This is a preprint of a book chapter:
Allen, S., Gutwill, J. P., Perry, D., Garabay, C., Ellenbogen, K., Heimlich, J., . . . Klein, C. (2007). Research in Museums: Coping with Complexity. In J. Falk, L. Dierking & S. Foutz (Eds.), *In Principle, In Practice: Museums as Learning Institutions* (pp. 229-246). New York: Rowman & Littlefield Publishers, Inc.

Research in museums: Coping with Complexity

Lead Authors: Sue Allen, Joshua Gutwill

Supporting Authors: Deborah L. Perry, Cecilia Garibay, Kirsten M. Ellenbogen, Joe E. Heimlich, Christine A. Reich, Christine Klein.

Introduction

Museums and other informal learning environments are challenging to study. One reason is the enormous variability in both the environment and the audience – researchers need to address a multitude of interacting factors in order to make sense of what visitors learn. Falk and Dierking's contextual model of learning (2000) underscores the complexity inherent in visitor studies by identifying four broad contexts or influences, each of which is a complicated world unto itself: the physical, personal, and sociocultural realms, and time. Their complexity and inter-relatedness can easily overwhelm a researcher. For example, how does one characterize a "typical visitor experience" even in terms of basic exposure, if each person's attention takes a unique path through a densely featured space of possibilities?

Fortunately, museum researchers and evaluators have adapted and developed methods for managing the complexity involved in studying free-choice learning. In this chapter, we categorize such methods as falling into two main approaches: reducing complexity to identify causal relationships, and embracing complexity to create a deep understanding from multiple perspectives.

These two approaches have deep roots in social science, and both are practiced by well-respected researchers (National Research Council, 2002). For each, we mention some of the underlying philosophical assumptions, show how the approach applies to visitor studies in particular, and highlight (with examples) some useful methods for maximizing quality. We do not advocate one approach over another; rather, we attempt to highlight some of the issues that arise in the challenging processes of study design and implementation.

Reducing complexity to identify causes and effects

Museum practitioners design learning environments intentionally, with some form of goal or purpose (however broadly defined). As a result, they often desire an approach that searches for cause-and-effect relationships to determine how well their goals are met and how their designs affect visitors. This is also desirable to those researchers and evaluators who believe in the possibility of general principles that characterize which environments and experiences will result in which outcomes.

Cause-and-effect research such as randomized clinical trials is treated in some circles as the "gold standard" of social science research because of its predictive and explanatory power (see Shavelson & Towne, 2004). However, it is not often used in its pure form in visitor studies for two main reasons, one theoretical and the other logistical. The theoretical question is whether it is possible to identify and measure key concepts such as motivation, engagement, or learning, and to assume they mean the same thing for a variety of visitors in a variety of situations. In a field where so much emphasis is given to visitors' personal interpretations and choices, can one legitimately combine different people's experiences and draw conclusions about characteristics of visitors and their learning? Researchers who support experimentation believe that this is indeed possible, and they point to social psychology, which has a history of taking complex situations in real-world settings, pulling out a small number of variables for study, and arriving at fundamental principles of social behavior that can then be applied back to the real-world situation. For example, a baffling case where a woman was murdered in full view of the windows of a large apartment building led to a set of laboratory experiments by Latane and Darley (1969) on altruistic behavior, ultimately yielding an explanation of why no one had called for help at the time.

Such experiments need to be done carefully, however, to have any credibility, and certain controls and comparisons are needed to rule out competing explanations. This leads to the second reason that the field of visitor studies lacks controlled experiments: the logistical challenges. Often researchers do not have the ability to "randomly assign" visitors to different versions of an exhibit or program, yet random assignment is a fundamental principle of this kind of work. Similarly, the experimental approach relies on comparisons: between visitors who were exposed to a program and those who were not, between visitors before and after using an exhibit, etc. Such comparisons are often expensive and difficult to arrange, and may intrude on visitors' natural behavior.

In spite of these challenges, some experimental research is being done in museums, and the principles can even be incorporated into regular evaluation procedures when the goal is to make causal claims about learning. Here we describe three ways to reduce complexity and increase the likelihood of identifying cause-and-effect relationships in free-choice learning studies: (1) simplify effects into a few measurable outcomes, (2) focus on a small number of potential causes, ,and most importantly, (3) reduce the number of competing explanations.

