## Issues and Methods of Summative Evaluation

The first article that follows is a summary of Hayward and Loomis' (1994) paper entitled, "Looking Back at Summative Evaluation." The next two papers by Serrell and Shettel provide an argument/counter-argument on the "51% Solution" approach. Beverly Serrell's article on the "51% Solution" is essentially a proposal for a type of standardized summative evaluation that sets criteria for determining the success of an exhibition. Harris Shettel provides a critique of the 51% Solution describing some of his concerns in detail.

Both Beverly and Harris hope that these papers will generate some healthy discussion of our evaluation methods. Readers who are interested in contributing to this dialogue are invited to send their comments to *Visitor Behavior.*

# Summary of "Looking Back at Summative Evaluation"

Jeff Hayward & Ross Loomis
From the 1994 Visitor Studies Conference
Raleigh, NC

Hayward and Loomis led a discussion at the conference that served to provide a meta-analysis of summative evaluation. The following is a summary of the paper that summarized this discussion. Four issues were discussed:

1. *Why is summative evaluation taking a back seat?* A number of disincentives of summative evaluation were listed:

- attitude that the project is over
- no resources left
- results won't have an impact
- fear of bad news
- threat of criticism
- attitude that it does not have practical application

To counter these points, seven positive benefits of summative evaluation were discussed:

- it offers a chance to learn about visitors
- it's important in assessing the project's goals
- it helps funding agencies demonstrate the educational value of exhibitions
- it can be an important part of the planning cycle
- it is a useful first step in visitor research
- it is the best opportunity to demonstrate learning
- it provides a counterpoint to staff impressions

2. *The debate between summative and remedial evaluation.* The question of whether remedial evaluation is distinct from summative evaluation was discussed, but no resolution was offered. Remedial evaluation was described as: a study that attempts to fix or fine-tune an exhibition; a less extensive study than summative; and may occur between final construction and a summative study.

Several questions were raised: "Is remedial just a small-scale summative?" "Should remedial be called 'final formative'?" "If remedial is not as extensive or systematic as summative, is the quality of it questionable?" " Is there a danger of remedial becoming a trade-off for summative?"

Some professionals feel that remedial deals with fixing an exhibition, while summative simply tells whether or not an exhibition is effective. Others believe that it is two sides of the same coin.

3. *Generally accepted strategy for summative evaluation.* The authors suggest that the research methods used depend on factors such as exhibition size, content, objectives, and the audience. Exit interviews, which seem to be the most common strategy for summative evaluation, may have some limitations (e.g., some visitors may need time to consolidate and reflect on what they've seen).

Other generally accepted procedures include: the use of large sample sizes, multiple methods (observations, pre- and post-visit interviews), and getting detailed feedback about visitor perceptions.

4. *Generalizability and usefulness of findings.* Although summative studies do not have to be generalizable to be useful, there is often value to others to share the findings.

# The 51% Solution Research Project: A Meta-Analysis of Visitor Time/Use in Museum Exhibitions

Beverly Serrell
Serrell & Associates
Chicago, IL

The "51% Solution" is a methodology that combines a systematic, summative evaluation strategy with criteria for assessing and comparing the effectiveness of a broad range of educational exhibitions. Within the context of this study, each of these items has particular meaning:

- "51%" represents a simple majority.
- "Solution" is a metaphor for mixture, as in "dilution," rather than "the one-and-only answer."
- "Systematic" means using the same definitions and techniques in consistent ways in a variety of museum settings so that the data will be comparable.
- "Summative evaluation" means evaluating the whole exhibition (all its parts in context) after the exhibition is open to the public.
- "Strategy" consists of a combination of two techniques — unobtrusive observations of visitor behavior and exit interview/questionnaire with open-ended

questions – and a variety of ways of interpreting the data (e.g., statistical analysis, content analysis, qualitative review).

