and practical knowledge is needed to conduct methodologically-rigorous and contextually-appropriate evaluations. Despite the availability of how-to manuals, workshops, guides, and standards, there remains a need for building evaluation capacity in ISE. We propose several strategies for enhancing the evaluation infrastructure.

Professional training. Those working in ISE draw from a diversity of educational and professional backgrounds. This disciplinary richness is a strength of the field (Ellenbogen & Grack Nelson, 2012), but multiple entry points means that individuals may lack common ground in their ISE or evaluation training. There are “a lot of methodologies and a lot of theoretical perspectives,” which make for “such a nuanced field, filled with all kinds of people problems
and embedded constructs that people don’t articulate” (E. Babcock, interview, June 20, 2013).

This can pose challenges for the development of ISE as a field:

So if I’m going to be an architect, I go to architecture school. If I’m going to be an economist, I get a degree in economics. If I’m going to be a lawyer, I go to law school. Where does somebody go to learn how to study informal learning environments? …There’s not a culture of ‘this is our research base, these are the textbooks we use, these are the methods that we use.’… [Informal learning is] very fragmented. (B. Kisida, interview, October 8, 2013)

Without compromising the richness of multi- and inter-disciplinarity, extended training programs focused specifically on ISE evaluation could enhance technical capacity. These could include degree programs with evaluation coursework and extended evaluation internships. Graduates would be trained in basic evaluation principles and also gain experience with the nuances of how these principles apply in the informal context. For example, at the University of Washington’s Museology Graduate Program, students specializing in evaluation must complete a yearlong evaluation project with a local museum (University of Washington, 2012). One interviewee emphasized the importance of mentoring new entrants to the field and exposing them to a range of work; beyond methodology, evaluators must learn “social skills and political nuances and ability to actually be a broker and a mediator and a strategic planner,” “negotiating client relationships,” and “business development” (E. Babcock, interview, June 20, 2013). Direct, hands-on experiences with doing evaluations—the apprenticeship model—is critical.
Short-term workshops and professional development courses may be helpful add-ons, but they cannot substitute for comprehensive formal training.

**Technical support.** ISE providers can team up with universities, researchers, evaluators, and others with technical expertise and interest in building knowledge for the field. In these mutually-beneficial collaborations, providers gain capacity for rigorous studies of their offerings and insights into their impactful design; researchers and evaluators gain access to sites, participants, and study opportunities (Callanan, 2012; Chesebrough, 2014).

Another model is that of a help desk or technical assistance center that provides just-in-time evaluation expertise tailored to individual problems. For example, NSF funded TEAMS, Technical Evaluation Assistance in Mathematics and Science, to build capacity and improve the quality of evaluations of its Math and Science Partnership (MSP) projects. TEAMS provides MSP evaluators, principal investigators, and others with a range of services including an online help desk, access to instruments and resources, direct help with employing experimental and quasi-experimental designs, assistance with instrument reliability and validity, review of evaluation plans and reports, and ongoing needs assessment (TERC, 2014, 2015). This type of direct support could be beneficial in ISE evaluation, particularly for those who are strapped for resources.

**Improved reporting.** Improvements in reporting could help position ISE evaluations as sources of learning for the field. The Building Informal Science Education (BISE) project is a leading example of efforts to improve and learn from ISE evaluation reports, having coded more than 500 evaluation reports, commissioned synthesis papers, and freely shared its database and coding scheme (Grack Nelson & Deng, 2014).
One indicator of quality reporting is careful language that provides an accurate account of the evaluation and its findings. In our review of reports, including reports of evaluations using non-experimental designs, we encountered overly causal and definitive language such as “as a result of,” “learned,” “effects,” and “effectiveness.” Such language reveals an assumption that the intervention caused the observed outcomes. What is the basis for attributing outcomes to the intervention being evaluated? How have intervening factors or counter-interpretations been ruled out? Authors often used measured language when reporting results, but we saw a tendency towards less restraint in the conclusion and recommendations sections. “Irrespective of the rigor of their study design, evaluators should be careful not to over-interpret their data or over-generalize their claims, lest they lead to misguided or simplistic policy decisions that may adversely affect learners in other settings” (Friedman, 2008, p. 33). Quality reporting requires evaluators to exercise caution in their interpretations, with caveats and explicit discussion of the study’s limitations.