Focus on a few measurable outcomes

Outcomes (or "dependent variables") are the central things being studied or measured, typically such things as visitors' knowledge, emotions, ,attendance, conversations, etc. One way to simplify an array of possible outcomes is to select certain aspects of visitors' behavior and ignore the rest,. For example, rather than analyze entire visitor conversations, Humphrey and Gutwill

(2005) chose to study only the questions visitors asked themselves and the statements that immediately followed, because one of their project goals was to shift authority from the exhibit's label text to the visitors' own driving questions. Similar selection of key outcomes happens every time researchers force visitors' responses into a predetermined numerical scale, and even interviews are best analyzed with respect to a dimension or variable that is motivated by theory or stakeholders' questions.

A second way to simplify outcomes is to give something a single score that captures the "big picture." For instance, rather than identify a dozen different aspects of visitor inquiry at a science museum, one could assign a single inquiry score to each visitor group, based on overall judgments of their behavior. Properly done, this kind of holistic scoring is easier to analyze and has been found to be just as effective at characterizing visitors' inquiry behaviors (Randol, 2005).

Finally, there are statistical techniques for simplifying the outcomes after a study has been run, such as cluster analysis and skree plotting. Cluster analysis is a process for finding patterns by clustering respondents into meaningful groups (e.g. Hui-Min & Cooper, 2001). Skree plotting and other scatterplot methods provide a visual and statistical means for observing the logical groupings of variables (e.g., Heimlich, Storksdieck, Barlage, & Falk, 2005).

Focus on a small number of potential causes

When trying to understand the reasons behind an outcome, it is often helpful to limit the number of potential causes ("independent variables") included in the study . Having fewer variables improves the statistical chances of finding a result if one exists, especially important in visitor studies that involve small samples of visitors. It also tends to improve the clarity and interpretability of the data analysis (Prosavac, 1998).

Factors that might affect the outcomes but are not the focus of a study can be dealt with in three ways: by fixing them at a single level, by randomizing, or by blocking. For example, if age is expected to affect visitors' experiences, but this is not the focus of the study, a researcher could fix this variable by including only visitors whose ages fall within a narrow range. Alternatively, visitors of all ages could be assigned randomly to different versions of the offering being tested, so that after many visitors have participated, the effect of age would "wash out," being roughly equal across all versions. Or, using the third technique, a researcher could 'block' the variable into categories such as child, adolescent and adult (ideally, natural groupings observed in pilot studies). Visitors would participate until a representative number of visitors existed in each block. These three techniques occupy different positions in the trade-off between deeper understanding of a particular group's experience on the one hand, and broader applicability of the study results on the other. Finally, factors that cannot logistically be controlled for using these techniques can often be dealt with statistically following data collection.

While reducing the number of variables (both dependent and independent) in a study is an effective way to cope with complexity, such simplification comes with two cautionary notes. First, it is always possible that the learning outcomes observed may actually be due to a variable that was not measured or recognized as important. Second, the approach puts limitations on the degree to which the results can be applied to other situations. For example, researchers studying exhibit narratives at the Exploratorium were so concerned about limiting the number of variables that they ended up studying narratives that lacked aesthetic variation and interest; later they revised their study design to allow for testing of narratives that were more representative of realistic museum creations (Allen, 2004). Even something as simple as fixing the exposure to an exhibit label by asking visitors to read all of it, may change the context of learning from that on the open museum floor. Researchers frequently face this kind of trade-off between "internal validity" (the design of a carefully controlled experiment), and "external validity" (the degree to which its results can be generalized beyond those specific circumstances). Often they compromise, limiting the number of variables but choosing them carefully to capture a few key causal relationships and not drift too far from the real-world context of museum learning.

Reduce the number of competing explanations: some useful study designs

A key characteristic of an experimental study is that it allows researchers to rule out alternative theories or explanations of an observed outcome. Only rarely are visitor studies created with this as a priority. Usually, the limited evaluation resources are focused on characterizing outcomes of interest (e.g., visitors' behaviors, or their reflections on a visit or program) rather than proving that these were generated by specific causes (e.g., proving that visitors' understandings came from experiencing a program rather than from their prior knowledge, or that visitors' increased engagement at an exhibit was due to improvements in its design, rather than changes in the population of visitors using it). However, in an era when museums are increasingly being asked for proof of their impact and the factors that contribute to it, we list several examples of ways to adjust some of the most frequently-used research and evaluation designs to support more rigorous cause-and-effect inferences. We draw on well-known overviews given by Campbell & Stanley (1963) and Cook & Campbell (1979).

