SUMMATIVE EVALUATION

Introduction and Issues in Summative Evaluation

Stephen Bitgood
Jacksonville State University

As stated by Hayward and Loomis (1994), summative evaluation is "a study of visitors' use and/or perception of a completed exhibit, generally conducted within the first year that the exhibit opens to the public, ... [it is] considered to be the end of a process." Summative evaluation is perhaps the longest-used type of evaluation. Before reading the articles in this section, the reader might wish to consider what comprises an effective summative evaluation.

Criteria for Summative Evaluation Measurement Systems

In order to evaluate an evaluation method, it is necessary to ask what are the characteristics of an ideal evaluation system. The following list of criteria are suggested for assessing the adequacy of both a single evaluation study and a general evaluation approach such as the one suggested by Serrell. Of course, the criteria will differ in terms of their importance. Nevertheless, each should be seriously considered.

1. *Inclusiveness of measures*: Does the evaluation include multiple measures (behavior, knowledge acquisition, and affective impact)? The inclusiveness criteria is important since there seems to be general agreement that a summative evaluation should include more than one type of measure.

2. *Reliability of measures*: Do independent observers agree on how to score a response? Qualitative outcomes (e.g., responses to open-ended questions such as "What did you learn in the exhibition?") can be particularly troublesome in obtaining agreement between independent observers.

Unwanted drifts in the data collection process present another problem that must be closely monitored. Interviewers, in their eagerness to produce an exciting finding, may subtlely change their criteria for categorizing a response.

3. *Construct validity*: Does the measurement really measure what you think it measures? In Serrell's paper below, she suggests a measure of exhibition attention she calls [visual] "sweeping" as measured by the total amount of time a visitor spends in an exhibition as a function of total square feet. (This measure was formally called "speed.") Whatever this measure is called, does it really measure the kinds of attention to exhibits that it is supposed to measure?

4. *Recording accuracy:* Does the measurement distort the actual occurrence of the behavior? Some measures tend to overpredict or underpredict behavior. For example, when asked, "How long did you spend in that exhibition?", visitors tend to overpredict time.

5. *Internal validity*: If the outcome measure (e.g., attitude toward the subject matter of the exhibition) is satisfactory, is this finding really a result of the exhibition experience or would nonvisitors show the same outcome? As Hayward and Loomis (1994) suggest, it is common practice to use a pre-visit/post-visit interview comparison in order to minimize the possibility that visitors came into the exhibition with the desired knowledge or feelings about the subject matter. If only exit interviews are used, the outcomes cannot always be attributed to the exhibition experience.

6. *Generalizability*: Can the results be generalized to other exhibitions and settings? Although generalizability is not among the most important criteria for a summative evaluation, it is certainly desirable.

7. *Prescriptive utility*: Does the evaluation contribute to general principles for future exhibition design? What is learned from an evaluation of one completed exhibition is often incorporated into the design of the next exhibition. For example, if the evaluation reveals that visitors do not read labels that are distant from the objects they describe, then future design might ensure that objects and labels are placed close together.

8. *Diagnostic utility*: Does the evaluation provide a diagnostic summary of exhibit strengths and weaknesses? In many cases, future changes are contemplated. A summative evaluation can reveal weaknesses that can be easily corrected.

9. *Resource/benefit ratio*: Did it produce the greatest "bang for the buck?" Ideally, an evaluation will provide as much information as possible at limited cost of resources.

10. *Sampling validity*: Were sampling procedures appropriate for the study? Did the sample represent the desired audience? Were sample sizes adequate? Was there bias in the selection of participants? Was there a low refusal rate?

11. *Acceptability*: How acceptable is the evaluation by peers and consumers? It is possible that peers accept a method, but consumers do not, or vice versa.

12. *Sensitivity of measure*: Some outcome measures are less sensitive than others. For example, free recall of exhibit content is less sensitive to knowledge acquisition than a recognition task where the respondent need only identify exhibit content instead of recall it.