What to ask and how to answer: a comparative analysis of methodologies and philosophies of summative exhibit evaluation

*Margaret Lindauer**

Virginia Commonwealth University

Abstract

This essay examines the question of how museum professionals select research methods for summative exhibit evaluation. It explores the ways in which this question historically has been answered in the United States, and it argues that selecting appropriate research methods depends upon understanding the interrelationship between research theories, methods, and designs. It also characterizes this interconnection in relation to different kinds of evaluative questions. The main purpose of the paper is to help museum professionals select an approach to summative evaluation appropriate to specific exhibitions and contexts.

Key words: Summative Evaluation, Research Theory and Practice, Methodological Disputes, Educational Merit

Introduction

When I first set out to conduct a summative evaluation of a museum exhibition, I looked at guidelines described in selected publications available through the American Association of Museums bookstore (Korn and Sowd 1990; Diamond1999; American Association of Museums 1999). On the one hand, I found clear distinctions among front-end, formative, remedial, and summative evaluations as well as ample description of data collecting methods that I might use—observing visitor behaviour, conducting interviews, gathering survey responses, or administering pre- and post-visit tests. On the other hand, I looked in vain for adequate discussion of *how* to choose among the options, upon *what* principles qualitative and quantitative methods might be combined, or *why* some data collection methods correspond to statistical analysis while others are more compatible with expository analysis.

I knew from my doctoral coursework and research in education that selecting methods for data collection and analysis ideally occurs in the course of articulating a research design. I turned for guidance to a broader body of literature on evaluation practice (including contributions from the fields of education, psychology, and sociology) and found a vast array of research designs including experimental, quasi-experimental, goals-based, naturalistic, and goal-free (House 1993; Pawson and Tilley 1997). I also found impassioned debate over the ways in which positivist and interpretivist research theories are or are not relevant to the design, implementation, and/or reception of evaluative research (Guba and Lincoln 1989; Pawson and Tilley 1997; Schwandt 2002). These differences of opinion fuelled methodological disputes that are sometimes characterized as 'paradigm wars' or 'paradigm debate' (Gage, 1989: 4; Pawson and Tilley 1997: 2; Schwandt 1997: 108).[1]

As I studied the evaluation studies literature, I realized that the range of research designs and the relevance of research theory to summative exhibit evaluation might be implicitly inscribed in the museum studies literature. When I was awarded a 2004 Fellowship in Museum Practice at the Smithsonian Center for Education and Museum Studies, I decided to explore past publications about, and examples of, summative exhibit evaluations, believing that an explicit discussion of the interrelationship of research theories, designs, and methods would answer the essential questions of *how* to choose among data collection options, upon *what* principles qualitative and quantitative methods might be combined, and *why* some data

collection methods depend on statistical analysis while others are more compatible with expository analysis.

In this paper, I present my research on how, when, and why standards of evaluative practice have developed and diversified in museum contexts, but my overall goal is to transcend mere description of past and current trends. I argue that enacting an appropriate approach to answering any evaluative question has to be firmly grounded upon a theoretical foundation. Different evaluative questions call for different methods, and research theories provide the methodological basis with which to assemble trustworthy data and draw well-founded conclusions. I do not endorse the application of any one theory, research design, or set of methods, but rather I characterize the relationship between 'what to ask' and 'how to answer' in such a way that museum professionals may decide for themselves which approach is most appropriate, to their specific circumstances, for carrying out a summative exhibit evaluation.

I organize my presentation of different research theories and designs as they historically and sequentially have been enacted through summative exhibit evaluations. I do not claim to provide a comprehensive history, particularly because most of the evaluators that I cite have been based in the USA, but rather I characterize broad developments that have generated diverse practices. I also focus on summative evaluations that assess the educational merit or an educational aspect of an exhibition, although not all summative evaluations focus on this. (See St. John and Perry [1993] and Pekarik, Doering and Karns [1999]). In the course of my discussion, I offer succinct definitions of key terms, which may seem overly basic to some readers but have to be included in order to distinguish clearly one approach from another. I conclude by describing four currently recognized approaches to evaluating educational merit and explaining the criteria for selecting one approach or another.

Early history of summative exhibit evaluation in the United States

The first systematic studies of museum visitors were published in the USA during the early twentieth century, when research that enacted a scientific approach to understanding human behaviour was *de rigueur* in the social sciences. The widespread interest in human behaviour was grounded on a presumption that solutions to social problems could be generated through scientific analysis (House 1993). Scientists identified the existence of a problem (e.g. illiteracy, high crime rates, entrenched poverty) through *basic* research, and they carried out *applied* research to examine the extent to which particular social programs effectively redressed the problem. Basic and applied research designs share the same principles for generating trustworthy findings, but have different aims from one another (Martella, Nelson and Marchand-Martella 1999). The aim of basic research is to develop universal laws, identify general patterns, or generate models that account for social or natural phenomena, while the aim of applied research is to understand events that occur in one particular circumscribed setting—a classroom, social service program, museum exhibit, etc. Applied research is evaluative when it focuses on whether or not, or in what ways, a program works (Pawson and Tilley 1997). It may draw criteria for success from basic research. For example, once a theory of museum learning is elucidated through basic research, it can be applied as a standard with which the educational effectiveness of an exhibition may be evaluated. This relationship between basic and applied research can be discerned in the earliest examples of museum-visitor studies, which identified and then set out to redress 'museum fatigue'.

The scientific study of museum fatigue

The phrase 'museum fatigue' was coined in the early-twentieth century when Benjamin Ives Gilman, Secretary of the Museum of Fine Arts in Boston, described photographs of museum visitors stooping over to view adequately various exhibit elements (Gilman 1916). He cited ergonomic principles and recommended that museums place display cases and labels at eye-level. Soon thereafter, a scientific interest in museum fatigue increased when the American Association of Museums was awarded a grant from the Carnegie Corporation of New York and contracted Edward Stevens Robinson, a psychology professor at Yale University, to determine scientifically the existence of the hypothesized phenomenon.

In his empirical identification of museum fatigue, Robinson explained that he applied 'a scientific attitude' to studying visitor behaviour 'because only in such a state of mind can we make common knowledge precise and only thus can we discover the more important and least distrusted of our erroneous opinions' (1928: 10). The 'scientific attitude' that Robinson invoked, refers to a positivist research theory, which historically has been the dominant, though not ubiquitous, research theory for evaluating museum exhibitions (Lawrence 1991). At its roots, positivism is based on the philosophical assumption that there is a concrete reality that exists beyond the human mind or experience. The aim of positivist research is first to discover empirically and then to manipulate predictably aspects of that objective reality. For example, museum fatigue was determined to be a concrete observable phenomenon that existed outside the constructs of social discourse or subjective opinion of the researcher. Upon verifying its existence, Robinson set out to identify ways to manipulate exhibit variables so that museum fatigue would decrease in a predictable way.