- "Criteria" are guidelines or benchmarks that set desirable, achievable levels of performance (the three criteria are listed on pages 7 and 8).
- "Assessing" means gathering data and comparing it to the criteria.
- "Comparing" means that the data gathered can be shared across exhibition types, sizes and disciplines.
- "Effectiveness" or success is defined by the degree to which the exhibition achieves its stated objectives with its intended audience.
- "Broad range" means a diversity of museum sizes, disciplines and budgets.
- "Educational" implies that the exhibition has stated specific learning goals, where "learning" is defined very broadly, e.g., "Visitors will find out about or realize something new about X"; "Visitors will make a personal connection with Y"; "Visitors will be inspired to wonder 'what if . . . ,' about Z, etc." as a result of experiencing this exhibition.
- "Exhibition" means a defined room or space (with known square footage), with a given title, containing elements that together make up a conceptually coherent entity recognizable as an exhibition of objects, interactives and/or phenomena. Some form of interpretation is present (e.g., text labels, graphics, videos, interactives) beyond mere identification of objects/art, etc.

## What the 51% Solution Is Good For

The 51% Solution provides tools (methods and criteria) for a goal-related investigation to answer the question: "How well is this exhibition working?" It includes ways to answer the following specific questions:

- How much time do visitors spend in this exhibition? (duration)
- What percent of the visitors pay at least some attention to the different parts of the exhibition? How many of the elements or stations do they stop at? (time allocation, utilization)
- What sorts of experiences do visitors have in this exhibition that they find meaningful and memorable? Do they get the main ideas? Can they remember any specifics? Did they make personal connections? (impact)
- How does the visitor use of this exhibition compare with use of other exhibitions? What is the impact on visitors relative to others? (comparisons)

The performance of the exhibition is measured by looking at a variety of visitor behaviors, involving time (duration and allocation), observable overt actions, and self-reported impacts and outcomes. This feedback can then be compared with the exhibition's stated communication objectives. In addition, the exhibition's time/behavior potential (size of exhibit, number of elements, modalities of elements, type and location of the host institution, etc.) can be compared with the data from other exhibitions, where similar evaluation methods have been used (see Note 1). The 51% Solution provides a simple yet rigorous approach to defining, collecting, and analyzing data; but at the same time, it lets visitors act naturally and normally. This approach dictates the researcher's behavior, not the visitor's.

The 51% Solution is unique because it provides a methodology that can be used across disciplines, and it will allow us to gather a large database to share and compare. We will be looking for broad trends and patterns that will provide useful information for making decisions about exhibition development and evaluation. The goal that is aided by the 51% Solution is the goal of improving the effectiveness of exhibitions.

## What the 51% Solution Is Not

It is not the ultimate or only way to look at what visitors get out of an exhibition. It is not based on a pre—post-knowledge gain model. It does not measure the long-term learning, but it does assess the potential and prerequisites for it. The 51% Solution is not focused on understanding how different kinds of people learn from exhibitions (i.e., looking at differences between special audiences, e.g., gender, social group, educational characteristics or variables), but it does look at some exhibition variables that may contribute to visitor learning (e.g., size, density, modalities present). It is not a methodology meant for doing formative evaluation nor for evaluating single exhibit elements—it is for whole exhibitions. It is not anecdotal, and it does not seek to predict what any one person will get out of any one exhibition.

## What Methods Does the 51% Solution Use?

The 51% Solution uses two methodologies: (1) unobtrusive observations of visitor behavior, and (2) feedback from individual visitors in an open-ended exit interview/questionnaire.

Unobtrusive observations of visitor behavior—through tracking and timing of visitors—provide information about how long visitors spend in the whole exhibition; what percent of the elements were utilized by visitors; the relative "popularity" of all exhibit elements (attracting potential—which ones were visited most, least); if visitors were using the interactives appropriately and completely; if they were reading the texts and following directions; if they were using exhibit elements repeatedly; and what percent of the visitors went through the whole exhibition vs. only part (e.g., exited at the first opportunity). Although for the purposes of this study it is not necessary to record specific behaviors of visitors, you will be able to see if

visitors used interactives appropriately and completely, if they were reading the text and following directions, and to what degree they engaged in social interactions (e.g., talking, pointing) as they went through the displays.

A randomly selected, representative sample of visitors (n=40 or more) is tracked and timed on weekends and weekdays. The data collector notes visitors' gender, approximate age, and social group makeup, where they go in the exhibition, and how long they stay. The demographic information helps determine if the sample was representative of the museum's "normal" visitor profile. Pathways through the exhibit and stops at elements are noted on the map. Data are tallied on a spreadsheet, listing each individual, total time, total stops and other behaviors, if systematically recorded. These unobtrusive, objective measures – which focus on time and attention—are valuable indicators of the exhibit's attractiveness and visitor interest, and have been found to correlate positively with learning and enjoyment (see Note 2).

