EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS IN VISITOR STUDIES: A CRITICAL REFLECTION ON THREE PROJECTS

Scott Pattison, TERC

Josh Gutwill, Exploratorium

Ryan Auster, Museum of Science, Boston

Mac Cannady, Laurence Hall of Science

This is the pre-publication version of the following article: Pattison, S., Gutwill, J., Auster, R., & Cannady, M. (2019). Experimental and quasi-experimental designs in visitor studies: A critical reflection on three projects. *Visitor Studies, 22*(1), 43–66. https://doi.org/10.1080/10645578.2019.1605235

Abstract

Identifying causal relationships is an important aspect of research and evaluation in visitor studies, such as making claims about the learning outcomes of a program or exhibit. Experimental and quasi-experimental approaches are powerful tools for addressing these causal questions. However, these designs are arguably underutilized in visitor studies. In this article, we offer examples of the use of experimental and quasi-experimental designs in science museums to aide investigators interested in expanding their methods toolkit and increasing their ability to make strong causal claims about programmatic experiences or relationships among variables. Using three designs from recent research (fully randomized experiment, post-test only quasi-experimental design with comparison condition, and post-test with independent pre-test design), we discuss challenges and trade-offs related to feasibility, participant experience, alignment with research questions, and internal and external validity. We end the article with broader reflections on the role of experimental and quasi-experimental designs in visitor studies.

Keywords: Experimental design, quasi-experimental design, visitor studies, methods, museums, validity

Experimental and Quasi-Experimental Designs in Visitor Studies: A Critical Reflection on Three Projects

As in other fields, identifying causal relationships is an important aspect of research and evaluation in visitor studies. For example, investigators often seek to make claims about the connections between a particular museum program or exhibit and learning outcomes for participants or test program and design components that are hypothesized to influence these outcomes. Experimental and quasi-experimental approaches are powerful tools for addressing these types of causal questions and have long been popular in the field of educational research more broadly.

Although the use of experiments is not new in visitor studies (e.g., Bitgood, 1988; Bitgood, Patterson, & Benefield, 1988; Hirschi & Screven, 1988), the approach is still relatively rare compared to other methods. The prevalence of experimental and quasi-experimental designs and the associated challenges within the field was well described by Alice Fu and colleagues (Fu, Kannan, Shavelson, Peterson, & Kurpius, 2016). In their work, they examined all the summative evaluation reports submitted to the website informalscience.org in 2012. Summative evaluations often aim to assess the outcomes or impacts of a particular intervention. Therefore, we would expect that these reports would focus on making causal claims about the impact of the interventions they evaluated. Within the 36 evaluation reports the authors examined, many described using multiple methods to address their research questions. However, across all of the studies and methods described, they found only one that used an experimental design, 18 that used some form of quasi-experimental design that lacked sufficient rigor to address causal questions, and 30 that used non-experimental approaches. The lack of study designs that support

causal claims highlights, at least in part, the difficulty of conducting these types of studies in museums and other out-of-school environments.

One way to address this difficulty is by examining visitor studies that do support causal claims about impact and reflecting on how these studies also honor the nature of informal learning environments. In this article, we offer examples and reflections on the use of experimental and quasi-experimental designs in three different science museums to aide investigators interested in expanding their toolkit of research and evaluation methods[1] and increasing their ability to make strong causal claims about programmatic experiences or relationships among variables. Our goal is not to advocate for the use of these designs in all visitor studies but rather to highlight the affordances, theoretical perspectives, limitations, and design considerations of such approaches. Through this article, we aim to better familiarize the field with experimental and quasi-experimental approaches in order to position investigators to make informed decisions about when or if such approaches are useful for answering their questions.

We begin with a brief overview of experimental approaches, including core concepts underlying the testing of causal relationships and eliminating alternative explanations. We then present detailed descriptions of three designs used in recent visitor studies: (a) a fully randomized experiment, (b) a post-test only quasi-experimental design with strategies to avoid selection bias, and (c) a post-test with independent pre-test design intended to minimize the burden on study participants. For each of these examples, we discuss challenges and trade-offs related to feasibility, participant experience, alignment with research questions, internal and external validity, and eliminating alternative explanations. Finally, we end the article with broader reflections on the role of experimental and quasi-experimental designs in visitor studies.

Introduction to Experimental and Quasi-Experimental Designs

Experimental and quasi-experimental designs are best aligned with studies focused on cause-and-effect relationships, rather than studies about how or why some effect occurs, sometimes called mechanistic relationships. The desire to make claims about cause-and-effect is common among summative evaluations that seek to determine if programs met their intended outcomes and research studies that aim to test a particular theory or hypothesis rather than explore a new area where theory may be lacking. Aligned with these goals, experimental and quasi-experimental designs aim to determine what caused an observed effect. They do this by attempting to account of control for all potential causes in an environment other than the intervention in order to remove or reduce alternative explanations for an observed result. The logic is as follows: "We observed this effect, and since we eliminated all possible causes of this effect other than the intervention, the intervention must have caused the observed effect." Making such strong and definitive claims often requires creative study designs to eliminate alternative explanations.

To communicate these designs to others, a common notation is used to describe a sequence of events, including randomization, observations, and interventions. For example, the basic randomized experiment (Shadish, Cook, & Campbell, 2001) requires initial random assignment (R) of individuals into one of two conditions, the treatment condition that receives or participates in an intervention (X) and a comparison condition that does not. Finally, an observation (O) or assessment is made of the two groups to determine if there is a difference between them (Figure 1).

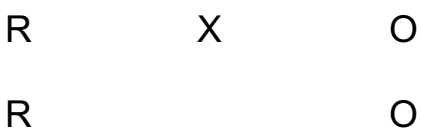

Figure 1: True experimental design with randomization (R) of participants into a treatment (X) group and a comparison group and observations (O) or measurements of outcomes in each group

The assumption underlying this design is that after the random assignment to the two conditions, the groups of individuals comprising the treatment condition are statistically equivalent to those in the comparison condition. Then the treatment condition participates in, or receives, the intervention while the comparison condition does not. Any differences later observed between the two conditions (e.g., science interest, dwell time, satisfaction) must have resulted from the intervention, since before the intervention the individuals in the two conditions were equivalent. This design is, quite literally, a textbook example of an experiment (Shadish et al., 2001) that eliminates all alternative explanations for a later observed difference between the two conditions. In practice, however, there are often many factors limiting our ability as researchers to conduct a study like this. For example, it is very challenging to randomly assign who visits museums and who does not, or even who visits an exhibit within a museum and who does not. Given that the choice to visit and the freedom to individualize the museum experience are fundamental characteristics of these settings (Falk & Dierking, 2013), attempting to control this dynamic can undermine the external validity of the research (Falk, Koke, Price, & Pattison, 2018). Instead, researchers often must be creative to work within these and other constraints to eliminate as many alternative explanations of later observed differences as possible while still attending to the informal and free-choice nature of the environments they are a part of.

Studies that are similar in purpose (i.e., looking for causal relationships) and structure (e.g., comparison groups, pre- and post-intervention data collection) to experimental designs but

lack random assignment are referred to as "quasi-experimental" (Campbell & Stanley, 1967). Quasi-experimental designs tend to generate less compelling support for observed causal effects, since it is very likely that the groups in each condition differed systematically at the beginning of the study, and often require more creativity and careful planning. As Shadish, Cook, and Campbell (2001) put it, "in quasi-experiments, the researcher has to enumerate alternative explanations one by one, decide which are plausible, and then use logic, design, and measurement to assess whether each one is operating in a way that might explain any observed effect" (p. 14).