Robinson's work is well known in the field of audience research, but a description bears repeating because it exemplifies the relationship between what to ask and how to answer within a positivist-research framework. In his initial study (an example of basic rather than applied/evaluative research), Robinson tracked visitor behaviour in four art museums—noting total time visitors spent in the museum, the number of rooms they entered, the percentage of displayed paintings they viewed, and the amount of time they lingered at each artwork. He found that visitors spent significantly less time looking at artworks toward the end of their visits than they had done at the outset. Robinson characterized his finding as 'an objective and clear-cut demonstration of the existence of a "fatigue" effect in the behavior of the museum visitor', for which, he postulated, there wasn't *one* organic cause (1928: 42). He subsequently set out to explore the effects of different variables on the lengths of time that visitors look at artworks—the size, display location, and relative isolation of artworks; label placement; installation style (art galleries versus period rooms); and architectural features such as room size. Two of Robinson's doctoral students identified other factors affecting attentiveness. Melton (1935) focused on the location of exhibits within a museum, the number and homogeneity of objects displayed, and the number and visibility of exits. And Porter (1938) studied the effects of signs and pamphlets on the length of time visitors spent in an exhibition.

Although their focus was strictly behavioural, Robinson, Melton, and Porter considered their research to be highly relevant to the educational function of museums. As Melton stated, 'the fundamental premise . . . is that public museums are institutions for the education of the public,' who would benefit from 'a science of museum education built on knowledge of the behavior of the museum visitor' (1935: 1). This so-called science was generated through investigation of cause-and-effect hypotheses (i.e., variable *x*, such as label placement, will cause behaviour *y*, such as increased attentiveness), based on the presumption that there was 'a positive relation between the degree of interest shown . . . in the exhibits [by the visitor] and number and quality of the more or less permanent educational results of the museum visit' (Melton 1935: 3).

Three features of scientific research design

Beginning in the 1930s, researchers who investigated cause-and-effect hypotheses to explain visitor behaviour relied on experimental or quasi-experimental research designs, which share three general features— objectivity, reliability, and validity—that, according to positivist theory, collectively ensure trustworthy findings (Martell, Nelson and Merchand-Martell 1999). *Objectivity* means that researcher bias is kept in check as variables are defined, isolated, measured, and statistically analyzed. It is often achieved, in part, by using a non-human instrument for measuring or mapping variables (exhibit elements and visitor behaviour) and by using random or representative sampling methods (described below).

Reliability means that the instrument used for collecting data does not influence visitors' behaviour or response, produces the same results irrespective of the researcher, and offers consistent measurements (e.g., a stopwatch that is not broken can accurately measure the amount of time a visitor looks at individual exhibit elements). As Diamond explains, a reliable instrument 'measures the same thing, . . . in the same way, each time it is used' (1999: 77).

Not all reliable instruments are mechanical. For example, surveys and structured interviews may meet the criteria for reliability if they have been pre-tested to ensure that the ways in which questions are phrased and sequenced don't skew the results.

There are two kinds of validity - internal and external. *Internal validity* means that the variables being measured represent what is being tested. For example, Melton considered the amount of time a visitor stands in front of an exhibition element to signify attentiveness, which in turn indicates 'the more or less permanent educational results of the museum visit' (Melton 1935: 3). But apparent time-on-task does not necessarily indicate attentiveness, thus the internal validity of his study may have been threatened if some visitors, for example, appeared to be engaged in looking at a particular exhibit element while actually thinking about what to pick up at the grocery store. Other visitors may have continued to think about an exhibit element beyond the amount of time they spent looking at it, thereby appearing as if they no longer were attentive. The fuzzy boundaries of attentiveness compound the potential threat to internal validity if visitor behaviour is considered to mark the 'educational effectiveness' of an exhibit. As Diamond cautions, 'If you are using a series of behaviors to determine whether learning has occurred, then you need to be sure that the behaviors are valid indicators of learning' (1999: 75).

External validity means that the research participants (visitors who are observed, tested, or interviewed) accurately represent an overall population to which the findings are generalized. This is achieved through random or representative sampling. The use of one of these two sampling methods distinguishes, in part, an experimental from a quasi-experimental research design. In classic experimental design, research subjects who share salient characteristics (e.g., intellectual ability, education level, prior museum experience) are *randomly* assigned into two groups (meaning that each member has an equally random chance of being placed in one group or another). One group receives a treatment or exposure to a programme, which is a hypothesized cause (independent variable) in a cause-and-effect relationship, while the control group does not. A marker of the hypothesized effect (dependent variable) is measured in all participants before and after the experimental group receives the treatment. Aggregate measurements of the experimental group are compared to those of the control group, indicating the extent to which the hypothesized cause did indeed generate the hypothesized effect. For example, Melton conducted a series of experiments wherein school children, each of whom were 'equivalent [to one another] in every important respect' (e.g., age, school grade and performance, intelligence quotient) were divided into two groups and offered exactly the same pre-visit experiences 'except for the alteration of one phase of the entire plan of instruction'. A post-visit test was administered to both groups and the average scores from each 'were used to determine the relative effectiveness of the altered phase of the plan of instruction' (Melton 1936: 9).

Within positivist theory, classic experimental design generates the most trustworthy findings (relative to other research designs), but it is not always tenable in real-life settings. For example, random sampling used in experimental research design is not feasible when the effects of one variable or another on *all* visitors to an exhibition are being studied. The entire population cannot be circumscribed and divided into comparable groups in the way that a classroom of fifth-graders, for example, can be separated into two or more groups of equal intellect, ability, and previous knowledge. Given these real-life circumstances, most exhibit evaluators who subscribe to positivist theory have used representative sampling and enacted quasi-experimental research design.

Representative sampling means that visitors selected for observation, survey, or interview represent salient characteristics of the entire population of visitors to an exhibition. Determination of salient characteristics - demographic, social, or behavioural features - depends upon the particular evaluative research question. For example, an evaluation may set out to explore whether an exhibition communicates more effectively with one category of visitors than with another. Demographic categories may be based on age, gender, ethnicity, education level, socio-economic status, and/or place of residence. The social context in which visitors attend an exhibit - alone, as a couple, in a group of peers, or with family members - may be a salient feature. A behavioural distinction, such as how much time a visitor spends at an exhibit, may also be a significant variable. Representative sampling is accomplished when the sample size of each category of visitors with pre-determined salient characteristics matches

the proportion of that group in the overall population of museum visitors, and when the selection of participants occurs through *probabilistic* sampling methods (e.g., a systematic selection wherein every fifth visitor fitting a particular category is observed or interviewed). When the evaluative research question does not articulate a hypothesized relationship among salient demographic, intellectual, or behavioural variables, participants may be selected to represent *all* visitors rather than categories of visitors. In this case, external validity also rests upon using probabilistic sampling methods, the particulars of which ideally are statistically determined in relationship to the size of the overall population.