For the initial research project, "A Search for Generalizability: Visitor Time/Use in Museum Exhibitions," we are collecting time and stops data from as broad a range of exhibition types and sizes as possible.

The second data-collecting method employed for this type of summative evaluation is an exit interview/questionnaire with cued visitors. Visitors are recruited as they enter the exhibition. Participants who agree to fill out the form afterward are given a slip of paper with a code number and their starting time, and are told to look for the evaluator at the end of the exhibition. Participants receive a gift for filling out the form. Some people decline, citing lack of time, presence of young children, or inability to speak English well (see Note 3).

Feedback from cued visitors who answer five open-ended questions (see note 4) in their own handwriting provides information about what visitors remember and find meaningful in the exhibition, and to what degree they understand the educational concepts and communication goals of the exhibition. Visitors fill out the questionnaires while seated at a table outside the exhibit area, taking as much time as they choose. They cannot look at or read the exhibits, however, while they answer the questions. (Correct spelling and punctuation are not important, and we have found that a lack of articulate writing skills is not a factor in communicating if they learned something new, made a personal connection, or had some existing feelings/knowledge reinforced.)

The advantage of letting people answer open-ended questions in writing is that the evaluator does not put words into their mouths or put pressure on them for a quick answer, or create categories that limit or direct visitors' responses. On the other hand, this form of data is more difficult to score and summarize, which is why the evaluation report should include several different ways of looking at the questionnaire feedback, as well as copies of all the original forms or transcriptions for interested staff/reviewers to examine for themselves.

A random sample of visitors (n=30 to 50, not the same people who were tracked) is asked to fill out the questionnaire after a very brief interview about their visitation and prior interest (see Note 5). Their responses are summarized by content analysis – looking at the words visitors used and how those words relate to the individual exhibit elements, the ideas communicated, and generally how visitors related to the stated goals of the exhibition. Even if a visitor's comment is extremely terse (e.g., one word), that word, if specific enough, can be matched with the exhibit or area's goal it is closest to. Individual questionnaires can also be rated or sorted according to how appropriate the person's reactions are as compared to the exhibit developer's hopes or intentions (e.g., "didn't get it," "so-so," and "OK").

In summative evaluation studies, cued testing is not normally recommended because cuing increases people's level of motivation and attention. It can be argued, however, that cuing is useful for museum settings. Visitors' recall levels are likely to be very low, due in part to the typically brief, incomplete, and informal visits people make to exhibitions, and, in some cases, confusing or unclear exhibits (see Note 6). In addition, visitors to the exhibition are under no obligation to learn anything.

Cuing provides a "best-case scenario." Thus, if *cued* visitors fail to notice, understand, or remember parts of an exhibition, one can assume that it it very likely that uncued visitors are not paying much attention to them either. The unusual and difficult-to-grasp concept for the questionnaire part of the evaluation is that missing data (no response, no recall, no meaningfulness) and/or a lack of patterns in the data where one might expect to find them, provide insightful information.

On the other hand, if cued visitors to an exhibition do have a high rate of recall, that level of response cannot be assumed to be typical for a population of "normal" visitors to the exhibition – people who are less motivated and spend less time. Cued visitors provide empirical evidence that visitors *can* learn from the exhibit, but do not prove how many actually do. Among exhibitions, however, the rate of recall from cued visitors can be compared systematically to the response rate of cued visitors in other exhibition studies.

## What Are the Criteria of the 51% Solution?

The 51% Solution has three criteria or guidelines for measuring and comparing the effectiveness of an exhibition, in the form of three questions:

- Do 51% of the visitors move through the exhibition at a rate of less than 300 square feet per minute (size of exhibit divided by total average time)?
- Do 51% of the visitors stop at 51% or more of the exhibit

elements?
- Can 51% of a random sample of cued visitors, immediately after viewing the exhibition, express general and specific attitudes or concepts that are related to the exhibition's objectives?