Interviewing after an experience

A common evaluation design is the exit interview, in which visitors are interviewed as they leave an exhibit or program, and conclusions are drawn about what they learned. While this design can efficiently capture visitors' reflections about what they did and thought during their experience, it usually relies on memory and verbal self-report of what has been learned, rather than directly assessing any change in visitors' knowledge, attitudes, or skills. Unfortunately, even an exit interview that does directly assess such things as visitors' knowledge as they leave the program exhibition is vulnerable to the competing explanation that the visitors already had such knowledge and abilities before they arrived.

Interviewing before and after ("pre versus post")

To reduce competing explanations, a somewhat stronger version of the exit interview involves interviewing visitors twice. For example, in personal meaning-mapping (Falk, Moussouri & Coulson, 1998), visitors represent their knowledge on concept maps before and after their visit. However, the design is still vulnerable to alternative explanations: learning might have been enhanced because the visitors knew they would be interviewed (e.g., Serrell, 2000), or because they got practice doing the assessment. Also, it is difficult to design pre-interviews that fully anticipate the kinds of learning an individual visitor might do, and because it takes twice as long as a single interview, it may be difficult to recruit a representative sample of visitors.

Interviewing two groups

Another slight improvement over the exit interview is the two-group interview. The researcher conducts only one interview, but with two sets of visitors: those who chose to explore the offering (e.g., a program) being assessed, and those who did not. Comparing the two groups gives some idea of the effects of the program as compared with the rest of the institution. However, this design is open to the competing explanation of visitor selection: perhaps the already-interested visitors went into the program and the disinterested visitors did not. This is a tricky case; museums may be legitimately reluctant to assign visitors to having or not having an experience, either because they consider it too disrespectful or because thy worry that taking away visitors' choice may limit the study's applicability to real world of the museum (e.g., How do you interpret a study that includes people who hate audio-tours but learn something when forced to use one?). However, letting visitors decide means that a heavy price is paid in terms of internal validity of the study. If researchers want to prove that a particular offering resulted in a particular outcome, they need to study a large enough sample of visitors, and randomly assign them to different versions of an experience (deciding, for instance, who will get an audio-tour and who won't).

TWO RECOMMENDED STUDY DESIGNS FOR HIGH INTERNAL VALIDITY

Exit interview with random assignment

In a rigorous version of the two-group study, visitors are randomly assigned to different versions of the experience (or one group gets nothing if the goal is to prove the experience had an impact). Only one exit interview is required because all other potential causes (such as day of the week, time of day, or a sudden busload of historians) will “wash out” across treatments if the sample is sufficiently large. This design could actually be used far more often in museum research and evaluation than it is, especially when testing minor changes in exhibits or labels where there are alternate versions that can be easily interchanged at short notice. Ideally, this could be done even in “uncued” studies (i.e., without alerting visitors ahead of time that they will be interviewed), by randomly revealing one version of the offering for discovery by the next visitor who spontaneously approaches. More realistically, if the effort of swapping versions is high, rigor may be slightly compromised by leaving each version available for a number of visitors to experience it, before changing versions. This method was used in the Finding Significance project (Allen, 2004), where cued visitors were shown one version of an exhibit, and the version changed after each family.[1]

Counterbalanced (requires cuing)

Another strong design is the counterbalanced design, in which all visitors see all versions of an offering, but in different orders. Gutwill (2006) used this method to test an exhibit label in three formats; visitors were asked to comment on each, and state their preference. This design rules out many competing explanations such as selection, practice effects, etc. The main limitations are that the entire experience must be cued rather than spontaneous, the versions have to be swapped or

[1] This study used another slight modification from the method described, in that the versions of the exhibit rotated in sequence. This block design was a departure from a strict randomization, but was done because the interviews took over an hour, and the research team wanted to ensure that every day each version was used at least once.

revealed very quickly while a visitor waits, and it requires asking time of visitors to complete the whole process (leading to possible visitor selection effects).