I belabour the criteria for generating trustworthy findings according to principles of a positivist theory in the same way that researchers who conducted studies of museum visitors from the 1930s through the early 1970s collectively explored and debated appropriate data collection and analysis methods. An annotated bibliography by Elliot and Loomis (1975) offers a snapshot of how researchers variously built upon one another's findings in relationship to four research problems or issues. Some studies set out to isolate the effects of various environmental variables (label size, colour, font, word count, orientation signs, lighting, and exhibit size) on museum visitor behaviour. Other evaluations focused on the effects of human factors (education level, existing interest or knowledge, social group, and gender) on information retention. A third type of question explored the reliability of various instruments (multiple choice questionnaire, survey, pre-test/post-test comparison, and tracking and timing visitor behaviour). A fourth group of studies examined the validity of a particular outcome as an indication of learning (e.g., information retention, correct responses to post-visit test, attitudinal change, or number of exhibit elements to which visitors attended).

The impact of accountability in the United States

By the end of the 1970s, summative exhibit evaluation had become a broadly endorsed (though not necessarily common) museum practice in the USA. This surging interest was not only because earlier studies set a precedent but was also in response to broader societal demands. The Great Society legislation passed during Lyndon Johnson's presidency in the 1960s mandated evaluation of federally funded social and educational programs (House 1993; Pawson and Tilley 1997).[2] At the same time that this legislation ushered in a broad societal focus on accountability, the American Association of Museums (AAM) was lobbying for museums to be federally recognized educational institutions so that they could qualify for education funding. In *America's Museums: The Belmont Report*, a 1969 study of USA museums presented to President Johnson, members of the Federal Council on the Arts and Humanities stated, 'Because museums are not defined by the federal government as educational institutions, they are denied certain tax concessions and foreclosed from certain federal grants provided to educational institutions' (AAM 1969: 38). This status changed when granting agencies began specifically listing museums among the institutions eligible for education program funding (Zeller 1996). As governmental recognition of the educative social purpose of museums was achieved in an era of accountability, the following logic took hold: if museums are recognized educational institutions and receive taxpayers' money for producing educational exhibitions, then they must demonstrate that exhibits educate effectively. This requirement generated a shift toward, as well as a critique of, goals-based evaluative research design (described below).

The response box

In the early 1970s, Chandler Screven, a faculty member in Department of Psychology at the University of Wisconsin, Milwaukee, conducted a series of exhibit evaluations in which visitors used a 'response box' at various points within an exhibition. The response box, which was also implemented at Smithsonian museums by Robert Lakota, took different forms—either an audio cassette player accompanied by punchboard or an automated slide machine that presented questions to which visitors responded by pressing one of several buttons placed alongside multiple choice answers. Visitors could use the response box to test their acquisition of factual knowledge as they proceeded through displays. The box recorded the results so that researchers could aggregate and statistically analyze the results.

Screven maintained that exhibitions ought to have specific quantifiable learning objectives because 'without predefined goals and some way of measuring whether or not they have been achieved, there is no scientific basis for evaluating existing displays or designing new ones' (Screven 1974: 11). He advised exhibit developers to predetermine the percentage of achieved objectives that would constitute an educationally effective exhibition. With clearly articulated learning outcomes and a predetermined numerical measure of success, the evaluator could then apply the principles of positivist research to assess the exhibit. Indeed Screven explicitly stated that he used 'experimental procedures in which particular variables were systematically manipulated to examine their effects' (1974: 9).

Screven's goals-based evaluation depended upon visitors independently feeling compelled (without invitation or urging from the evaluator) to use the response boxes situated at several points in an exhibit. Thus he also researched ways in which different visual appearances and placements of the response boxes affected the percentage of visitors who engaged the device. To a certain extent, the need to assess could become like the tail wagging the dog, as exhibit designs were shaped by the requirements for (as opposed to the results of) scientific investigation. This would be analogous to a classroom instructor teaching for the test rather than teaching to the skills and interests of learners or according to the richness and complexity of a particular topic. This relationship between teaching and testing has been criticized widely in the field of education because it limits the range of teaching methods and subjects, even as (or perhaps because) it is ubiquitously imposed upon public school teachers. Similar criticism was invoked at the1977 Museum Evaluation Conference (MEC), as the presumption that evaluation necessarily required a strict adherence to scientific principles for investigation was disputed.

Smithsonian Museum Evaluation Conference

In 1977, the Smithsonian Office of Museum Programs sponsored a conference that focused solely on the issue of evaluation, inviting selected university professors, researchers, and museum professionals to contribute to a roundtable discussion. Most of the participants, whose comments were published in the Museum Evaluation Conference proceedings (1977), indicated their affiliation with positivist research theory as they expressed concern for accountability and acknowledged particular challenges associated with evaluating museum learning. For example, Harris Shettel, a psychologist at the American Institute of Research in the Behavioral Sciences, forewarned, 'Public accountability will inevitably become an integral part of . . . museum operations' (MEC 1977: 6). Screven reiterated a vision of the exhibit development process as being analogous to articulating and testing a hypothesis through a goals-based research design built upon positivist principles. He asserted that evaluation is grounded in three questions, 'First, what impact do you want and on whom? Then, how are you going to attempt to achieve this impact via your program or exhibit? Thirdly, how will you know if you have the desired impact on the audience?' (MEC 1977: 3). Minda Borun, an evaluator at the Franklin Institute Science Museum and Planetarium in Philadelphia, also noted the evaluative challenge posed by the fact that museums attract 'a heterogeneous audience . . . [who] come by choice, seeking pleasure and diversion, as well as information' (MEC 1977: 7).