Key considerations pertaining to all study designs, whether they are experimental or quasi-experimental, center on *internal validity* and *external validity*. Internal validity focuses on the extent to which the study itself has eliminated alternative explanations for an observed phenomenon. Said another way, internal validity is the strength of the claim that differences in interventions among the experimental conditions caused differences observed across groups. External validity, on the other hand, describes for whom and under what conditions we expect study results to have relevance. In other words, external validity is the strength of claims that research findings from a particular study can be expected to generalize to other settings, participants, or contexts.

Another consideration across experimental and quasi-experimental designs is the ability to detect a difference between two groups or between pre- and post-test conditions. *Statistical power* is the probability of detecting a difference between the two groups if one truly exists in the sample. This probability depends on two things: (a) the size of the difference and (b) the precision with which that difference is measured. In studies examining the impact of a particular program, there is a higher probability of measuring a difference if the impact is large. Since a

researcher may not be able to influence the size of the impact, the other way to improve statistical power is to increase the precision of the measured difference. This is done by using a high-quality measurement tool and increasing the sample size of the study. A *power analysis*, as described in the examples below, is a technique used to calculate the statistical power achieved in a study or, as is often done before a study begins, to estimate a necessary sample size to achieve sufficient statistical power (Cohen, 1988; Faul, Erdfelder, Lang, & Buchner, 2007).

Scholars have described numerous variations on experimental and quasi-experimental designs. In particular, researchers have explored different types of quasi-experimental designs, ranging from a post-test only study without a comparison condition to a complex array of comparison conditions and multiple tests and observations before, during, and after the treatments (Shadish et al., 2001; West, Biesanz, & Pitts, 2013). These variations differ in their complexity, how well they control for various threats to internal and external validity, and how feasible or appropriate they are to different research contexts and questions. For example, a post-test only design is often criticized for having many plausible threats to internal validity that undermine evidence for the causal impact of the treatment, such as history (the changes would have happened with or without the treatment) and selection bias (the sample represents a particular group that already exhibited the desired outcomes). However, from our perspective it is problematic to claim that one design is superior to another, since the choice of design depends on the context, research question, and plausibility of different validity threats (i.e., the likelihood that a particular threat is actually salient in a particular study or context). A pre- and post-test design without a comparison group might be a very appropriate (and efficient) approach if it is highly unlikely that the desired outcomes would occur without the treatment (e.g., very specific knowledge about a particular content domain), the entire sample of interest is accounted for in

the study (e.g., all visitors that enter a particular museum exhibit), and the pre-test instrument is unlikely to produce a testing effect (although this threat may be less easily explained away without a comparison group).

Examples of Experimental and Quasi-Experimental Designs

There are many resources to dive more deeply into these topics beyond the scope of this article (e.g., Allen et al., 2007; Brewer & Crano, 2013; Fu et al., 2016; Morgan, 2014; Shadish et al., 2001). Instead of providing additional explanation, we highlight three studies as examples of different strategies for working within real-world constraints while aiming to provide credible causal inferences and support internal and external validity. We encourage the reader to consider how each research study was designed to remove or account for alternative explanations of impacts, and how widely these results might have meaning beyond the subjects that participated in the study. In each example, we attempt to offer an honest account of the challenges and decisions that the teams wrestled with as they determined what design was most appropriate to achieve their research goals. Therefore, we also encourage the reader to attend to why the researchers made the design decisions that they did and how they navigated the tension between controlled conditions and naturalistic settings in their institutions.

The first study (described by the second author and principle investigator on the project)[2] most closely represents the textbook experiment described above. In this example, the researchers aimed to develop strong inferential evidence for a cause-and-effect relationship between a particular intervention, inquiry games, and visitor inquiry behaviors. The researchers used random assignment to place groups of museum visitors into one of four experimental

conditions. In this example, we see a clever use of comparison groups to narrow in on the specifics of the intervention, which gives the study strong internal validity.

The study (described by the first author and lead researcher and co-principle investigator on the project) highlighted in the second example sought to produce strong causal evidence of the impact of facilitation on family learning at exhibits while remaining true and representative of a naturalistic museum visit. This example highlights attention to participant self-selection bias across the study conditions, which is one often one of the principle threats to internal validity in quasi-experimental designs. It offers an example of a post-test only design that maintains a naturalistic museum feel. This approach, with no random assignment, has less internal validity than the previous study, but represents a more authentic museum experience, which may support stronger external validity claims.

The final study (described by the third author and lead researcher and co-principle investigator on the project) sought to determine the impact of facilitated engineering design challenges on youths' perceived success, self-efficacy, and attitudes toward engineering. The study also aimed to gather evidence on the mechanism (e.g., frequency of interactions, when interactions occur, and content of the interactions) through which the facilitation led to observed changes. This example highlights attention to creative comparison group selection to minimize the burden on research participants. This approach also lacked random assignment, making it quasi-experimental, but the study design demonstrates ways to address potential threats to internal validity and maintain the authenticity of the museum experience to support external validity claims.

While all three examples presented in this article were conducted in the context of science centers, we hope the descriptions and reflections are useful for professionals in other

contexts across the visitor studies field. We also acknowledge that the three examples were all part of large, federally funded projects and, like many experimental and quasi-experimental studies, involved considerable time and resources. In the conclusion section, we return to this issue and discuss ideas for implementing less expensive experimental studies.

Example 1: True Experimental Design

In 2005, the Exploratorium launched a National Science Foundation (NSF)-funded project—*Group Inquiry by Visitors at Exhibits* (GIVE)—that sought to create and study short programs to teach museum visitors how to engage in deep inquiry at science museum exhibits (Gutwill & Allen, 2010b, 2010a, 2012). The goal was to coach family and field-trip groups on how to ask and answer their own questions at interactive science exhibits. In order to investigate the qualities of such coaching, the GIVE team developed two programs, calling them Inquiry Games. One game, named "Hands Off!" encouraged individuals in each group to retain control and take turns leading the group in inquiry. The other game, "Juicy Questions," utilized a more collaborative pedagogy by asking group members to generate questions and choose one to pursue together.

The project set out to test three hypotheses about Inquiry Games:

1) *Group Inquiry can be taught.* Learning Inquiry Games will improve family and field-trip groups' science inquiry at novel exhibits, specifically groups' ability to ask relevant questions, conduct multiple, related experiments to answer those questions, and draw inferences that build on one another's observations.
2) *Facilitation matters.* Interacting with a friendly, knowledgeable museum educator will improve groups' inquiry practices. Even facilitation that does not provide inquiry

coaching could inspire groups to spend more time and pay greater attention at a novel exhibit.

3) *Pedagogy affects outcomes.* The particular pedagogy used in an Inquiry Game—supporting either individual or collaborative experimentation—will affect groups' inquiry practices (although the team had competing views about the relative merits of the two pedagogical approaches). The collaborative approach might support greater shared attention and thinking but feel too formal and school-like. The individualized approach may be easier to learn and empower quieter or younger voices to be heard, but also allow some group members to pay less attention to others' questions and experiments.

Study Design

To test these hypotheses and develop strong inferential evidence for cause-and-effect relationships between the Inquiry Games and groups' inquiry behaviors at a novel exhibit, the team employed a design called "multiple treatments and controls with pretest" (Shadish et al., 2001), assessing changes in groups' inquiry behaviors before and after learning and practicing an Inquiry Game (see Figure 2).

R O1 XHO O2 O3

R O1 XJQ O2 O3

R O1 CET O2 O3

R O1 CPC O2 O3

Figure 2: Experimental design for the GIVE study. R=Random Assignment; O1 = Pretest exhibit, X = Treatment, C = Control, HO = Hands Off, JQ = Juicy Questions, ET = Exhibit Tour, PC = Pure Control, O2 = Posttest exhibit and Exit interview, O3 = 3-weeks post interview.