A note about variability across interviewers and data coders

A final challenge to cause-and-effect experiments arises if multiple people collect or code the data. For example, one interviewer might inadvertently ask visitors more follow-up questions than another, or different coders may see different things in the same data. Even a single coder may drift over time, influencing the results in a certain direction. Most of these effects can be eliminated by having each person work equally across all variations tested, and by doing periodic inter-rater reliability tests: Multiple people record or code a subset of the data, and check their findings against one another. If the agreement drops below an acceptable level, the coding rubric is revisited (see Bakeman and Gottman, 1997). Needless to say, such testing and adjustment requires significant time.

In summary, reducing complexity is a powerful approach that can, at its best, provide evidence of the causes that lead to key visitor outcomes. Even relatively simple evaluation studies can often be strengthened by the use of appropriate control groups and comparisons, allowing evaluators to make stronger claims that learning has happened in the setting of interest, or that one design is more effective than another. At the same time, adding such comparison groups almost always uses more resources than a simple evaluation of outcomes, and the controlling of variables puts limits on the situations to which the conclusions can be applied.

Embracing complexity to understand multiple perspectives more deeply

In this section we describe methodological approaches that embrace, rather than reduce, the complexity inherent in informal learning settings. Naturalistic inquiry, case studies, and culturally responsive studies fall within a research paradigm for fostering deep understanding of the multiple realities of visitors' experiences, and all have been applied in museum contexts.

Naturalistic Inquiry

Naturalistic Inquiry is a methodology that embraces the variability of actual environments. It is grounded in the belief that the best way to study a research question is to look at many aspects of it in as much detail as possible in the natural setting (Lincoln & Guba, 1985).

> Naturalistic Evaluation takes a broad, holistic view of the program, exhibit or institution being studied, is more interpretative than judgmental, and requires participation from a wide range of people who are to be served by the study effort....The purpose is to uncover the multiple realities and multiple perspectives that exist and are provoked as people experience the museum environment—it reveals the configuration of meaning that emerges when different people are exposed to a common stimulus. (Wolf & Tymitz, 1979, p. 2-3).

Rather than search for cause and effect relationships by averaging across visitors, naturalistic inquiry seeks to understand the mutually influencing factors that yield a range of individual visitor experiences. This approach tends to enhance perception of the richness, complexities, and intricacies of museum environments.

Naturalistic and experimental inquiry often use similar methods (e.g., interviews, observations) but the exact procedures may differ. For example, because naturalistic research is seen as fundamentally interactional, interviews often grow out of observations and previous interviews rather than following a standard protocol. Also, naturalistic inquirers tend to employ purposive sampling (selecting a particular respondent for a particular reason), rather than sampling randomly. For example, the researcher may have just finished observing and interviewing an adult with a child, and now she wishes to see how two children working together without an adult might use the exhibit. Or perhaps the researcher overhears a particular group engaged in a heated debate at one exhibit, so asks to join their conversation. The goal of purposive sampling is to ensure that a broad range of visitor experience is included in the study, and that the interactions with any particular respondent are extended and rich, even if this results in small sample size.

This raises an important question: How does one know if a given naturalistic study is good, or at least good enough? With naturalistic inquiry, researchers employ four criteria for establishing trustworthiness: Credibility, Transferability, Dependability, and Confirmability (Gyllenhaal, 1998; Lincoln & Guba, 1985; Williams, n.d.).

Credibility

To achieve credibility, a naturalistic study should be believable to readers and approved by the people who provided the information gathered during the study. There are several procedures for enhancing credibility: During the data collection phase, *prolonged engagement* helps to build trust with respondents and allows researchers to experience variation over time. Ideally, this can be done over a period of years (e.g., Schaefer, Perry, & Gyllenhaal, 2002) but can even make a noticeable difference over a single evening (e.g., Gyllenhaal, Perry, & Cheng, 2005). *Persistent observation*, exploring details of the phenomena under study to a deep level, may reveal important or surprising results. For example, during the evaluation of a set of science exhibits at the Exploratorium, persistent observation of visitors led to the emergence of an unanticipated factor influencing the visitor experience, viz., the reasons that visitors chose to end their engagements with an exhibit element (Tisdal, 2004). Inquirers ought to verify their findings through *triangulation* by referring to multiple sources of information (including the literature); using multiple methods of data collection, such as interviews and observations; and using multiple inquirers when possible to diversify interactions with study respondents.