As conference participants discussed the challenges associated with scientifically evaluating museum exhibitions and addressing the federal mandate, Robert Wolf, an evaluator at the Indiana Center for Evaluation, in the School of Education at Indiana University, explicitly dissented from the majority opinion that positivist theory offered the most suitable principles for exhibit evaluation. As reported in the conference proceedings, Wolf cautioned that studies focusing on 'measurement, quantification, [and] causality . . . detract from the creativity and spontaneity that occurs in educational settings' and 'are often very insensitive to the multiplicity of goals . . . [and to the diversity of visitors'] aspirations and expectations' (MEC 1977: 20). He accordingly endorsed 'naturalistic inquiry', inscribing what later would be called interpretivist research theory (described below) and using qualitative data collection and analysis methods. He argued that naturalistic evaluation was better suited to 'understanding actualities, social realities, and human perceptions' than 'the obtrusiveness of formal measurement or preconceived [survey] questions' (MEC 1977: 22).[3]

Positivist versus interpretivist research theories

On the surface, the differences between positivism and interpretivism appear to be purely methodological, but underneath a positivist *tendency* toward quantitative data collection versus an interpretivist *tendency* toward qualitative data collection rests a philosophical distinction. Whereas positivism is grounded on the assumption that an objective world exists outside of human constructs, proponents of an interpretivist theory contend that, in the social realm, there are multiple realities that cannot be fully or objectively understood, partly because they are constantly changing in relationship to one another and also because they always are observed through a cultural lens (Jacob 1992). The aim of interpretivist research therefore is to understand the complexity of what variously occurs in a specific social context by taking account of participants' multiple subjectivities, opinions, and perspectives regarding particular events, processes, or outcomes.

Both positivist and interpretivist theories begin with philosophical assumptions and overarching aims; neither is better nor worse than the other in this respect. And it is from these distinctly different philosophical footings that methodological differences unfold. Whereas a positivist theory rests on the criteria of objectivity, reliability, and validity for judging the trustworthiness of findings, an interpretivist theory calls for credibility, authenticity, and comprehensiveness as markers of trustworthy research (Guba and Lincoln 1986).[4]

The two sets of criteria—'objectivity, reliability, and validity' versus 'credibility, authenticity, and comprehensiveness'—are parallel to one another. The positivist emphasis on objectivity is analogous to the interpretivist emphasis on credibility insofar as both characterize the relationship between researcher and research participants. *Credibility* rests upon the evaluator demonstrating that his/her interpretation of what's happening is grounded in data—observations, interviews, and written documents—rather than simply personal opinion. As Wolf and Tymitz (1979) suggest, an exhibit evaluation ideally is a descriptive report written in the language of program participants. The researcher accordingly sets out to understand, to describe, and to analyze opinions and perspectives among both visitors and exhibit developers without predetermining specific indices that signify success. Thus the process is akin to investigative reporting, focused on *understanding* what visitors and developers value rather than *judging* their actions and interpretations, and it is directed toward generating a report that gives enough detail to instill the readers' trust in the researcher's findings.

The positivist criterion of reliability is analogous to the interpretivist criterion of authenticity insofar as both characterize the means through which trustworthy data is collected. *Authenticity* refers to the researcher's ability to elicit participants' opinions, interpretations, and values. In an interpretivist evaluation, the researcher is the instrument for collecting and analyzing data. This means, as Wolf and Tymitz (1979) explain, the evaluator aims to make visitors and exhibit developers at ease by talking in their own terms about their expectations, hopes, opinions, and experiences. It also means adjusting the pace and rhythm of a conversation based on the interviewees' apparent comfort level and altering the content or order of interview questions to address what interviewees say. Each interview unfolds in a way that feels more like a casual (albeit focused) conversation rather than a list of questions. Thus each interview is unique because each visitor and each member of an exhibit development team is unique.

The positivist criterion of validity is analogous to the interpretivist criterion of comprehensiveness insofar as both articulate the relationship between collected data and breadth of analysis. *Comprehensiveness* refers to the evaluator showing that s/he has not based an interpretation or conclusion on any single piece of evidence or simply one visitor's experience. In a museum context, the evaluator ideally looks at a range of data sources, speaking individually to people involved in developing an exhibition, looking at written accounts of policies and/or process, examining the exhibition itself, watching visitors go through the exhibit, and speaking to a wide range of visitors (people attending alone, as family groups, or with friends; people of diverse age, racial, or ethnic groups; and people from various geographic locations) in order to elicit a wide range of perspectives. The point however is not to represent social and demographic diversity (in the sense of a positivist principle of representative sampling) but rather to talk to an array of visitors as a means to discover and understand the diverse ways in which visitors experience and/or value an exhibition.

Comprehensiveness rests on analyzing a *range* of exhibit developers' and visitors' opinions, accounts, and expectations. And it is analogous to a positivist concern for generalizability in terms of 'case-to-case transfer' rather than external validity (Lincoln and Guba 1986). To accommodate case-to-case transfer, an evaluative report includes sufficient detail for the reader to recognize general ways in which findings from one context can be applied to another context.

These overarching analogues mark the distinguishing features between the two theories. Data collection methods—quantitative versus qualitative— do not in-and-of themselves distinguish a positivist from an interpretivist evaluation (Patton 2002). Researchers subscribing to either theory may gather numerical data and/or participants' opinions or recollections. A positivist researcher collecting qualitative data uses a formal instrument (e.g., pre-tested survey), transforms that data into numerical indices that can be statistically analyzed, and generalizes findings beyond the specific context or sample from which data was collected. An interpretivist researcher who collects numerical data incorporates it into an expository analysis only insofar as those data contribute to a descriptive account of the evaluative context, process, and/or findings. Quantitative and qualitative data collection *methods* can be used in tandem for summative exhibit evaluation. But positivist and interpretivist *theories* are philosophically inconsistent with one another; as positivist research seeks to understand an objective phenomenon that exists outside of human constructs while interpretivist sets out to understand a social realm as it is experienced through human filters. These philosophical differences historically have fuelled impassioned debate.

Methodological disputes

Within the social sciences, researchers who have subscribed to either positivist or interpretivist theories have engaged in methodological disputes or, *pace* Gage (1989), 'paradigm wars' in which the advantages of one theory have been cast alongside the disadvantages of the other. Facets of these disputes can be found in museum studies literature beginning in 1920s and continuing through the 1990s. For example, Robinson cast doubt upon research that does not meet the positivist criterion of objectivity when he asserted:

> The casual visitor to the museum has not usually had psychological training and there are few reports so untrustworthy as those of an unpracticed observer regarding what he thinks he thinks and what he feels he feels. . . . Why should we seek such personal revelations when we know from sorrowful experience that they are sure to be more false than true? (Robinson 1928: 11)

Nearly seventy years later, Roger Miles, a palaeontologist and former director of the Department of Education and Exhibitions at the Natural History Museum in London, declared that 'the best of evaluation practice can indeed be said to be positivist' (Miles 1993: 29) because findings can be generalized, whereas interpretivist findings 'can reasonably be suspected of eccentricity' because 'they're subjective' (Miles 1993: 30). He cautioned that interpretivist approaches posed a 'danger to exhibit evaluation' insofar as 'objective assessment becomes impossible, argument and counter-argument are reduced to a slanging match, and any hope of progress is crippled' because evaluators subscribing to either theory become distracted by 'silencing the opposition' (Miles 1993: 30).