To rigorously test the three hypotheses, the team created two treatment and two control conditions for the study:

- *Hands-Off Inquiry Game.* While group members play with the exhibit, anyone may shout, "Hands Off!" and take control of the exhibit. They may then either propose a plan of action for the group to implement or declare an observation or reflection. This condition embodies an individualized pedagogical approach.
- *Juicy Questions Inquiry Game.* After becoming familiar with the exhibit, the group stops and every member tries to generate a "juicy question," defined as a question that can be answered at the exhibit and to which no one knows the answer. The group chooses one question to pursue first, conducts experiments to answer it, and stops again to share observations and reflections. This condition employs a collaborative pedagogy.
- *Exhibit Tour Control.* Groups meet with an enthusiastic museum educator who focuses solely on explaining the exhibits' science content and development histories. No support is offered for doing inquiry.
- *Pure Control.* Groups use all exhibits without interacting with an educator.

Regardless of condition, each family or field trip group used four exhibits in a fixed sequence while being video recorded: (1) Pretest exhibit, where groups use the exhibit as they normally would, before an educator interacts with the group; (2) Coaching exhibit 1, where groups learn either an inquiry game or something else according to control condition; (3) Coaching exhibit 2, where groups practice whatever they had learned at coaching exhibit 1; and (4) Posttest exhibit, where the educator leaves the group and asks participants in treatment conditions to play the Inquiry Game they had learned. After using the four exhibits, one adult

and one child were randomly chosen from the group to participate in an exit interview and a delayed post interview 3 weeks after the experience.

The study randomly assigned each family or field-trip group to one of the four conditions. The team also blocked for educator, meaning that each educator facilitated equally across all four conditions (called a block) even though some educators facilitated more blocks than others.[3]

Design Trade-Offs and Validity Considerations

The team took several actions to reduce threats to internal validity, including random assignment of groups to conditions, pretesting groups on inquiry behaviors, and utilizing control conditions. Randomization helped the team draw causal inferences about the effects of treatment by increasing the likelihood that participating groups would have the same average characteristics across conditions at pretest (Shadish et al., 2001). Still, randomization was no guarantee that conditions would contain similar groups. (Indeed, by chance, fewer boys ended up in one of the treatment conditions in the study of families.) The inclusion of a pretest exhibit offered further protection against *a priori* differences in groups across conditions,[4] by allowing the team to compare changes in outcomes across conditions, rather than comparing only post-test outcomes.

Adding control conditions removed confounding effects that could undermine causal inferences about the impact of the Inquiry Games on group inquiry. For example, the Exhibit Tour Control condition accounted for the effect of facilitation—the potential influence of interacting with an educator on a group's motivation to conduct inquiry at a subsequent exhibit. To test for this effect, one of the planned comparisons analyzed the difference between changes

in outcome measures from groups in that condition to the changes in outcome measures of groups in the Pure Control condition. (No differences were found, effectively ruling out the effect.) The Pure Control condition also acted a baseline, protecting against the threat of practice effects: What if groups became more adept at doing inquiry simply by using more exhibits? Any pre- to post-test increases found in groups from the Pure Control condition could be used as a reference for all other pre- to post-test increases. (There were no such increases in the Pure Control groups, again ruling out the threat.)

Ensuring internal validity was expensive. Indeed, adding a pretest exhibit and two control conditions more than doubled the size and cost of the study. Randomization and blocking for educator, though fiscally inexpensive, required careful planning and significant additional effort by team members. However, the threats of non-comparable groups, facilitation effects, and practice effects were deemed important enough to justify the increased costs.

In regard to external validity, the team made several decisions that simplified the study but at the cost of potential generalizability of study findings. For example, the study employed only one type of hands-on, interactive exhibit: active, prolonged engagement (APE) exhibits designed to involve multiple users in self-directed exploration (Humphrey & Gutwill, 2005). This maximized opportunities for groups to ask and answer their own questions in the pursuit of scientific inquiry. However, this decision also narrowed the results to apply directly only to inquiry at APE exhibits. In another decision, the team exposed all participants to the same sequence of exhibits, rather than counterbalancing (swapping) the exhibit order, which would have required larger sample sizes in each condition to achieve the same statistical power. The study procedure was already complicated, involving random assignment to condition and blocking for educator. The team felt that counterbalancing the exhibit order could increase

human error. The fixed exhibit sequence reduced generalizability, because interaction effects between order and treatment (e.g., practice effects) were not assessed. Would the same results have surfaced had the pre- and post-test exhibits been swapped?

Finally, the study was held in a specially equipped laboratory space just off the Exploratorium's public floor in order to retain better procedural control. The lab was quiet, allowing for easy audio recording of groups' conversations, a necessity for analysis. Exhibit order could be strictly maintained by covering and revealing each exhibit at the proper moment. And groups could take as long as they wished at an exhibit, without concern for other museum visitors. The cost of these advantages, once again, came in the form of a loss in generalizability and external validity, or the degree to which the findings represented what might occur outside a research laboratory. Could the Inquiry Games be learned and applied in the buzzing setting of a science museum floor? By leaving this important question unanswered, the results of the study were limited to the "best case" scenario of a lab environment. Actually, the team attempted to improve external validity by subsequently teaching the more successful game, Juicy Questions, on the floor with families and field trip groups in a follow-up study (Gutwill & Allen, 2010b).

Practical Considerations

The project was expensive at 1.2 million dollars in 2005. The price tag came not only from experimental design issues like collecting and coding video data across four comparison conditions, but from developing and formatively testing multiple Inquiry Games (Allen & Gutwill, 2009). Designing the games to fit the constraints of a typical family or field-trip visit required a great deal of time and effort. Any game would have to be simple and memorable enough to be learned in about 20 minutes; accessible and appropriate for non-scientists with a

broad range of ages, interests, and prior knowledge; intrinsically enjoyable so groups would want to play on their own; and applicable to a wide variety of exhibit types and topics. The first game developed was none of these things. Excessively complicated, it focused on six different inquiry skills and assigned group members different roles such as questioner, experimenter and observer. Ultimately, the team reduced the number of skills to two and cut all the roles except facilitator, which was played by one adult in the group in both the Hands Off and Juicy Questions games.

Another practical problem arose in randomly assigning groups to condition while keeping the recruiter blind to that condition. The team was concerned that any foreknowledge of the next groups' assigned condition could unconsciously bias the recruiter as she sought out the group. This predicament was exacerbated by a desire to have educators double as recruiters so as to reduce project costs. How could the educator recruit a group without knowing their assigned condition, but then facilitate the appropriate condition once the group joined the study? The answer lay in having a research assistant (who was monitoring the audio/video feeds) randomly choose the condition for the next group while the educator was recruiting that group from the museum floor. When the educator returned to the lab with the group in tow, she found a note revealing the assigned condition and acted accordingly.

Example 2: Quasi-Experimental Post-Test Design with Comparison Condition

We now turn to an example of a quasi-experimental study that used a carefully designed comparison group[5] to help eliminate threats to internal validity from participant self-selection bias. The study was part of *Researching the Value of Educator Actions for Learning* (REVEAL)—a three-year, NSF-funded research study carried out by the Oregon Museum of Science and Industry (OMSI) between 2013 and 2017. In collaboration with TERC and Oregon

State University, the team explored the role of museum educators in deepening and extending family engagement and learning at interactive math exhibits. Although museum educators are a common component of the museum experience at many institutions across country, leading school group tours and classes, presenting demonstrations and stage programs, and facilitating learning for visitor groups at activities and exhibits, there is very little research directly measuring the impact of museum educators on visitor engagement and learning or identifying effective facilitation strategies in these contexts (Pattison & Dierking, 2013; Pattison et al., 2017). To support the work of museum educators and the growing number of professional development efforts for these individuals, the REVEAL project was intended to develop a model of facilitation for educator supporting family learning at interactive exhibits and to rigorously test this model using an experimental design.

