Looking Back at Formative Evaluation

Jeff Hayward
People, Places and Design Research
Northampton, Massachusetts

Ross J. Loomis
Colorado State University
Fort Collins, Colorado

For many, formative evaluation is a *must do* kind of evaluation. Exhibits and programs can be tested with an iterative process during design and construction stages. Results of such testing can be applied immediately and aid designers and program planners in finding the best alternatives for ensuring exhibit effectiveness. Typically, small convenience samples are used in concert with mock-ups of design or program alternatives.

The basic idea of formative evaluation is not new, and is certainly not limited to education or exhibits and programs. Engineering and construction work makes use of mock-up testing, and Petroski (1992) notes that formative types of studies have a broad and rather long history. For example, he observes that skepticism surrounded the building of the London Crystal Palace in the 1850s. Critics felt the building would not survive the tramping of crowds. To quiet the critics, a formative test was done with a sample of workers tramping across a section of gallery floor set up for the test. Apparently, the sample of workers could not trample in proper cadence to provide a good test, so a detachment of soldiers was employed and a more orderly demonstration resulted. This bit of formative evaluation was reported to have been observed by none less than Queen Victoria herself. Scientists present concluded that the vibration produced by the marching soldiers did not exceed that found during evening parties at common London houses.

This conference session was planned as a follow-up to last year's *Looking Back at Front-End Studies*, to review significant issues in the practice of exhibit evaluation. The co-chairs defined four issues for discussion concerning formative evaluation that included:

- deciding what should be evaluated;
- using small, but representative samples;
- defining "effectiveness" of mock-ups; and
- recommendations and follow-through.

The Co-chairs summarized a number of points for each issue, and then a discussion followed with members of the audience sharing ideas and suggestions. By this process an informal "meta-analysis" of formative evaluation was accomplished during the session.

What Should Be Evaluated?

Most projects consist of many individual components which could benefit from formative evaluation. Which components should be studied, and how extensive should the evaluation be? Often, the focus of what gets evaluated is decided by staff conflicts. The evaluator is apt to be in the middle with expectations that evaluation will show who was right. Unfortunately, deciding staff arguments is not the best reason for doing evaluation. Evaluators have the responsibility to talk early and often, and to encourage staff to set overall goals and objectives for evaluation. It is important to clarify the evaluator's role and to explain that results are apt to show strengths and weaknesses for both sides of any disagreements. Evaluation must be objective and may include topics and problems that none of the design team thought about. Evaluators must have the independence to report when a design is not working. It is generally accepted that formative evaluation can help with practical problems, such as how an interactive element is used, but formative evaluation can also help explore conceptual issues and the likely effectiveness of the basic message or theme of an element.

Audience members emphasized that an important step in keeping formative evaluation on track is to clarify objectives and set them down in writing. It was also suggested that evaluations should be planned with quick turnaround times to provide information that can be applied as decisions are being made.

Using Small and Representative Samples

Perhaps one of the most controversial aspects of formative evaluation is the use of small samples. Conventional wisdom is to use large, scientifically composed samples, but the practical demands of formative evaluation preclude the time needed for large samples. Some administrators and decision-makers are apt to be skeptical about results based on small samples, so evaluators need to keep this skepticism in mind when reporting on formative studies.

A number of things can be done to strengthen the validity of small sample outcomes. For one, the sample can be carefully described to make clear just what people are represented. Graphic summaries, such as are provided in a histogram, can be used to make it clear who was included. Samples can be constructed to represent a cross section of visitors, or focused to assess specific groups such as families, first time visitors, repeat

visitors, museum members, or any other defined population. Small sample statistics, such as can be found in programs that compute exact probabilities of outcomes (for example, StatXact), can increase the power of tests with limited samples. Calculating power of a test using generous effect size and conservative alpha levels (Kraemer & Thiemann, 1987), and employing nonparametric statistics are other aids in this category. These aids are found on most statistics software programs. As a rule-of-thumb, a conservative 70/30 or even 80/20 criterion could be used. That is, 70% or 80% of a sample must show a specific outcome (i.e. understand a term, look at an orientation panel, correctly activate a display) to support using that design alternative. Finally, the co-authors recommended using a minimum of *50* subjects per sample to allow for good power for testing *conceptual* understanding.

Members of the audience pointed out that a sample of 50 may not be needed for diagnosing mechanical problems or more obvious changes. Also mentioned was the use of predetermined quota samples for testing specific audience groupings. Keeping outcomes simple, such as in a yes-no format, can make it easier to compare results across sub-groups of the sample.

Defining "Effectiveness" of Mock-Ups

Mock-ups used in formative testing range from paper and pencil tests to highly realistic simulations. Some designers are uncomfortable with mock-ups because they lack realism. It is important to emphasize that mock-ups can be effective without absolute realism if they contain critical cues or features of that which is being tested. Helpful ideas for making mock-up testing effective include using appropriate visual aids if the element is not an interactive one, and testing design alternatives. Alternatives should be planned as part of the overall evaluation strategy. In addition, while testing mock-ups for attractiveness can be problematic, mock-ups can work well for testing the conceptual content of interpretation. Finally, it is important to set or define the context for the mock-up and be realistic about what can be tested. For example, can the mock-up be built full size, rather than using a model or reduced scale presentation?

Audience members commented that involving exhibit designers in the mock-up testing is very important. Designers, including those with backgrounds in industrial design, can have valuable input into the often highly visual nature of mock-up materials. Fitting design criteria to communication objectives was emphasized by audience members.

Recommendations and Follow-Through

Difficulties can arise when the testing is completed and it is apparent that some things are not working as planned. A natural inclination is to look for someone to blame. Old debates about who is right may resurface.

Timing of evaluation information may be a problem, also; results often have a way of trickling in a bit at a time. Partial results can give rise to rumors or premature changes without the benefit of more testing, but there may be resistance to further testing.

There are a number of things an evaluator can do at this stage of a formative evaluation:

- Data summaries should be accompanied by interpretation of results, to minimize speculation and rumors;
- The evaluator can be careful to *interpret*, not decide, results of studies;
- Criteria and alternatives can be suggested that would help resolve problems. Here it is important for the evaluator to focus attention on producing the best program or exhibit, rather than on proving who was right;
- Evaluators have a responsibility to not be shy about recommending further testing. Furthermore, such additional tests should not be done for free! It is not the evaluator's fault that more testing is needed. Rather, it is the strength of the interactive nature of formative testing that additional studies can improve the final product. In addition, these studies may more than pay their way in helping to avoid costly post-installation changes.

The audience made several valuable suggestions, including involving staff in the evaluation process to educate them about the process and prepare them for outcomes. Some discussion revealed the problem of people not understanding raw data, and the need for interpretation of results. Timing of information was also mentioned as being very important. Administrators and project staff need to understand up front the time needed to do evaluation, and that schedules should include adequate time for formative studies.

Ideas brought up in this informal meta-analysis can help strengthen formative evaluation. There seems to be general agreement that this kind of evaluation will continue to be an important tool in visitor studies.

References

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects?: Statistical power analysis in research.* Beverly Hills, CA: Sage Publications.

Petroski, H. (1992). Making sure. *American Scientist, 80,* March-April, 121-124.

StatXact. Cambridge, MA: Cytel Software Corporation.