After the data have been collected, naturalistic researchers engage in several procedures for checking the credibility of their conclusions. *Negative case analysis* is used to refine conclusions until they account for all known cases without exception, and *progressive subjectivity checks* require a researcher to document their changing expectations for the study. During *peer debriefing,* the inquirer meets with one or more colleagues not involved in the study who can question the methods, emerging conclusions, and biases of the inquirer, and can sometimes offer a fresh interpretation. Finally, when possible, naturalistic researchers engage in *member checking* by having their data, interpretations, and reports reviewed by the respondents to determine whether their perspectives have been adequately and credibly represented. All these procedures help to ensure that the study is credible.

Transferability

Transferability means the degree to which findings from one context or setting (i.e., where the research was conducted) can be applied to others. Whether findings can be transferred or not is an empirical question which cannot be answered by the inquirer alone; people who read naturalistic inquiry reports have to make this determination themselves. To judge the transferability of a particular study, *thick description* is necessary, meaning a clear, rich, and complete description of the time, setting, and context during which the study took place (Geertz, 1983). In practice, because of budgetary and other constraints, description is often not as thick as future readers might need, but limited to comments such as data were collected on a particularly crowded, free day.

Dependability

Dependability is the stability or consistency of the inquiry processes used over time. To check the dependability of a naturalistic inquiry, an independent auditor reviews the activities of the researcher (as recorded in an audit trail in field notes, journals, and reports) to see how well the criteria of credibility and transferability have been met. In practice, this is rarely done due to budget constraints, but dependability audits could be done by graduate students learning to conduct evaluation and research in museums.

Confirmability

Confirmability is how well the *results* are supported by events that are independent of the researcher. In studies in museums, the main method for achieving confirmability is reference to the literature or previous studies.

Case Studies

Another way of embracing complexity in museums is the use of case studies, which traditionally focus on a small number of subjects and involve highly contextualized data collection and analysis. Case studies can offset the marginalizing effects that can result simply from focusing on averages when analyzing a set of data. For example, universal design (the creation of products and environments to be usable by all people without the need for adaptation) stems from recognition that there is no average person, and that designers should design environments that represent the diversity of size and ability that exists within the human population.

Many of the issues that are critical to conducting case studies are the same as those already described for Naturalistic Inquiry (in fact, case study may be regarded as a method falling under this broader umbrella). Researchers conducting case studies also face challenges that are specific to the approach, such as selecting the case and establishing its boundaries, collecting data appropriately and accurately from multiple sources over time, and interpreting and analyzing context-specific data (Anderson & Arsenault, 2004).

Selecting the case

Arguably, the most critical challenge for a case study is the selection of the case(s). Sometimes selection is driven by a research question or problem, such as: How are museum experiences incorporated into a family's day-to-day life and ongoing learning activities? A case for such a study should not be selected to be representative, but because it yields the most useful data. Specifically, a good case is cooperative, accessible, and active, and provides data relevant to the study's purpose. For example, in order to select good cases for an ethnographic case study of the role of museums in family life, Ellenbogen (2002) selected only families who went to museums six or more times per year, even though this was twice the typical rate defining a frequent museum-goer (Hood, 1983). The high visitation rate ensured that the case would be active enough to generate the needed data.

Collecting and analyzing data over time

Case studies may be particularly effective in addressing the complexity presented by the context of time. The rich detail uncovered through a case study can reveal both macro-level learning mechanisms (e.g., how school curricula and museum visits can mutually support learning (Anderson, Lucas, Ginns, & Dierking, 2000)) and the moment-by-moment interactions involved in learning in museums (Rowe, 2002). Whether macro or micro, the case study allows an exploration of why one particular interaction leads to another or how a learner's identity can be traced back through a developmental pathway of social interactions.