One way to do this is to be explicit about the logic underlying the evaluation and the links from questions through findings and conclusions. In other words, clearly state the evaluation questions; and, for each question, delineate how data were collected and analyzed to provide answers. This seems obvious, but many reports do not articulate the links at this level of detail. This issue is not unique to ISE or evaluation, as highlighted by a recent article in *Science*: “Determining which question is being asked can be even more complicated when multiple analyses are performed in the same study or on the same data set…. each step in the analysis should be labeled according to its original intent” (Leek & Peng, 2015, p. 1315).
We found that a common pattern was to list evaluation questions, summarize methods, and then present a series of findings. Readers are left to tease apart the linking logic.\footnote{Our experience echoed that of Falk et al. (2012), who examined a purposive sample of 10 evaluation reports in ISE. Most of the reports failed to outline a study rationale, made no connections to prior work or relevant literature, did not articulate study questions, provided no justification for the choice of study methods, and were not transparent about their methods and instruments.} For example, we wondered:

- Which questions does each method address? (Methods may address different but overlapping questions and sub-questions.)
- How is each instrument administered? When, where, to whom? (Within a single evaluation, there may be several study designs; for example, some instruments might be used only once, others used in pre-post comparisons, and still others used with comparison groups.)
- What is the validity and reliability evidence for each instrument? (The quality of such evidence is often variable; instruments themselves are not always shared.)
- How are data analyzed and interpreted? (Data from multiple sources may be interpreted separately and together; and findings that do not map to the original evaluation questions may also be presented.)

To leverage summative evaluation reports as a source of learning about ISE, its impacts, and evaluation itself, reports should provide an accurate and thorough account of the methods and how they tie to questions and claims.

**Conclusion**

Our review of reports corroborated the common impression of over-reliance on a few evaluation designs and methods in ISE summative evaluation. If a primary function of summative evaluations is to estimate impact, and doing so requires making causal inferences and
measuring outcomes validly and reliably, evaluation resources need to be invested in these areas. The field needs to (a) plan for experimental and (strong) quasi-experimental studies that can address causal questions and balance out the mostly descriptive studies that currently exist; (b) develop better measures, particularly those that are direct measures of outcomes, embedded in the informal experience, or common to facilitate comparisons; and (c) enhance professional training, technical support, and standards of reporting to build an infrastructure of evaluation expertise.

Rigor is not denoted by the choice of particular designs or methods but by how the evaluation is carried out and the match to the evaluation purpose and questions. Getting this right requires technical evaluation capacity; in ISE, this encompasses both mastery of the “scientific principles” of evaluation and the artful (sensitive and responsive) application of these principles in informal contexts. Developing evaluation capacity necessitates developing conceptual clarity about the purposes and possible uses of summative evaluation; and about the constructs, outcomes, and interventions to be evaluated. There is room for rigor in ISE summative evaluation—for summative evaluations to better meet the purpose of examining project impacts and thereby contribute to the evidence base for the value of informal learning.
References


Ellenbogen, K., & Grack Nelson, A. (2012). *Evaluation under pressure: Balancing the needs of the ISE field with the needs of individual projects*. [Paper commissioned for Summit on Assessment of Informal and Afterschool Science Learning, convened by Board on


DESIGNS AND METHODS IN ISE EVALUATION


Shah, A., Robertson, D., Carvalho, K., & Noam, G. (2014). *Internal and external observer use of the Dimensions of Success (DoS) observation tool: A collaborative study in northern and southern California*. [Study conducted by PEAR at Harvard University and the School of Education at University of California, Irvine.]


TERC. (2014). *TEAMS Showcase: The TEAMS project has launched two new services for MSP Project Evaluators, PIs, Co-PIs, and Project Directors: Instrument reliability and validity assistance and a document review service*. Retrieved from http://teams.mspnet.org/index.cfm/27328