REVEAL built on the NSF-funded *Access Algebra* project, which created a large traveling exhibit, Design Zone, capitalizing on visitors' interest in design, engineering, art, and music to create engaging and memorable learning experiences with math (Garibay Group, 2013). Specifically, the exhibit was designed to engage visitors, especially family groups, in algebraic thinking—a type of mathematical reasoning, similar to scientific inquiry, involving the exploration of mathematical relationships in the world around us and the use of these relationships to understand and create (Greenes & Rubenstein, 2008; Kaput, Carraher, & Blanton, 2008; Moses, 1999). The REVEAL project built on staff facilitation techniques and visitor learning measures piloted through the Design Zone project and used three specific exhibits that were designed to support staff facilitation for families around the topic of mathematics (Pattison, 2011). However, the overarching goal of the project was to explore staff facilitation strategies that might generalize across exhibits and content domains.

The REVEAL study consisted of two stages: (1) a design-based research (DBR) study with two expert educators (Pattison et al., 2017) and (2) a quasi-experimental study comparing facilitated and unfacilitated family interactions at exhibits (Pattison et al., 2018). During the DBR stage, a cross-disciplinary team of educators and researchers collected and analyzed data from hundreds of staff-family interactions over the course of six months. These efforts produced a model of staff-facilitated family learning at exhibits, including facilitation strategies for supporting mathematical reasoning and adapting to the needs and interests of different family groups. During the second stage, the team trained four new educators and conducted a quasi-experimental study to test the REVEAL facilitation model and assess the impact of staff facilitation on family learning across five distinct outcome variables: engagement time, intergenerational communication, visitor satisfaction, mathematical reasoning, and math awareness.

Study Design

The goal of the second REVEAL study was to produce strong, causal evidence of the impact of facilitation on family learning at the exhibits, compared to the comparison condition (i.e., internal validity), while at the same time ensuring that the study sufficiently mirrored naturalistic interactions between educators and families in science centers in order to be useful for practitioners (i.e., external validity). To do this, the team used a quasi-experimental design with two conditions: (a) facilitation and (b) greeting. In the first condition, trained educators provided full facilitation for families based on the REVEAL training they had received. In the second condition, educators simply greeted families as they approached and allowed them to engage with the exhibits on their own (see Figure 3).

NR T O

NR C O

Figure 3: REVEAL Quasi-experimental design used in the REVEAL study. NR=Non-Random Assignment; T = treatment condition with full facilitation; C = comparison condition with greeting only; O = videotaped observations and post-interaction surveys.

The team tested the impact of staff facilitation with four different educators using three different Design Zone exhibits. Data were collected almost every weekend day over the course of approximately 6 months, with the educators, exhibits, and study conditions rotated systematically by weekend day and morning and afternoon shift (taking care to ensure that educators, exhibits, and condition were equally represented during each time period). During each shift, one of the three exhibits was set up outside one of the museum's main exhibit hall and one educator was assigned to either greet visitors or facilitate learning for all groups that chose to approach the exhibit. Visitor groups were not actively recruited to participate in the study but instead were free to choose whether or not to engage with the exhibit, using a posted signage process of informed consent (Gutwill, 2003; Sindorf, Gutwill, & Garcia-Luis, 2015). For all eligible family groups that approached the exhibit, the team collected video and audio data of the exhibit interactions and post-interaction survey data from one adult visitor in the group in order to assess engagement time, intergenerational communication, visitor satisfaction, mathematical reasoning, and math awareness. Based on an initial power analysis (Faul et al., 2007), the team collected data from 263 family groups (171 in the facilitation condition, 92 in the control condition)[6] in order to reliably detect small effect size differences between conditions and medium effect size differences within the facilitation condition.

Design Trade-offs and Validity Considerations

In developing and implementing the research design described above, the REVEAL project team made several key design decisions with implications for feasibility, internal validity, and external validity. These included the initial decision to structure the study as a quasi-experimental design, rather than a true experiment, the selection and design of the comparison and treatment groups, and the intentional inclusion of different degrees of variability for important study context variables. Below we discuss each of these in turn.

The most significant design decision happened early in the project when the team decided to use a quasi-experimental design, without randomization of participants across conditions, rather than a true experiment. From the outset, a primary goal of the team was to make the study setting as authentic and naturalistic as possible so that results would be of practical significance to museums and educators (i.e., supporting external validity). As noted above, one potential threat to external validity is participant reactivity, or ways that the behavior of participants changes in response to the research context or perceived expectations from researchers (Brewer & Crano, 2013). Outside the field of visitor studies, this has been shown to be an important consideration, especially when the stakes are higher for participants (e.g., Eckmanns, Bessert, Behnke, Gastmeier, & Ruden, 2006; Kohli et al., 2009; Weiss, O'Mahony, & Wichchukit, 2010). Only a few studies in museum contexts have addressed reactivity (e.g., Serrell, 2000), and we still know little about its influence in these settings. One study (Pattison & Shagott, 2015) showed that reactivity had a strong influence on the outcomes of an experimental study looking at two versions of an exhibit, influencing both the magnitude of findings across both conditions (i.e., external validity) and the comparison of outcomes between conditions (i.e., internal validity). Prior research and piloting by the REVEAL team also suggested that studies of staff

facilitation at exhibits might be particularly susceptible to reactivity effects, since interactions with researchers could influence the natural role negotiation processes between visitors and educators (Pattison & Dierking, 2013).

Given this prior work, the team decided to eliminate this threat to external validity by not actively recruiting visitors into the study or randomly assigning them into different conditions Instead, the conditions were rotated over the course of the study, and visitors were allowed to self-select to approach the exhibits during each data collection session. This minimized interactions with researchers before visitors engaged with staff at the exhibits, but also had several other trade-offs. First, the team had to consider internal validity issues related to self-selection bias between the two conditions, which we discuss more below. This is one of the most common threats to internal validity in a quasi-experimental study and became a primary design consideration after the team chose a quasi-experimental approach. Second, since visitors were not actively recruited, the team had less control over which visitors participate in the study, what was known about visitors before they interacted with staff, and what happened to visitors after the interactions. The visitors that approached the exhibits possibly represented a particular group of museum visitors (e.g., visitors more interested in interacting with staff members), thus limiting the generalizability of study findings (i.e., external validity). No pretests could be conducted with visitors, which are often used to control for variability across participants and thus strengthen the statistical inferences made with the data. And because of the "hands-off" approach, the team did not follow up with visitors beyond videotaping the interactions and conducting post-interaction surveys. Therefore, the study did not shed light on how staff facilitation might impact subsequent exhibit experiences during the museum visit. The design did lower the burden on educators, however, since they did not need to switch between facilitation

and no-facilitation conditions in between each interaction, thus making the experiences more comfortable and potentially contributing to external validity.

The decision to use a quasi-experimental approach was closely linked to another decision: the final design of the comparison condition used to estimate the impact of staff facilitation on visitor learning at exhibits. A natural comparison group for the facilitation treatment condition described above would be visitors interacting with the exhibits without the presence of educators. However, this introduced a serious threat of self-selection bias (internal validity), since visitors who chose to approach the exhibits when an educator was present might differ systematically from those who approached without an educator. In fact, some prior research had suggested that certain visitor groups might be wary to engage with an educator and would prefer to experience the exhibits on their own (Marino & Koke, 2003). Therefore, to help eliminate this threat, the team designed a different comparison condition: "greeting," in which the educator started at the exhibit when visitors approached but then only greeted the group and allowed them to interact with the exhibit on their own.