Systematic analysis

Performing a case study requires that the rich data set be analyzed systematically, often using what is referred to as the constant comparative method within Grounded Theory (Glaser & Strauss, 1967). In this iterative method, each incident or data point gathered through interviews or other assessments is compared with all others. An explanatory theory soon emerges, which drives further data collection. Simultaneously, the researcher begins to generate categories or codes that group the incidents according to the theory. Once new data cease to add diversity to a category, the code is considered saturated. When all codes are saturated and the theory seems stable, writing begins. This entire process is similar to that undertaken in Naturalistic Inquiry.

Culturally Responsive Research and Evaluation

The diversity of cultures within the visiting (and potentially visiting) audience is one kind of variability that plays a foundational role in museum research and evaluation. By "culture" researchers mean a group's beliefs, attitudes, customs, values, ways of thinking, communication patterns, and frames of reference. While it often refers to ethnic or racial background, any group that has some shared affiliation and characteristics (e.g., people with disabilities, teenagers) can be seen as having a common culture.[2] In museum research and evaluation, many consider it vital to embrace this type of variability, both to accurately reflect the experiences of a wide range of visitors, and to address the frequent calls to diversify museum audiences (e.g., American Association of Museums, 1992).

Culturally responsive approaches argue that no culture-free research or evaluation exists. For example, Ricardo Millett (2002) suggests that most evaluation instruments are developed for (and by) people who are employed, acculturated, well-educated, English-proficient, and enjoy moderate-to-high incomes. To this list could be added mobile, sighted, and hearing. In this view, part of the researcher's task is to actively cultivate one's own awareness and appreciation of cultural issues and their complexity, remaining vigilant to our own potential cultural biases. Kirkart (1995), among others, states that for an evaluation to have validity and utility, cultural perspectives must be addressed.

A culturally responsive framework affects all aspects of a study—from the formation of the research team to the dissemination of findings.

[2] Despite the use of this term as a simplification, a cultural group is never homogenous; within-group diversity is often based on factors such as socio-economic and educational background.

Inclusive composition of the research team

In culturally responsive research and evaluation, some members of the research team are members of, or at least deeply understand, the specific cultures included in the study. This ensures that the research questions are appropriately framed, and that the data collectors can recognize and interpret data appropriately. A variation of this type of inclusiveness is "participatory design," in which visitors are included as members of the design team rather than the research team, and are actively solicited for feedback during the creative process (Ringaert, 2001). Such a process may help to focus on the defects of the environment or exhibit, rather than the "defects" of the visitors, and is often used for research and evaluation relating to universal design (Gill, 1999).

Appropriate design of research instruments

In culturally responsive research and evaluation, the instruments used to collect data are framed in a culturally appropriate fashion. Sometimes this involves minor revisions to a previously-designed instrument to ensure that questions and scales are interpreted in similar ways across all cultural groups in the study ("cross-cultural equivalence"). Instruments can be developed in multiple languages using a "decentering translation" process, where an instrument is developed in one language, translated to the second language, and then translated back to the original language to verify the quality of the original translation. Alternatively, existing methods and framings may be broadened to encompass a wider range of norms and values, such as redefining one's expectations of what "parental involvement" might look like for a particular cultural group (Garibay, 2006a).

In other situations, translations or adjustments may not adequately address cross-cultural differences, and more innovative methods may be useful. For example, storytelling may be an appropriate research technique for cultures rooted in oral traditions, and photographic methods may be more accurate for cultures that are not highly verbal. Participant observations, in which people share their thoughts with the researcher at times of their own choosing, may be more culturally appropriate for deaf visitors at hands-on exhibits than "think-aloud" protocols (Ericsson and Simon, 1993) where visitors talk and manipulate the exhibit simultaneously. This is because Sign Language, the primary language of the Deaf culture, requires hand movements that can conflict with visitor manipulation of the exhibit (Reich, 2005).

Dissemination of the findings to the communities

Finally, adherents of culturally responsive research believe that disseminating data to the communities who participated in the research is important, and that the results of such studies should be easily accessible to respondents. Dissemination of results improves the validity of the findings by obtaining the community's input (e.g., Reich, 2000), and also serves as a political act by empowering community members with knowledge about themselves (Garibay, 2006b).