Within the interpretivist camp, positivist studies similarly were judged to be lacking. In the Museum Evaluation Conference proceedings Wolf bemoaned that a scientific approach to evaluation, which asks all visitors the same questions, 'presume[s] . . . that only those questions are critical and that each respondent finds them provocative, interesting, and important' (MEC 1977: 22). Mary Ellen Munley, who worked as an assistant to Wolf and later became the Director of Education at the Field Museum of Natural History, Chicago, characterized the questions explored from a positivist approach as being analogous to looking for lost keys under the light of streetlamp after dark, not because that's where you dropped them but rather because that's where the light is best (Munley 1987: 120). And Ghislaine Lawrence, Senior Curator at the Science Museum in London, bemoaned the fact that, despite a methodological shift in sociology fuelled by 'the inappropriateness of using methods modeled on those of the

natural sciences to study social phenomena' (Lawrence 1991: 15), evaluation practices within museums had changed very little. Lawrence disparaged, 'A large number of museum evaluators . . . remain largely preoccupied with the accurate observation of behaviour' (Lawrence 1991: 25), which is well suited to scientific research methods but not necessarily fruitful for exploring the complexity of visitors' experiences.

The vituperative tone of methodological disputes has gradually diminished as the two theories are recognized to be like apples and oranges; neither is *inherently* better than the other. Rebuking an interpretivist research design for lacking objectivity or generalizability is like discarding an apple because its peel is red, not orange. Likewise, castigating a positivist research design because it pre-determines specific variables to be measured is like rejecting an orange because it already is divided into discernible sections whereas an apple is not. The once-rancorous disagreement over methodologies in fact has subsided to the point that some evaluators argue that identifying a research theory is unnecessary and even misguided when theoretical affiliation is privileged over pragmatic issues, such as available time and resources for carrying out a study (Pitman and Maxwell 1992). But the merits of individual evaluative studies can be judged only according to criteria for trustworthiness outlined in the research theory with which the evaluation has been designed (even if the relationship between theory and design has not been formally articulated). Thus methodological disputes are resolved not by denying the relevance of theory but rather by understanding how different theories relate to different kinds of evaluative research questions.

What to ask and how to answer

In the current era, the decision regarding what to ask can be more daunting than it was when a researcher fervently subscribed to a singularly 'correct' approach to evaluation. Not only are there more professionally recognized options for research design, but also the standards for 'good' design have evolved within both positivist and interpretivist paradigms. One way to navigate through options for developing (or selecting) a research design is to examine ways in which the question 'what counts as learning?' is currently answered (either implicitly or explicitly) within different evaluative approaches and how those approaches correspond to research theories and methods.

From positivist to postpositivist research designs

An evaluation grounded on the question 'did the exhibition effectively communicate the main idea or message?' corresponds to a definition of learning as the reception of pre-defined knowledge. The question is well suited to a positivist research design (experimental, quasi-experimental, or goals-based) because it poses a cause-and-effect relationship—attending an exhibit will cause visitors to acquire particular knowledge or information. It therefore alludes to dependent and independent variables that can be isolated, quantified, and statistically analyzed. The only remaining issue is to determine a numerical definition of the term 'effectively communicate'. This issue can be addressed by answering the question, 'What percentage of visitors must correctly identify the main idea or message in order to signify that the exhibit accomplished its educational goal?'

Beverly Serrell, director of the private consulting firm Serrell and Associates, has addressed this question through a meta-analysis of data from numerous behavioural studies of museum visitors. Serrell (1998) asserts that visitors are best equipped to identify the main theme or idea when they are adequately engaged in an exhibit. She quantifies the notion of engagement by measuring the amount of time visitors spend per square foot (sweep rate index) and the number of stops they make in an exhibition (percentage of diligent visitors). After analyzing results from over eighty exhibit evaluations, she determined both average and exceptional rates for each of these factors, with which any exhibition may be compared. She also established a desired percentage of visitors who ought to be able to identify the main theme or idea, which she correlates with the other variables. She maintains that exhibitions in which visitors spend more time and engage with more components are more apt to elicit correct identification of a main idea or theme. Her approach exemplifies a deductive reasoning process, beginning from a *general* stated premise—an exhibit can be deemed a success if at

least thirty percent of visitors stops at fifty percent of the exhibit elements, traverses an exhibit at a rate of no more than three hundred square-feet-per minute, and correctly identifies a main theme or idea that exhibit developers intended to communicate. After collecting and analyzing data, the evaluator moves from the general stated premise to a *particular* conclusion—the exhibition exceeded, matched, or fell short of the average rate.

While a number of evaluators have used the sweep rate index and percentage of diligent visitors formula, others have criticized it for being too narrow (Serrell 1998). Among the latter, some evaluators maintain that because visitors enter a museum exhibition with unique sets of prior experiences and knowledge, they may experience a variety of learning outcomes, including but not limited to exhibit developers' intended message or main idea. The term 'free-choice learning', which was coined by John Falk and Lynn Dierking (2002), director and associate director of the Institute for Learning Innovation, alludes to the difficulty of narrowly predicting educational outcomes.

Evaluation based on predetermined teaching objectives, which is standard practice in formal educational institutions, would not necessarily assess the new knowledge that visitors take away from an exhibition. Thus some evaluators trained in a positivist approach have explored methods of assessment that recognized a range of outcomes, and a postpositivist orientation has emerged.

Postpositivism developed in response to critiques of positivism and maintains that knowledge is conjectural rather than built upon absolutely secure foundations, and the purpose of scientific research is to generate warranted assertions (Phillips and Burbules 2000).[5] In a postpositivist approach that adopts scientific data collection and analysis methods, the three criteria for ensuring trustworthy results—objectivity, reliability, and validity—are generally maintained but strict adherence to them is relaxed. For example, researchers at the Institute for Learning emphasize validity over reliability, in assessing learning. They have characterized 'learning' as *change* among visitors' understanding of a key concept or idea, and they have used an instrument called 'personal meaning maps' to measure change between visitors' pre- and post-visit comprehension (Adams, Falk and Dierking 2003). As the researchers acknowledge, personal meaning maps may not produce consistent results irrespective of the individual researcher (as prescribed by traditional positivist standards for reliability) but they do allow for a broadened definition of 'learning' that takes into account the fact that visitors 'make what meaning they can from museum experiences regardless of . . . what museum staff intended or expected them to make' (Adams, Falk and Dierking 2003: 16). Strict positivist standards would fault the research for threatening objectivity to the extent that researchers might implicitly direct (perhaps even predispose) visitors to attend to a concept and idea and then ask how the exhibition affects visitors' understanding of that concept. But a postpositivist approach pragmatically accepts that social phenomena, such as 'making meaning', do not exist objectively (outside human experience and interaction).