Again, this decision created several trade-offs. As noted, the likelihood of self-selection bias was greatly diminished, since visitors walking by the study area would see the same set up in both conditions—an educator standing by an exhibit. Thus, the primary threat to internal validity was mitigated. However, the study was no longer a direct comparison of facilitated and non-facilitated interactions. Instead, the study measured the nature and outcomes of family interactions with educator facilitation compared to when an educator was present but not providing facilitation support. The team hypothesized that the actual impact of staff facilitation compared to no facilitation would be greater than any differences found in the REVEAL study, since the presence of the educator in the comparison condition might compel families to spend

longer and be more diligent in their use of the exhibits. Additional studies are required, however, to verify this hypothesis. Furthermore, the new comparison condition created several feasibility challenges. During data collection sessions with the comparison condition, the educator was required to be present and to do something very unnatural for them—greet visitors but not offer support, regardless of visitor needs. The team developed a protocol to ensure that educators could provide support if visitors demanded it and the interaction was becoming a negative customer service experience. This protocol was not ultimately required, but if these situations had arisen, they would have undermined the fidelity of the quasi-experimental design.

Finally, the REVEAL project team made a number of critical decisions about how variability was accounted and controlled for in the study, with the hopes of increasing the generalizability and external validity of findings. The REVEAL study hoped to shed light on the impact of staff facilitation on family learning at exhibits. However, the universe of possible situations involving staff and visitors at interactive exhibits is infinite. These interactions vary by type and focus of exhibit, staff member and facilitation approach, visitor and group characteristics, context of other exhibits around the interaction, noise and crowding within the area, time of day and day of week, and more. By using exhibits from a previous project, all focused on algebraic thinking, the team had avoided issues with problematic interactives, but they had already greatly narrowed the potential generalizability of the findings. To counteract this challenge and strengthen external validity, the team used three different exhibits and four different educators and collected data during morning and evening shifts on both Saturdays and Sundays over the course of six months. All of these variables were systematically rotated among the treatment and comparison conditions so that they would not become confounding variables, or alternative explanations to differences seen across the two conditions (i.e., internal validity).

In other words, these variables were both measured and controlled. Several other variables, such as visitor characteristics (e.g., number of prior visits, age, gender, education level, and language spoken at home) were free to vary but were measured so that they could be included in the analyses. And an infinite number of other context variables were free to vary but were not measured, such as weather and crowding. All of these variables collectively describe the context within which the REVEAL project was conducted and potentially constrain the degree to which findings represent what might happen in a different museum, at different exhibits, with different staff members, on different days, etc. (i.e., external validity).

Practical Considerations

Although the REVEAL project, like GIVE, was part of a large grant project (approximately $800K budget over three years), the research team had to consider many practical considerations beyond the specifics of the study design. The quasi-experimental study was highly resource and time intensive, requiring the team to make practical trade-offs along the way, such as choosing to only work with three different exhibits and four different educators and determining the number of videotaped interactions that could be realistically analyzed. In this respect, the power analysis was essential, since it would not have been possible to analyze more videos within the scope and budget of the grant-funded project. The nature of the study, and the project more broadly, also required close collaboration between educators and researchers, which influenced all aspects of the study goals, methods, and analysis process. This collaboration was partly achieved by bringing museum educators on to the research team from the very beginning of the project and by closely coordinating with museum education managers to determine which staff members would be available to participate in such a long-term, time-intensive project, and

how the team could collect data without disrupting other education activities and staffing needs throughout the museum.

Related to this, the study also required the team to recruit and train four new educators, which therefore necessitated the development of a training program and materials. The REVEAL study was not intended to test the impact of professional development on staff facilitation, but the team did need a group of educators trained in a standardized way to serve as the treatment condition for the quasi-experimental design, similar to the GIVE project. These educators needed to have some consistency and fidelity in the way they facilitated family learning, based on the model of facilitation that had been developed in the previous phase of the project. However, the team hoped that the training provided to the educators would not be so far beyond what an educator at another museum might realistically receive that the results of the study would have little practical significance for the field. Therefore, the team designed a two-week, intensive training that was presented to the four educators as a group immediately before the start of data collection. Researchers also conducted a preliminary review of video data and checked in with the four educators at several points during data collection to reinforce key ideas and facilitation strategies. This training process was documented and now serves as a resource for the field (https://REVEAL.terc.edu). It also represents a limitation to the external validity of the findings, since study results are clearly specific to this group of educators who experienced this type of training designed for these specific exhibits and this overall facilitation model.

Example 3: Quasi-Experimental with Independent Pre- and Post-Test Design

Finally, we present an example of a quasi-experimental design that used independent pre- and post-test groups to examine causal claims, while minimizing the burden for study

participants. *Facilitation Research for Engineering Design Education (FREDE)* was a two-year Institute of Museum and Library Services (IMLS)-funded study carried out by the Museum of Science, Boston (MOS). The project had two parts: (1) a quasi-experimental research study designed to examine the impact of educator facilitation on children's engineering perceptions in Design Challenges, a drop-in activity that engages visitors in the engineering design process; and (2) a two-day professional development workshop for educators focused on the development and facilitation of hands-on design-based activities similar to Design Challenges. FREDE built on several previous studies, including the IMLS-funded Engaging Girls in Engineering Design (Auster & Lindgren-Streicher, 2013), as well as an internally supported study focused on the effect of competition in engineering activities (Beyer & Auster, 2014). The study's focus on educator facilitation was driven by the Design Challenges staff's expressed desire to better understand which facilitation techniques result in a positive, engaging engineering experience for children. In alignment with program goals and the mapping of activities to local educational standards for students in grades 2 through 8, participants aged 7 to 14 were included in the study.

The quasi-experimental, mixed methods research study sought to determine if participation in a facilitated Design Challenges activity led to changes in engineering self-efficacy (Bandura, 1977; Paris & Paris, 2001) or attitudes toward engineering (e.g., Martin, Mullis, Foy, & Stanco, 2012). Aspects of interactions with educators—the frequency of the interactions, the phase of the design cycle in which they occurred, and the content of these interactions—were also observed to determine if they influenced any changes in participant perception. Three different engineering activities ("Echo Base Bobsleds," "Ships Ahoy!" and "Create-A-Claw") were selected by educators for inclusion in the study, and participants were evenly divided among the three, as well as between the research conditions described below.

Study Design

Neither researchers nor museum educators desired or felt it was ethical to remove staff facilitation from the Design Challenges experience to create a true control group, nor was it deemed feasible to request multiple surveys from adolescents in this context. As a result, the study was set up to create a separate but statistically comparable pre-test group to allow for the exploration of facilitation effects in more typical pre- to post-test fashion. The design was therefore what has been called a "post with independent pre design" (Fu et al., 2016), as shown in Figure 4.

NR O1 C —

NR — X O2

Figure 4. Quasi-experimental design for FREDE project. NR = no random assignment; O = measurement/observation (O1 = pre, O2 = post), C = comparison group, X = treatment group.

Using this set up, the pre-test (comparison) groups were selected after they approached the Design Challenges activity but prior to actually engaging in the engineering experience. This ensured that these groups had similar motivations and interest in participating as those who would end up taking part in the activity. (Groups who were selected based on this criterion but did not end up participating in the engineering design experience were excluded from analysis.) The post-test (treatment) groups were observed for the duration of their Design Challenges experience and then approached for additional data collection afterward. Following the Institutional Review Board (IRB)-approved protocol for this study, signs were placed in visible locations in the Design Challenges space, notifying visitors that observations were occurring.