Conclusion

In this chapter, we presented two broad approaches for dealing with museum complexity in research and evaluation: reducing it to support cause-and-effect understanding, or embracing it to create a deep understanding of multiple perspectives. Pragmatic considerations such as constraints of time and budget make it virtually impossible to conduct the ideal versions of these approaches in real museums. But researchers who are familiar with a variety of techniques can make informed decisions about when to push for higher quality and greater understanding as resources permit. For example, evaluators working within an experimental paradigm may be able to include a small control group in their summative evaluations, and many formative evaluations can be adjusted to compare alternative offerings rather than just assessing visitors' experiences at one. Researchers wanting to conduct a more rigorous naturalistic inquiry can usually spend longer times with visitors to build more trust, and can always confirm with them at least the most important or controversial interpretations. Case studies can be scrutinized for the credibility of their arguments, not just the thickness of their descriptions. And researchers from *any* paradigm who want to incorporate culturally responsive frameworks may at least be able to diversify their research team and use methods that include the experiences of a previously excluded audience group.

While the two main approaches to complexity have been listed sequentially and separately in this chapter, several researchers and theorists (e.g., Guba & Lincoln, 1989; Patton 2002) support the use of multiple methods to strengthen and triangulate interpretations of data, as long as the combination does not undermine any fundamental theoretical assumptions. For example, researchers who want to determine average holding times in an exhibition should use random sampling techniques when they track and time visitors. However, the results of such a study could be used in conjunction with a naturalistic study that purposively samples visitors for qualitative interviews. The union of these findings could yield an overview of typical traffic patterns and a range of the individual visitor experiences that motivate them. As another example, developers at the Museum of Science in Boston used universal design techniques to modify a diorama-based exhibition to be more accessible for all visitors, including those with disabilities. The summative evaluation (Davidson, 1991; Davidson, Heald, & Hein, 1991) included two main components: First, researchers conducted naturalistic interviews with persons with disabilities and found that the exhibit was accessible and engaging for that population. Next, researchers used random sampling and a pre/post experimental design with typical visitors, and concluded that holding time and conceptual understanding increased as a result of the exhibition modifications. The thoughtful and selective use of the two methodologies produced findings in support of the notion that universal design techniques may benefit all museum visitors.

As a final note, we encourage researchers and evaluators to discuss the values of these various approaches with stakeholders. Often museum practitioners will expect a certain study design because it is simple or familiar to them, but they may quickly embrace a different approach if they understand its purpose and efficacy. More importantly, stakeholders are the people making difficult decisions for action based on the completed studies, so it is important for them to appreciate the limitations of even the best work in social science. Descriptions of context are always limited, generalizations are tentative at best, resources are finite, and human behavior is always more complex than any single study can reveal.

Acknowledgments

This chapter benefited from the contributions and comments of Linda Deacon, Doreen Finkelstein, George Hein, the editors, and three anonymous reviewers.

References

Allen, S. (2004). *Finding significance*. San Francisco: Exploratorium.

American Association of Museums (1992). *Excellence and equity: Education and the public dimension of museums*. Washington DC: American Association of Museums.

Anderson, D., Lucas, K.B., Ginns, I.S., & Dierking, L.D., (2000). Development of knowledge about electricity and magnetism during a visit to a science museum and related post-visit activities. *Science Education, 84*(5), 658-679.

Anderson, G. & Arsenault, N. (2004). *Fundamentals of educational research, 2nd Edition*. London: Routledge.

Bakeman, R. & Gottman, J.M. (1997). *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge: Cambridge University Press.

Campbell, D.T. & Stanley, J.C. (1963) *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Davidson, B. (2001). *Universal Design (Accessibility)*. Retrieved December 30, 2003, from http://www.mos.org/exhibitdevelopment/access/

Davidson, B., Heald, C. L., & Hein, G. (1991). Increased exhibit accessibility through multisensory interaction. *Curator*, 34(4), 273-290.

Ellenbogen, K.M. (2002). Museums in family life: An ethnographic case study. In G. Leinhardt, K. Crowley, & K. Knutson, (Eds.), *Learning conversations: Explanation and identity in museums* (pp. 81-101). Mahwah, New Jersey: Erlbaum.

Ericsson, K.A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Falk, J., Moussouri, T., & Coulson, D. (1998). The effect of visitors' agendas on museum learning. *Curator* 41: 107-120.

Falk, J.H. & Dierking, L.D. (2000). *Learning from museums*. Walnut Creek: AltaMira Press.