Fifty-one percent was empirically derived from evaluation data (not whimsy or intuition) of a critically acclaimed "model" exhibition (see Note 7). Also, 51% represents a simple democratic majority as well as a realistic standard that exhibitions can strive for, considering the diversity of visitors' demographic and psychographic characteristics. Fifty-one percent provides a reference point, e.g., is the data below 51%, or does it exceed 51%? It is not an end point.

The first two criteria are measured unobtrusively by tracking and timing. The third is measured by one-on-one exit interview/questionnaires. Complete definitions (not covered earlier) for the terms used above for the criteria are:

- "Visitors" are people who enter the exhibition and appear to be in it because they are using it (not lost, not using it as a hallway). Only adults (ages 16+) are the subjects; only individuals are tracked (regardless of the number of other people they are with); similarly, only casual visitors are observed, not people in tour groups. For each study, the recommended sample size is 40 to 75 (100 or more is overkill).
- "Randomly selected" means adult casual visitors selected by a specified mechanism (e.g., every Xth visitor) without bias for gender, age, race, or social group. Subjects are selected over representative periods of days of the week and time of day.
- "Exhibition objectives" are the educational objectives (see "educational" defined earlier) that the exhibition's developers have clearly identified. Exhibitions without learning goals or any interpretive components will not be included in this study.
- "Exhibit elements" are defined as discrete, conceptual units, experiences, or components within the exhibition layout. They may vary widely in size and type, e. g., a panel, a case, a diorama, a set of artifacts, a video theater, a computer, an interactive device, etc. They should be defined by the in-house staff who are familiar with the exhibition (see Note 8).
- "Stop" equals both feet of the visitor coming to a full halt for 2-3 seconds while the person's body and/or head is oriented toward the exhibit element (see Note 9). Stops are used to derive the number of exhibit elements "used" by each visitor (a very simple and admittedly generous interpretation) (see Note 10). Multiple stops at a single element are counted as only one stop. A stop at every element would mean that total stops and number of elements are the same.
- "Total time" is the time elapsed as the visitor entered the exhibition, looked around, made stops, and left. (Time at individual elements need not be recorded (see Note 11) unless the element has a clearly definable amount of time to be used completely, such as a 3-minute video.)
- "Average time" is the sum of the total times for all visitors in the sample, divided by the number of visitors in the sample. "Outliers" –unusually high times (e.g., twice or three times the average) for one or two visitors in the sample — should be dropped because they skew the average so that it does not realistically represent the sample. They should be reported, but not figured into the statistical analysis.
- "Square feet per minute" is a figure derived by dividing the total square footage of the exhibition by the average time, which allows exhibitions of different sizes to be compared against one another (see Note 12). As visitors stroll through the exhibit, they look around, visually sweeping the area, stop occasionally, and sometimes stop long enough to look closely, watch, read, and/or interact with an exhibit element, or interact socially (e.g., talk, point, read out loud). Also called sweep rate.

The 51% Solution postulates that good exhibitions are well-utilized ones: a low sweep rate and a high-percent utilization are considered positive behaviors for visitors in most exhibitions ("They stayed a long time and looked at almost everything!")

For the purposes of The 51% Solution Research Project, museum practitioners who are participating will conduct a study at their own institution using the same definitions and methods described in this paper and will send Serrell & Associates the following tracking and timing data:

1. Name of institution; name of exhibition where tracking and timing was done;
2. Square footage of the exhibition;
3. Number of elements in the exhibition;
4. Data in the form of a list with the total time, and total number of stops for each visitor observed;
5. A sample of the data sheet used (showing floor plan of exhibition).

Only tracking and timing data are necessary for this part of the project. Even though it does not provide as complete a picture of visitor use/impact, it is a place to start, to gather a substantial data base from which to look for generalizable trends and patterns.

If you have any questions about how to get started, or would like to bounce some ideas off me as you go along, or want to make sure that your format and definitions will be compatible with the rest of the participants, please feel free to give Beverly Serrell a call at 312-643-5922 in Chicago (or fax at 312-643-8460).

## Notes

1. A number of commonly used visitor survey methods describe demographic and/or psychographic characteristics of museum audiences (e.g., ages, reasons for visiting), but no widely used methods or criteria exist that allow for the comparison of whole-exhibit evaluation data (e.g., visitor behavior or exhibition impact) across exhibition types, between institutions, or among evaluators. "Attracting power" and "holding power" are usually used to measure visitor response to single exhibit elements.