Groups who were observed but then declined to participate further were also excluded from the study.

For both pre- and post-test conditions, one focal child per group was given a survey asking about future participation in engineering activities, self-efficacy, and interest. These children were also asked several brief interview questions designed to elicit additional understanding of these constructs.

Before the study, the team conducted an *a priori* power analysis to ensure that, given the logical constraints of the study in terms of both data collection hours and budget, medium to low effect sizes could be detected with sufficient power. A total sample of 300 study participants was found to be optimal, so a balanced design was created by including 100 participants for each of the three engineering activities (50 each in the pre- and post-test conditions). Given the extensive body of literature on gender differences in STEM interest and performance (e.g., Goodman, 2002; Legewie & DiPrete, 2011) and program educators' explicit desire to increase engineering opportunities for girls, purposeful sampling of boys and girls led to blocks of 25 of each gender within both pre- and post-test groups for each engineering activity. The selection of participants produced a roughly normal distribution of age within the target range.

Design Tradeoffs and Validity Considerations

Similar to the other studies described in this article, the FREDE team made several decisions regarding the research design that had critical implications for internal and external validity. Primary among them were the choice not to collect pre- and post-test measures from the same individuals, the decision not to create a true control group, and the desire to use instrumentation that aligned with previous Design Challenges research.

Unlike the GIVE and REVEAL projects, which used data from visiting groups that may have contained children, FREDE relied almost exclusively on data from children within groups. Given that the target age range (7 to 14 years old), researchers felt it was impractical to collect data from the same adolescent participants before and after their engineering experience. Previous studies on Design Challenges have found that the average stay time is roughly 20 minutes, so asking these youth to participate in both pre- and post-test surveys and interviews would have created an excessive burden for such a short free-choice learning experience. Additionally, researchers were concerned about the possible threat to the quality of participant responses due to a testing effect—a form of response bias in which respondents remember their earlier answers and respond based on that memory, rather than based on the effect of the intervention.

Creating an independent pre-test group may have minimized the burden for participants, but it also reduced the study's internal validity, as the same individuals were not being measured before and after their experience. The threat to claims about the effectiveness of the Design Challenges "intervention" was obvious: the two groups (pre and post) could easily differ in some fundamental ways *other* than having participated in a Design Challenges activity, thus causing the observed differences. In an attempt to mitigate this concern, the two groups were compared based on personal characteristics that were countable or measurable, including age, gender, and prior participation in an MOS Design Challenges activity (a proxy for engineering activity experience). The resulting sample was indeed sufficiently close across all of these variables. Of the 301 youth participants (151 pre, 150 post), 51% of the pre-test group and 50% of the post-test group were girls, the mean and median ages were the statistically equivalent, and 41% of the pre-

test group and 38% of the post-test group had prior Design Challenges experience (not a statistically significant difference).

In testing for differences between the pre- and post-test groups according to these criteria, researchers were able to make claims about the comparability of the two groups, defending the choice to have independent groups measured before and after participating in Design Challenges. Although not truly a "matched pairs" design, in which individuals who match across key variables are paired and then randomly assigned to experimental groups, the similarities between groups helped eliminate at least some of the internal validity threats. Furthermore, the sampling strategy of only recruiting those adolescents who approached Design Challenges and would later participate for the comparison group reduced the concern that the two groups differed in terms of interest in the activity (seen as a proxy for interest in engineering). Despite this, potential confounds remained: did the pre- and post-test groups differ in terms of engineering self-efficacy? Were youth with more formal engineering education unknowingly but systematically recruited for the post-test group? Were adolescents with parents working as engineers somehow only selected for the pre-test group? Questions such as these went untested and unanswered.

Given these challenges, it was clear that creating a true control condition would be extremely difficult. The research team acknowledged that the internal validity of the study would be strengthened with the inclusion of a control, but the task of doing so seemed daunting and unethical—requiring the recruitment of separate sample of children who had not participated in Design Challenges at any point during their museum visit, assessing these children upon their entry to the museum, again at exit, and verifying that they had in fact not participated in Design Challenges (and likely losing a considerable amount of data for groups that did do the

engineering activity). Logistical challenges and the implications on data collection time prevented researchers from pursuing this option.

In regards to external validity, one of the ways in which this aspect of the study was strengthened was through the inclusion of the professional development workshop offered as part of the FREDE grant award. Designed to bring many educators together from museums offering facilitated, hands-on activities delivering visitors of all ages the opportunity to design, build, and engineer within certain constraints and toward a specific goal (similar in nature to Design Challenges), the two-day workshop allowed for conversation and in-person demonstrations from more than 40 professionals at 19 different institutions. Together, these educators jointly learned about the impact of facilitation and vetted the research activities by discussing exactly how results from the study would have relevance to them and apply to their museum environments.

Instrumentation also played a large role in the consideration of internal and external validity concerns. Both the survey questionnaire and interview protocol were similar in design to those used in earlier studies on Design Challenges, which meant that the findings from FREDE could contribute to a growing understanding of the effects of participating in a Design Challenges activity across a broader range of MOS youth participants. These questions were originally designed with the goals of the program in mind and with consideration to the age range of the intended participants, increasing the likelihood that they would provide data consistent with the intentions of Design Challenges. On the other hand, the measures had technical limitations: no psychometric work had been done on any of the questions or scales used and visitors to MOS have been found to rate themselves consistently high on attitudinal and self-efficacy measures. The ability of some questions to discriminate between response options – for example, between “Sort of Agree” and “Really Agree” on a 4-point agreement scale in response

to a question about wanting to be an engineer – was not sufficient to allow differences to be detected. This meant that some of the scales used, while parallel to earlier studies, lacked the statistical precision necessary to detect change effectively between pre- and post-test groups, despite the power analysis that was conducted. While it was argued that the continuation of measures across studies strengthened the overall research, it most likely weakened the internal validity argument for FREDE, as measurement error could have influenced the findings.

Finally, the use of observational data during the Design Challenges activity for the post-test group contributed to the overall strength of study findings. Observations allowed the team to develop an understanding of the variability of facilitation across groups and to explore how aspects of facilitation related to visitor outcomes. Although these analyses were non-experimental, since they were conducted within the treatment group alone, they contributed to an understanding of the underlying mechanisms explaining the impact of facilitation, which is often lacking in a typical experimental or quasi-experimental design. For example, while all educators were trained in facilitation, the experience was not scripted. Therefore, not all groups received the same level of facilitation or feedback, which likely contributed to the impact on visitors. Moreover, conducting observations necessitated operationalizing the study constructs and treatment group design more concretely, particularly defining exactly what an "educator interaction" consisted of, while being very specific about the timing of events in relation to the engineering design cycle. All data collectors participated in the creation of a data collection protocol outlining these definitions after piloting the observation form, as well as undergoing interrater reliability testing prior to beginning data collection.

Practical Considerations

Compared to GIVE and REVEAL, the budget for FREDE was relatively modest. Of the $260,000 in funding (approximately $129,000 awarded by IMLS and a matching amount contributed by MOS per grant requirements), the quasi-experimental research study comprised less than half of the total, with the majority of funds used to plan and host the professional development workshop. Had more rigorous selection criteria been imposed on study participants, alternate measures with psychometric testing been developed exclusively for the study, or additional time-consuming methods such as video analysis been used in analysis, this figure easily could have doubled or tripled. Instead, the choice was made to keep the study small enough in scale that the second FREDE component (the professional development workshop) could remain a focal point of the grant.