Garibay, C. (2006a). *Primero la Ciencia remedial evaluation*. January 2006. Unpublished manuscript. Chicago Botanic Garden.

Garibay, C. (2006b). *Latino audience research for the program in Latino history and culture. Nation Museum of American History*, Smithsonian Institution. Spring 2006.

Geertz, C. (1983). *Local knowledge: Further essays in interpretative anthropology*. New York: Basic Books.

Gill, C. J. (1999). Invisible ubiquity: The surprising relevence of disability issues in evaluation. *American Journal of Evaluation*, 20(2), 279-289.

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Guba, E.G. & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

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

Gyllenhaal, E. D., Perry, D. L., & Cheng, B. (2005). *Sugar from the Sun* formative evaluation (Unpublished manuscript). Chicago, IL: Garfield Park Conservatory Alliance.

Gyllenhaal, E.D. (1998). *Trustworthiness of naturalistic inquiry* (Unpublished manuscript). Chicago: Selinda Research Associates, Inc.

Heimlich, J.E., Storksdieck,M., J. Barlage, J., & Falk, J. F. (2005). *Catalina Island Conservancy: A triangulated study of conservation stakeholders*. Annapolis, MD: Institute for Learning Innovation.

Hood, M. G. (1983). Staying away: Why people choose not to visit museums. *Museum News, 61*(4), 50-57.

Hui-Min, C. and M.D. Cooper (2001). Using clustering techniques to detect usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*. 52(11): 888-904.

Humphrey, T. and J. P. Gutwill (2005). *Fostering Active Prolonged Engagement: The Art of Creating APE Exhibits*. San Francisco, Exploratorium.

Kirkhart, K.E. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice*, 16, 1-12.

Latane, B., & Darley, J. (1969). Bystander "Apathy". *American Scientist*, 57, 244-268.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage Publications.

Millett, R. (2002). Missing voices: A personal perspective on diversity in program evaluation. *The Non-profit Quarterly E-newsletter* 1(12). Retrieved June 18, 2006, from http://www.nonprofitquarterly.org/section/309.html

National Research Council (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Patton, M.Q. (2002). *Qualitative Research and Evaluation Methods.* Thousand Oaks, CA: Sage Publications

Prosavac, E.J. (1998). Toward more informative use of statistics: Alternatives for program evaluators. *Evaluation and Program Planning*. 21, 243-254.

Randol, S. M. (2005). *The nature of inquiry in science centers: Describing and assessing inquiry at exhibits*. Unpublished doctoral dissertation, University of California, Berkeley.

Reich, C. A. (2000). The power of universal design: Building an accessible exhibition. *Dimensions,* July/August, 14-15.

Reich, C. A. (2005). *Universal design of interactives for museum exhibitions*. Unpublished master's thesis, Lesley University, Cambridge, MA.

Rowe, S. (2002). The roles of objects in active, distributed, meaning making. In S. Paris (Ed.), *Perspectives on object-centered learning in museums,* (pp. 19-36). Mahwah, NJ: Erlbaum.

Schaefer, J., Perry, D. L., & Gyllenhaal, E. D. (2002). *Underground Adventure: Final summative/remedial evaluation* (Unpublished manuscript). Chicago: The Field Museum.

Serrell, B. (2000). Does cueing visitors increase the time they spend in a museum exhibition? *Visitor Studies Today* 3(2): 3-6.

Shavelson, R. & Towne, L. (2004). What drives scientific research in education: Questions, not methods, should drive the enterprise. *APS Observer, 17*(4).

Tisdal, C. (2004). *Phase 2 summative evaluation of Active Prolonged Engagement at the Exploratorium*. Chicago, Selinda Research Associates: 115.

Williams, D. D. (n.d.) *Educators as inquirers: Using qualitative inquiry*. Retrieved June 2, 2006, from Brigham Young University, Department of Instructional *Psychology* and Technology, Qualitative Inquiry in Education Web site: http://msed.byu.edu/ipt/williams/674r/TOC.html

Wolf, R. L., & Tymitz, B. L. (1979). *A preliminary guide for conducting naturalistic evaluation in studying museum environments* (Unpublished manuscript). Washington, DC: Office of Museum Programs, Smithsonian Institution.