Personal meaning maps capture the ways in which visitors 'make meaning' while also providing a method for generating quantitative data that can be statistically analyzed. Data collection begins before visitors enter an exhibit, when they are offered a blank piece of paper; told a specific concept, word, or phrase that has been pre-tested to elicit meaningful responses; and asked to write down the words, images, phrases, or thoughts that come to mind. After going through a display, a visitor is given the personal meaning map s/he already created and asked to make changes, annotations, deletions, etc. During both pre- and post-visit exercises, the researcher asks visitors for elaboration or clarification, noting salient points on the same piece of paper. After data collection is complete, researchers examine the personal meaning maps in order to determine categories of changes in knowledge, 'not just what visitors learn but also how much and to what depth and breadth learning occurs' (Adams, Falk and Dierking 2003: 23). They then quantify the degree of changes within those categories and determine aggregate change between visitors' pre-visit and post-visit comprehension. Unlike the traditional positivist deductive process, this postpositivist research design involves both inductive reasoning (moving from the particular data to determine general data categories) and deductive reasoning (moving from the general principle—that change in visitors' knowledge of a key concept or idea signifies the educational value—to a particular conclusion regarding the 'success' of an exhibition).

From interpretivist to humanist research designs

The principles of free-choice learning that precipitated development of postpositivist exhibit evaluation also resonate with interpretivist research theory, since the latter allows for an open-ended range of educational outcomes. In an interpretivist approach, however, research designs are goal-free as opposed to goals-based; rather than articulating a definition or index of learning in advance of collection data, evaluators sets out to explore the range of ways in which visitors are engaging with an exhibition (which may involve cognitive, affective, and/or personal understanding).

Interpretivist research designs are often iterative, meaning that they take shape as data collection and analysis proceed. Specific features of interpretivist evaluations are therefore less codified than features of positivist or postpositivist approaches, although there are overarching commonalities among them. An interpretivist evaluator typically begins with *open-ended exploratory questions* such as 'What is the range of ways in which visitors use and understand exhibit components?' Or, 'What is the range of educational experiences that this exhibit accommodates or elicits?' In an inductive reasoning process, the evaluator works from the specific collected data (generating data categories during, rather than in advance of, data analysis) to articulate a general assessment of the way in which an exhibit functioned educationally.

Insofar as data collection does not focus on predetermined behavioural, cognitive, or affective variables, an interpretivist summative evaluation does not set out to *judge* whether or not an exhibition effectively accomplished developers' goals/objectives but rather to *identify* educational experiences in which visitors engaged. For example, when Eric Gyllenhaal and Deborah Perry of Selinda Research Associates conducted an evaluation of computer interactive components in the Science Museum of Minnesota's *Atmospheric Explorations* exhibit, the questions that they articulated at the outset of the study asked 'how visitors were using the components' and what 'messages they were taking away from their experience' (Gyllenhaal and Perry 1998: 25). The evaluators used two methods of data collection—visitor observation and in-depth open-ended interviews—neither of which had a predetermined set of data categories or questions. Thus interviews proceeded in 'unanticipated directions and yielded information that . . . otherwise [would have been] missed' (Gyllenhaal and Perry 1998: 27). In an inductive process, Gyllenhaal and Perry generated data categories from which they drew their conclusions. Visitors displayed and/or described three primary modes of engagement, namely: exploration, broadly-focused play, and goal-directed play. Visitors also collectively took away three kinds of messages: a concrete cognitive understanding of the phenomena illustrated in the interactive computer programs, a visceral understanding of the phenomena, and/or personal associations or memories linked with the phenomena.

The study by Gyllenhaal and Perry exemplifies the interpretivist principles described earlier in this paper (credibility, authenticity, and comprehensiveness). Other studies illustrate that, like the relationship between positivism and postpositivism, interpretivism has been a point of departure for other theoretical approaches. For example, in an open-ended interpretivist stance toward what counts as 'learning', knowledge of oneself may be just as educationally valuable as knowing about someone or something else. Thus data collection methods may include autobiographical reflection or artistic production in response to a museum experience. When that reflection occurs *during* the learning/teaching experience and evaluation is folded into rather than occurring after a pedagogical process, a humanist research theory is applied (Schwandt 2002).

Irene Stylianides, a primary school teacher in Paphus, Cyprus, carried out a humanist evaluation of her museum experience in the Sainsbury Centre gallery at the University of East Anglia in Norwich. In her study (2003), she set out to understand her intensely negative and visceral reaction to two specific works of art. Her study was self-reflexive, but her motivation was pedagogical rather than solipsistic, as she wondered whether or not (or in what ways) to convey to students that aesthetic experiences are not always pleasant. Her research process involved prolonged visual study of artworks that she repeatedly had rushed past in order to enjoy others. She kept a diary of her stream-of-consciousness thoughts, which drifted toward painful personal memories that she had not anticipated remembering. Only through prolonged

looking and reflecting was she able to recognize the subconscious association she had been making as she previously avoided the artworks. Once she recognized it, she was able to distinguish her reaction to artworks from her reaction to recalling life experiences.

Her autobiographical research was pedagogically useful insofar as it made her 'more able to help children standing in front of an art object . . . accept that what is happening is connected to their experiences elsewhere, beyond the moment in the gallery' (Stylianides 2003: 164). She thereby would teach students that one of the reasons that viewers have unique museum experiences is because artworks sometimes stir deeply held beliefs, memories, or feelings. The simultaneous activities of experiencing artworks, evaluating the experience, and applying evaluation findings to teaching strategy exemplifies a humanist evaluation wherein, 'Neither the evaluator nor the practitioner is thought to face a problem to be solved as much as a dilemma or mystery that requires interpretation and self-understanding' (Schwandt 2002: 69).

Humanist research theory, like interpretivism, typically involves an iterative process and accommodates a wide range of research designs. Thus a description of one humanist evaluation does not represent standards for all, but it can provide a sense of the ways in which humanism and interpretivism are similar to and different from one another. The trustworthiness of Stylianides's conclusions can be judged according to the spirit (rather than strict application) of interpretivist principles. She establishes credibility (e.g., her research is not merely autobiographical but also relevant to pedagogical practices) by situating her study within a context of scholarly works. As she explains, her work contributes to a body of evaluative studies that emphasized the relationship between visitors' previous experiences and exhibit outcomes by setting out 'to appreciate more fully the relationship between . . . prior experiences . . . [and] new experiences in the gallery' as well as 'the correlation between these [new experiences and] . . . anticipation of future experiences' (Stylianides 2003: 155-6). The authenticity of her data is conveyed through detailed description of how it was generated by diary entries. The trustworthiness of her data is judged in the same way that a reader assesses the believability of any memoir or autobiography. In this sense, a humanist approach to evaluation is more akin to artistic narrative than it is to traditional social science. Comprehensiveness is least applicable among interpretivist principles to a humanist approach insofar as a single data source is perfectly acceptable. However a formal report should be purposefully 'useful' in some way to museum educators, even if only as inspiration or motivation to try something new.