As is typical in free-choice learning environments such as Design Challenges, maintaining a positive visitor experience was of primary interest to both educators and researchers. While signs were posted for IRB purposes, participants were not contacted by researchers until they had completed the activity in full in order to avoid participant reactivity and mitigate self-selection bias. In addition to losing data for those groups who were observed and then declined to participate, this passive recruitment approach required an extended data collection timeline in order to balance the participant samples within each group by gender and activity. Moreover, data collection was complicated by the operational scheduling of the program, which was open in two-hour windows anywhere from one to three times per day, depending on the day of the week and the season. Because education staff members wanted to rotate Design Challenges activities for visitors, this meant that, of the nine activities offered, the three included for study were not available for data collection every day. Last-minute changes to

scheduled activities based on large field trip groups, staffing needs, and activity materials led to additional challenges with the data collection schedule. As a result, the data collection that was originally scheduled for five months took more than 12 months to complete. During this time, staffing at the Design Challenges fluctuated, which was a limitation to the study that the team was not able to address.

Conclusion

In this article, we have presented an overview of the logic underlying experimental and quasi-experimental designs and provided examples that demonstrate how these approaches can be implemented in the field of visitor studies while attending to the free-choice and informal learning ethos of the study context. Experimental studies are certainly not the best or the only approach to investigating research and evaluation questions in museums and other informal learning environments. In fact, when investigators have descriptive questions about what visitor experiences look like in these settings, hypothesis generation goals to identify emergent factors and patterns for future research, or explanatory questions about the complex mechanisms and processes that underlie what goes on in these settings, an experimental approach is likely not appropriate. However, providing strong causal evidence of the impact of a specific program, intervention, or design is a common goal in visitor studies, and a growing focus of funders, especially in summative evaluations. In these cases, in which the results of the study can inform programmatic decisions, experimental and quasi-experimental designs are powerful tools, which we believe the field can use more frequently.

As examples in this article demonstrate, even when investigators select an experimental or quasi-experimental approach, there are many choices to be made that will ultimately affect the

strength of the causal claims made by the investigators and the extent to which findings can be argued to transfer or generalize to other contexts and settings. We want to emphasize that there is no perfect design, even a true experimental study, that does not involve trade-offs. In the GIVE project, the team created a strong experimental design that allowed for the randomization of visitor groups and both (a) provided evidence of the causal link between facilitation using the "juicy questions" game and visitor outcomes and (b) eliminated possible alternative explanations, such as the general benefit of staff facilitation. However, the team recognized that questions remained about whether or not these findings could be replicated in the often-chaotic environment of a museum since the experiment was conducted under strict controls. In contrast, the REVEAL and FREDE project teams chose to use quasi-experimental designs to maximize the authentic, naturalistic context of the studies, but these presented their own challenges. For REVEAL, the team developed a comparison condition that eliminated plausible alternative explanations related to visitor self-selection bias, but ultimately had to accept that the comparison between the facilitation and greeting conditions did not reflect the true difference between visitor experiences with and without facilitators. For FREDE, the non-equivalent pre- and post-test groups greatly reduced the burden on participants and the likelihood of reactivity effects, but also left open many questions about potential differences between the groups that were not measured by the team. Despite these difference, all three projects used a variety of approaches to ensure the studies were authentic and relevant to the museum context, such as working closely with educators throughout the design and implementation process, attending to the comfort of visitors and staff during data collection, aligning the study design and focus to the educational goals of the experience, and using multiple methods to capture a broad perspective on visitor outcomes.

The three examples presented in this article all focus on staff-facilitated visitor experiences in designed settings (more specifically, in large science centers), and therefore the trade-offs we reviewed only scratch the surface of the potential challenges that visitor studies professionals face in investigating different types of settings and programs. For example, investigators might want to test the causal impact of a museum-based afterschool program with groups of youth who are preselected to participate through a partnership with another organization and repeatedly engage in the program multiple times over many months. This situation presents a whole new set of challenges not discussed in this article, including issues of interdependence (i.e., outcomes for participants are interrelated, since they participate in the program together), questions about randomization (e.g., is it ethical to exclude some participants from aspects of the experience in order to form a comparison group), design decisions about clustering (i.e., participants could be randomized at the individual level or "cohorts" of participants could be randomized at the group level), and feasibility requirements for tracking the same participants over a longer time span.

One critical challenge is that experimental and quasi-experimental designs are often expensive to implement. All three of the examples represented here were part of relatively large, federally funded grant projects and involved a substantial number of participants and complex data collection and analysis. However, experiments and quasi-experiments can also be conducted with fewer resources. Alice Fu and her colleagues suggest that by planning ahead and leveraging existing resources during program development, evaluators can reduce costs of experimental designs (Fu et al., 2016). For example, the first author has incorporated quasi-experimental studies into larger projects and used volunteers to support planning, data collection, and analysis (Pattison, Ewing, & Frey, 2012; Pattison & Shagott, 2015). In another example, the second

author used a quasi-experimental design in an inexpensive formative evaluation study of question-asking in labels by comparing several versions of the same label at an exhibit (Gutwill, 2006).

Regardless of the research questions or the program context, we encourage investigators to carefully consider the trade-offs of the study design as they relate to internal validity, external validity, and other practical considerations. These decisions must also take into account ongoing developments in experimental design, such as new work on "retrospective pretest" designs (e.g., Mueller & Gaus, 2015; Pratt, McGuigan, & Katzev, 2000) and advances in statistical methods accounting for systematic differences between treatment and comparison groups (e.g., Dong, 2015). Ultimately, for the field of visitor studies to advance and to be helpful in improving visitors' experiences, we must develop a robust toolkit of methods, study designs, and measures that allow us to answer different research questions and adapt to the needs and challenges of different projects.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grants 1321666 and 0411826 and the Institute of Museum and Library Services under grant MG-10-13-0021-13. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

References

Allen, S., & Gutwill, J. P. (2009). Creating a program to deepen family inquiry at interactive science exhibits. *Curator: The Museum Journal, 52*(3), 289–306.

Allen, S., Gutwill, J. P., Perry, D., Garibay, C., Ellenbogen, K., Heimlich, J., … Klein, C. (2007). Research in museums: Coping with complexity. In J. H. Falk, L. D. Dierking, & S. Foutz (Eds.), *In principle, in practice: Museums as learning institutions* (pp. 44–56). Lanham, MD: AltaMira.

Auster, R., & Lindgren-Streicher, A. (2013). *Engaging girls in engineering design*. Boston, MA: Museum of Science, Boston.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 191–215.

Beyer, M., & Auster, R. (2014). *Assessing competition in engineering*. Retrieved from Museum of Science, Boston website: http://www.informalscience.org/sites/default/files/2015-06-15_2014_Assessing_Competition_in_Engineering.pdf

Bitgood, S. (1988). An overview of the methodology of visitor studies. *Visitor Behavior, 3*(3), 4–6.

Bitgood, S., Patterson, D., & Benefield, A. (1988). Exhibit design and visitor behavior: Empirical relationships. *Environment and Behavior, 20*(4), 474–491.

Brewer, M. B., & Crano, W. D. (2013). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd ed., pp. 11–26). New York, NY: Cambridge University Press.

Campbell, D. T., & Stanley, J. C. (1967). *Experimental and quasi-experimental designs for research* (2nd ed.). Boston, MA: Houghton Mifflin Company.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, N.J: Erlbaum Associates.

Dong, N. (2015). Using propensity score methods to approximate factorial experimental designs to analyze the relationship between two variables and an outcome. *American Journal of Evaluation, 36*(1), 42–66.