2. Time sets the precedent for and is indicative of many other desirable outcomes. In order for long-term learning to occur, there must be short-term learning; in order to have short-term learning, there must be attention, and attention takes time. Time alone is not the sole measure of a visitor's interest, enjoyment, or learning, but time not spent at a particular exhibit element can indicate that no interest was shown and learning was highly unlikely to have occurred then, nor will it occur in the future. While we do not have the tools to measure everything that happens to a visitor in an exhibition, we do know that attention is necessary and time is an important indicator of it.

3. For an explanation of the data sheet, see Raphling and Serrell's "Capturing Affective Learning" in *Current Trends in Audience Research and Evaluation*, Vol. 7, 1993, published by American Association of Museum Committee on Audience Research and Evaluation. (See #5 below).

4. Answers to open-ended questions produce different kinds of data from answers to closed questions, and the two often are not comparable. For example, the open-ended question, "What would you say this exhibit was about?" will produce qualitatively different information than the closed question, "Were you aware that the exhibition was about X?" The percent of visitors reported to have understood the main point of the exhibition under the two different situaitons is likely to be quantitatively very different.

5. The interview questions are: "Is this your first visit?" "Do you have any special interest, knowledge or training in (the subject)?" The questionnaire questions are: "What is the main idea of this exhibit?" (Prompts: "to show...", "to make people ...") "What is one new idea you are taking away with you?" (Prompts: "I never knew...", "I never realized that ..." and/or, "It reminded me ..." ) and "Anything else?"

6. Probably the most common reasons that many exhibit evaluation studies have failed to document or demonstrate postive learning impacts on visitors are that the exhibition was unclear (i.e., the learning objectives were not obvious), the amount of information or experiences were overwhelming, the space was uncomfortable (too crowded, noisy, hot), and/or not attractive enough to command visitors' attention long enough to inspire a meaningful intellectual or emotional connection.

7. The exhibition was "Darkened Waters: Profile of an Oil Spill." It was evaluated in 1992 by Serrell & Associates for the Pratt Museum while the exhibit was traveling to the Oakland Museum in California. It received positive reviews in *Museum News*, March/April, 1992.

8. For the sake of reliability, at least two different staff members should agree on the definitions and number of elements in the exhibition.

9. This is a very short time, but it is operationally definable and indicative of possible interest and learning – measured in seconds, not minutes. If you think 51% is too low a criterion, raising the number of seconds to be counted as a "stop" to 5 or 10 would lower the number of visitors who "use" exhibits even more. This strategy assumes and accepts that visitors will use many exhibits quickly and incompletely.

10. To be included in the final data analysis for each exhibition studied, visitors must make at least one stop. Visitors who do not make any stops in the hall or even pause to glance at some exhibits (that is, who appear to be using the hall as a passageway or are lost) are not included in the data analysis. Therefore, "non-user" data does not influence (i.e., lower) the average time/stops data. There are exceptions to this "rule": (1) if it is not uncommon for visitors to walk very slowly down the center of the hall while looking at the cases, clearly paying attention to the exhibits, but not coming to a full stop, or (2) if it is not uncommon for visitors to walk very slowly around all sides of an exhibit element, clearly paying attention to it, but without coming to a full stop. Those two exceptions could be called a "stop."

11. Recording time at individual exhibit elements is not recommended for the following reasons: (1) it makes the overall job of tracking and timing much more complicated, and therefore, prone to error; (2) it tempts evaluators to place value judgments on visitors' behavior at individual elements (e.g., relative level of engagement or interest shown), which tend to be more subjective than overall time, and (3) it collects more data than is necessary for the purposes of the study.

12. This has been the most problematic and controversial measure of the 51% Solution. At first we called it "speed," but ran into logical problems with its linear nature. We are working on this definition and struggling with the issue of trying to compare exhibitions containing elements that have vastly different-sized "footprints," such as paintings on a wall and cases of small objects vs. large dioramas or re-creations of environments (e.g., period room, rain forest). Although visitors cannot walk into large dioramas or fish tanks, they search those areas with their eyes. Thus, we currently use a notion called "sweep rate" (borrowed from mathematic's random search theory) rather than "speed" to describe visitors' movements. With a large database we will be able to see if any patterns hold true across exhibitions of different "footprints" or element types.