In humanist approaches, evaluation is not designed to determine whether or not desired outcomes are achieved. In fact it need not take on even the vestiges of a formal study and can be carried out in the form of day-to-day judgments regarding what has 'worked' or what would be a 'good' way to convey an idea or accommodate a range of meaningful experiences. In the words of Elizabeth Vallance, former Director of Education at Saint Louis Museum of Art:

> We know, as any good teacher does at the end of the day, whether the program has been 'good' or disappointing, using qualitative measures and sheer professional judgment. And we do not believe all of our programs are equally good: we make distinctions routinely and can assure any potential funder about which sorts of programs are most likely to be worth an investment. (Vallance 1996: 236)

Deciding what to ask

In his classic account of what he called paradigm wars in educational research Gage argued 'that programs of research that had often been regarded as mutually antagonistic were simply concerned with different, but important topics and problems' (Gage 1989: 7). The same holds for museum visitor research: contrary to the tenor of late-twentieth century paradigm wars, there is no exclusively 'correct' approach to summative exhibit evaluation. Positivist, postpositivist, interpretivist, and humanist approaches to designing evaluative research can all generate trustworthy results. But the nature of those results, and the criteria for judging their trustworthiness, varies from one approach to another. When an evaluation focuses on assessing educational merit, the process of deciding what to ask and how to answer should be grounded in identifying

the educational principles with which an exhibit was developed. For example, an exhibit with pre-determined quantifiable goals/objectives calls for a positivist or postpositivist approach to addressing a cause-and-effect or goals-based question: Does exhibit attendance bring about the desired cognitive, affective, and/or behavioural outcomes among visitors? Whereas an exhibition developed with a more loosely circumscribed or open-ended stance toward what counts as learning calls for an interpretivist approach to addressing a goal-free question: What is the range of educational experiences that the exhibit accommodates or elicits? When an evaluation is carried out in order to generate ideas for new pedagogical practices, a humanist approach to addressing a self-reflexive question can generate useful findings.

Deciding what question to ask, however, does not rest solely on understanding the distinctions between various research theories and methodologies. In the broadest sense, 'evaluation' refers to the process through which people characterize the 'value' of a planned program, such as a museum exhibition. Different stakeholders (exhibit developers, museum administrators, funding agencies, and audience members) may espouse divergent values. Policymakers and exhibition-developers can always find reasons (pursuant to their divergent values) to disregard findings that do not support their institutional, pedagogical, or curatorial agendas (Krug 2001). Thus the most *useful* evaluative question (one that will generate a report that stakeholders value) is articulated in response to a number of other questions: For whom and/or for what purpose will the evaluation be conducted? Does the funding agency endorse specific markers of educational success? If so, does the agency attach those markers to a judgment of fiscal accountability? Exhibit development team members do not necessarily share pedagogical values (Lindauer 2005); will they look to evaluative findings to adjudicate their philosophical differences? What are the fiscal implications for future curatorial and educational programs; will administrators look to the findings as they make budget allocations? Once audience and purpose are identified, further questions arise: Do some stakeholders place more stock in one research approach than another? When stakeholders espouse divergent values and/or endorse different research approaches, can negotiation lead to consensus? If not, whose values matter most?

The evaluator is not a stakeholder but rather a service provider. His/her methodological preferences or demonstrated skills ought not determine the evaluative question. Instead a museum should hire an evaluator with training and skills appropriate to answering the evaluative question that stakeholders have deemed to be most useful. Stakeholders who are involved in deciding what to ask, and museum staff members who hire evaluators, should therefore understand the interrelationship between research theories, methods, and designs. They should understand that after an overarching evaluative research question is articulated, the research theory to which it correspond indicates appropriate data collection and analysis methods. After evaluation is complete, the trustworthiness of findings can be judged according to criteria associated with the corresponding theory.

Notes

[1] The term 'paradigm wars' was adopted among educational researchers in the USA during the 1980s to characterize communities of researchers aligned along theoretical differences (positivist versus interpretivist) and engaged in methodological disputes (Gage, 1989:10). Although the term 'paradigm' calls to mind Thomas Kuhn's 1962 monograph *The Structure of Scientific Revolutions*, and although Gage mentions Kuhn in passing these educational researchers appear to be using the concept in their own distinctive way.

[2] Similar government-mandated emphasis on accountability among cultural institutions occurred in Canada and the United Kingdom in subsequent decades in the late-twentieth century. (See Galloway and Stanley 2004).

[3] The term 'naturalistic' was coined to emphasize the difference between conducting studies in environments where phenomena occurred 'naturally' (e.g., schools, social service agencies, etc) as opposed to 'artificial' settings of laboratories. The term 'interpretivist' replaced 'naturalistic' to emphasize the theoretical and analytical differences rather than

data collection context or methods. Hereafter and for the sake of consistency I use the more current term, 'interpretivist'.

[4] The criteria associated with interpretivist theory are less codified than criteria outlined in positivist prescriptions. Indeed interpretivist researchers collectively do not necessarily all share the terms 'credibility', 'authenticity', and 'comprehensiveness' to characterize their methodological standards or principles. But they do generally agree on the nature of those principles and how those principles compare to positivist principles.

[5] Postpositivism is not a 'unified "school of thought", for there are many issues on which postpositivists disagree' (Phillips and Burbules 2000: 25-6). In fact, the term 'postpositivism' sometimes is used to refer to researchers 'who seek to bury the positivist tradition' as well as to those who acknowledge the limitations of positivism but nonetheless seek to redeem aspects of it (Baronov, 2004: 58). It is the latter group to whom I refer in this paper.

References

Adams, M., Falk, J. H. and Dierking, L.D. (2003) 'Things Change: Museums, Learning, and Research', in Maria Xanthoudaki, Les Tickle and Veronica Sekules (eds) *Researching Visual Arts Education in Museums and Galleries: An International Reader*, 15-32, Dordrecht, The Netherlands: Kluwer Academic Publishers.

American Association of Museums (1969) *America's Museums: The Belmont Report*, Washington, DC: American Association of Museums.

American Association of Museums (1999) *Introduction to Museum Evaluation*, Washington, DC: American Association of Museums.

Baronov, D. (2004) *Conceptual Foundations of Social Research Methods*, Boulder, CO: Paradigm Publishers.

Diamond, J. (1999) *Practical Evaluation Guide: Tools for Museums and Other Informal Educational Settings*, Walnut Creek, CA: Altamira Press.

Elliot, P. and Loomis, R. (1975) *Studies of Visitor Behavior in Museums and Exhibitions: An Annotated Bibliography of Sources Primarily in the English Language*, Washington, DC: Office of Museum Programs.