Eckmanns, T., Bessert, J., Behnke, M., Gastmeier, P., & Ruden, H. (2006). Compliance with antiseptic hand rub use in intensive care units: The Hawthorne Effect. *Infection Control and Hospital Epidemiology, 27*(9), 931–934.

Falk, J. H., & Dierking, L. D. (2013). *The museum experience revisited*. Walnut Creek, CA: Left Coast Press.

Falk, J. H., Koke, J., Price, C. A., & Pattison, S. A. (2018). *Investigating the cascading, long term effects of informal science education experiences report*. Beaverton, OR: Institute for Learning Innovation.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fu, A. C., Kannan, A., Shavelson, R. J., Peterson, L., & Kurpius, A. (2016). Room for rigor: Designs and methods in informal science education evaluation. *Visitor Studies, 19*(1), 12–38.

Garibay Group. (2013). *Design Zone exhibition summative report*. Retrieved from OMSI website: http://informalscience.org/evaluation/ic-000-000-008-817/Design_Zone_Exhibition_Summative_Report

Goodman, I. F. (2002). *Final report of the women's experience in college engineering (WECE) project*. Retrieved from Goodman Research Group website: http://grginc.com/WECE_FINAL_REPORT.pdf

Greenes, C. E., & Rubenstein, R. (Eds.). (2008). *Algebra and algebraic thinking in school mathematics: Seventieth yearbook*. Reston, VA: National Council of Teachers of Mathematics.

Gutwill, J. P. (2003). Gaining visitor consent for research II: Improving the posted-sign method. *Curator: The Museum Journal, 46*(2), 228–235.

Gutwill, J. P. (2006). Labels for open-ended exhibits: Using questions and suggestions to motivate physical activity. *Visitor Studies Today, 9*(1), 1–9.

Gutwill, J. P., & Allen, S. (2010a). Facilitating family group inquiry at science museum exhibits. *Science Education, 94*(4), 710–742.

Gutwill, J. P., & Allen, S. (2010b). Group inquiry at science museum exhibits: Getting visitors to ask juicy questions. In Exploratorium Museum Professional Series. San Francisco, CA: Exploratorium.

Gutwill, J. P., & Allen, S. (2012). Deepening students' scientific inquiry skills during a science museum field trip. *Journal of the Learning Sciences, 21*(1), 130–181.

Hirschi, K. D., & Screven, C. G. (1988). Effects of questions on visitor reading behavior. *ILVS Review, International Laboratory for Visitor Studies, 1*(1), 50–61.

Humphrey, T., & Gutwill, J. P. (2005). *Fostering active prolonged engagement: The art of creating APE exhibits*. San Francisco: Exploratorium.

Kaput, J. J., Carraher, D. W., & Blanton, M. L. (2008). *Algebra in the early grades*. New York, NY: Erlbaum/National Council of Teachers of Mathematics.

Kohli, E., Ptak, J., Smith, R., Taylor, E., Talbot, E. A., & Kirkland, K. B. (2009). Variability in the Hawthorne Effect with regard to hand hygiene performance in high- and low-performing inpatient care units. *Infection Control and Hospital Epidemiology, 30*(3), 222–225.

Legewie, J., & DiPrete, T. A. (2011). *High school environments, STEM orientations, and the gender gap in science and engineering degrees*. Retrieved from https://academiccommons.columbia.edu/catalog/ac:139666

Marino, M., & Koke, J. (2003, February). Face to face: Examining educational staff's impact on visitors. *ASTC Dimensions*. Retrieved from http://astc.org/pubs/dimensions/2003/jan-feb/index.htm

Martin, M., Mullis, I., Foy, P., & Stanco, G. (2012). *TIMSS 2011 international results in science*. Boston, MA: TIMSS & PIRLS International Study Centre.

Morgan, D. L. (2014). *Integrating qualitative and quantitative methods: A pragmatic approach*. Thousand Oaks, CA: Sage Publications.

Moses, B. (Ed.). (1999). *Algebraic thinking, grades K-12: Readings from NCTM's school-based journals and other publications*. Reston, VA: National Council of Teachers of Mathematics.

Mueller, C. E., & Gaus, H. (2015). *Assessing the performance of the "counterfactual as self-estimated by program participants": Results from a randomized controlled trial*. American Journal of Evaluation, 36(1), 7–24.

Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist, 36*(2), 89–101.

Pattison, S. A. (2011). *Access Algebra staff facilitation: A formative evaluation report*. Retrieved from http://www.omsi.edu/sites/all/FTP/files/evaluation/algebrastafffacilitation.pdf

Pattison, S. A., & Dierking, L. D. (2013). Staff-mediated learning in museums: A social interaction perspective. *Visitor Studies, 16*(2), 117–143.

Pattison, S. A., Ewing, S., & Frey, A. K. (2012). Testing the impact of a computer guide on visitor learning behaviors at an interactive exhibit. *Visitor Studies, 15*(2), 171–185.

Pattison, S. A., Randol, S. M., Benne, M., Rubin, A., Gontan, I., Andanen, E., … Dierking, L. D. (2017). A design-based research study of staff-facilitated family learning at interactive math exhibits. *Visitor Studies, 20*(2), 138–164.

Pattison, S. A., Rubin, A., Benne, M., Gontan, I., Shagott, T., Francisco, M., … Dierking, L. D. (2018). The impact of facilitation by museum educators on family learning at interactive math exhibits: A quasi- experimental study. *Visitor Studies, 21*(1), 4–30.

Pattison, S. A., & Shagott, T. (2015). Participant reactivity in museum research: The effect of cueing visitors at an interactive exhibit. *Visitor Studies, 18*(2), 214–232.

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. *American Journal of Evaluation, 21*(3), 341–349.

Serrell, B. (2000). Does cueing visitors significantly increase the amount of time they spend in a museum exhibition. *Visitor Studies Today, 3*(2), 3–6.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Sindorf, L., Gutwill, J., & Garcia-Luis, V. (2015). Gaining visitor consent for research III: A trilingual posted-sign method. *Curator: The Museum Journal, 58*(4), 369–381.

Weiss, B. H., O'Mahony, M., & Wichchukit, S. (2010). Various paired preference tests: Experimenter effect on "take home" choice. *Journal of Sensory Studies, 25*(5), 778–790.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2013). *Causal inference and generalization in field settings: Experimental and quasi-experimental designs*. In H. T. Reis & C. M. Judd (Eds.), Handbook of research methods in social and personality psychology (Second edition, pp. 40–84). New York, NY: Cambridge University Press.

Notes

[1] In our discussion, we intentionally do not distinguish between the use of experimental design in research and evaluation. Although the purposes of evaluation and research are often different, in either case the strength of experimental and quasi-experimental methods is for testing causal relationships, whether those be about the outcomes of an intervention or a theory about education and learning.

[2] Although each of the examples is described by one of the authors who played a central role on the project, the reflections and overall framing of the article represent the thinking of the entire author team.

[3] Blocking is another strategy for increasing statistical power, since it ensures (beyond randomization) that other variables potentially influencing the outcomes of the study (e.g., individual educator or teacher, gender of participant) are equally represented across conditions.

[4] In other words, differences in participants across the conditions that existed before the treatment and might offer an alternative explanation to the outcomes of the experimental study.

[5] We use the term “comparison group” instead of “control” in the quasi-experimental design examples to emphasize the distinction between a true control group in a randomized design.

[6] The larger sample size for the treatment condition was designed to allow for more sensitive, exploratory analyses within that condition, such as testing the relationship between particular staff facilitation strategies and visitor outcomes.