Falk, J. H. and Dierking, L. D. 2002, *Lessons Without Limits: How Free-Choice Learning is Transforming Education*, Walnut Creek, CA: Altamira Press.

Gage, N. L. (1989) 'The Paradigm Wars and Their Aftermath: A "Historical" Sketch of Research on Teaching, Since 1989', *Educational Researcher* 18 (7) 4-10.

Galloway, S. and Stanley, J. (2004) 'Thinking Outside the Box: Galleries, Museums and Evaluation', *Museum and Society*, 2 (2) 125-146.

Gilman, B. I. (1916) 'Museum Fatigue', *Scientific Monthly* 12, 62-74.

Guba, E. and Lincoln, Y. (1989) *Fourth World Evaluation*, Newbury Park, NJ: Sage Publications.

Gyllenhaal, E. D. and Perry, D. L. (1998) 'Doing Something About Weather: Summative Evaluation of Science Museum of Minnesota's *Atmospheric Explorations* Computer Interactives', *Current Trends in Audience Research and Evaluation* 11, 25-35.

House, E. (1993) *Professional Evaluation: Social Impact and Political Consequences*, Newbury Park, NJ: Sage Publications.

Jacob, E. (1992) 'Culture, Context, and Cognition', in Margaret D. Le Compte, Wendy L Millroy and Judith Preissle (eds) *The Handbook of Qualitative Research in Education*, 293-335, San Diego: Academic Press.

Korn, R. and Sowd, L. (1990) *Visitor Surveys: A User's Manual Resource Report*, Washington, DC: American Association of Museums.

Krug, K. (2001) 'Our Colleagues, Our Selves: Modeling Museum Worldviews in the Process of Change', *Curator* 44 (3) 261-73.

Lawrence, G. (1991) 'Rats, Street Gang and Culture: Evaluation in Museums', in Gaynor Kavanagh (ed) *Museum Languages: Objects and Texts*. Leicester: Leicester University Press.

Lincoln, Y. and Guba, E. (1986) 'But is It Rigorous? Trustworthiness and Authenticity in Naturalistic Evaluation', in David D. Williams (ed) *Naturalistic Evaluation*, 73-84, San Francisco: Jossey-Bass.

Lindauer, M. (2005) 'From Salad Bars to Vivid Stories: Four Game Plans for Developing "Educationally Successful" Exhibitions', *Museum Management and Curatorship* 20, 41-55.

Martella, R.C., Nelson, R. and Marchand-Martella, N. E. (1999) *Research Methods: Learning to Become a Critical Research Consumer*, Boston: Allyn and Bacon.

Melton, A. W. (1935) *Problems of Installation in Museums of Art*, Washington, DC: American Association of Museums.

Melton, A. W. (1936) *Measuring Museum Based Learning: Experimental Studies of the Education of Children in a Museum of Science*, Washington, DC: American Association of Museums.

Miles, R. (1993) 'Grasping the Greased Pig: Evaluation of Educational Exhibits', in Sandra Bicknell and Graham Farmelo (eds) *Museum Visitor Studies in the 90s*, 24-33, London: Science Museum.

Munley, M.E. (1987) 'Intentions and Accomplishments: Principles of Museum Evaluation', in Jo Blatti (ed) *Past Meets Present: Essays About Historic Interpretation and Public Audiences*, Washington, DC: Smithsonian Institution Press.

Museum Evaluation Conference (1977) *An Abstract of the Proceedings of the Museum Evaluation Conference*, Washington, DC: Office of Museum Programs, Smithsonian Institution.

Pawson, R. and Tilley, N. (1997), *Realistic Evaluation*, London: Sage Publications.

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

Pekarik, A., Doering Z. D. and Karns, D. (1999) 'Exploring Satisfying Experiences in Museums', *Curator* 42 (2) 152-173.

Phillips, D. C. and Burbules, N. C. (2000) *Postpositivism and Educational Research*, Lanham, MD: Rowan and Littlefield Publishers.

Pitman, M. A. and Maxwell, J.A. (1992) 'Qualitative Approaches to Evaluation: Models and Methods', in Margaret D. Le Compte, Wendy L Millroy and Judith Preissle (eds) *The Handbook of Qualitative Research in Education*, 729-771, San Diego: Academic Press.

Porter, M. (1938) 'Behavior of the Average Visitor in the Peabody Museum of Natural History, Yale University', *Publications of the American Association of Museum New Series* 16, Washington, DC: American Association of Museums.

Robinson, E. S. (1928) *The Behavior of the Museum Visitor*, Washington, DC: American Association of Museums.

Schwandt, T. (1997) *Qualitative Inquiry: A Dictionary of Terms*, Thousand Oaks, CA: Sage Publications.

Schwandt, T. (2002) *Evaluation Practice Reconsidered*, New York: Peter Lang.

Screven, C. G. (1974) *The Measurement and Facilitation of Learning in the Museum Environment: An Experimental Analysis*, Washington, DC: Smithsonian Institution Press.

Serrell, B. (1998) *Paying Attention: Visitors and Museum Exhibitions*, Washington, DC: American Association of Museums.

St. John, M. and Perry, D. (1993) 'A Framework for Evaluation and Research: Science, Infrastructure and Relationships', in Sandra Bicknell and Graham Farmelo (eds) *Museum Visitor Studies in the 90s*, 59-66, London: Science Museum.

Stylianides, I. (2003) 'Significant Moments, Autobiography, and Personal Encounter with Art', in Maria Xanthoudaki, Les Tickle and Veronica Sekules (eds) *Researching Visual Arts Education in Museums and Galleries: An International Reader*, 153-165, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Vallance, E. (1996) 'Issues in Evaluating Museum Education Programs', in Douglas Boughton, Elliot Eisner and Johan Ligtvoet (eds) *Evaluating and Assessing the Visual Arts in Education: International Perspectives*, 222-36, New York: Teachers College.

Wolf, R. and Tymitz, B. (1979) *New Perspectives on Evaluating Museum Environments,* Washington, DC: Smithsonian Office of Museum Programs.

Zeller, T. (1996) 'From National Service to Social Protest: American Museums in the 1940s, 50s, 60s, and 70s', *Museum News*, 75 (2), 38-47.

* **Margaret Lindauer** is Associate Professor and Museum Studies Coordinator in the Department of Art History at Virginia Commonwealth University. She holds an interdisciplinary PhD in Curriculum Studies from Arizona State University, where she was the Curator of Exhibitions at the Museum of Anthropology. Her current research, for which she was awarded a 2004 Fellowship from the Smithsonian Center for Education and Museum Studies, focuses on the social, cultural and curatorial implications of the way in which museums historically have been characterized as educational institutions.