Websites: A Guiding Framework for Focusing Website Evaluations

Carey E. Tisdal

Tisdal Consulting

Abstract

The aim of this study was to explore 22 Web site evaluation reports, or sections of larger evaluation reports centering on a Web site, to identify, define, and provide examples of the range of evaluation focus areas to inform the design of Web site evaluation studies. The sample included a group of reports contributed to the Informalscience.org online database. Prior to this study, staff members at the Science Museum of Minnesota organized and coded the database of evaluation reports as part of the Building Informal Science Education (BISE) project funded by the National Science Foundation (NSF). In this analysis, grounded theory methodology and the constant comparative method (Glaser & Strauss, 2009) were used to identify and define nine major evaluation focus areas that appeared in one or more of the 22 reports: Target Audience and User Characteristics, Awareness, Motivation, Access, Usability, Use, and User Impact and System Effectiveness. In addition, the analysis identified connections among these elements to present a guiding framework for website evaluation design. The guiding framework displays 7 major evaluation focus areas as sequential, necessary steps to accomplish User Impacts and System Effectiveness.

Keywords: evaluation design, websites

Websites: Evaluation Focus Areas

Along with other authors, I was invited to synthesize a group of evaluation reports from the Informalscience.org database. Staff members from the Science Museum of Minnesota had organized and coded these reports as part of the Building Informal Science Education (BISE) project funded by the National Science Foundation (NSF). Since I had recently completed two Web site evaluation reports as well as two other projects where websites were one of several media studied, I chose a group of reports in which evaluands had been identified as a specific Web site. After an initial review, I decided that mining the reports for any generalizable knowledge about websites was not something that appeared either practical or appropriate for this sample. Like other evaluation reports, the findings among this group focused, implicitly or explicitly, on the merit or worth of a *specific* set of products and actions within a *limited* time and place to inform decisions of one or more stakeholding groups. This is the defining feature of evaluation, and it is one way that evaluation is different from research or policy studies (Guba & Lincoln, 1989). For this reason, reviews of evaluation reports for generalizable knowledge about specific media or technology-based formats may not be always be productive.

Even so, for evaluators, reviewing reports by our colleagues is a useful exercise and the BISE database provided a valuable resource. One thing I could explore was how my colleagues focused their inquiry, that is, what questions they asked and what areas were they able to provide information to clients and other stakeholders (e.g. funders, program participants, and the community) about the value or worth of an evaluand. The aim of this study was to analyze 22 Web site evaluation reports, or sections of larger evaluation reports centering on a Web site, to identify, define, and provide examples of the range of evaluation focus areas such that they might inform the design of other Web site evaluation studies.

One reason I selected this aim was its usefulness to my own practice of evaluation and, potentially, to that of others. Defining the overarching questions an evaluation will address is a collaborative process and generally involves working with stakeholders from a wide range of professional groups and with a wide range of perspectives. Professional groups may include exhibition designers, program developers, film-makers, Web site designers, scientists, science-

content specialists. Other important stakeholders include funding agencies, audiences of films and television programs, Web site users, and museum visitors. Each group shares a language and world view shaped by their personal, educational, and professional experiences. This means that people tend to have different perspectives about what is valuable and important about an informal learning product or experience. As an evaluator, I have the responsibility to identify the needs and interests of such a range of stakeholders. My role is often to help clients ask questions they may not have previously considered. Through this study, therefore, I hoped to develop a guiding framework for myself, my clients, and other evaluators that would help us consider a wide range of questions to explore when evaluating websites.

As I began this study, I had two initial overarching questions; a third question, to be addressed later, emerged during analysis. The three questions are:

- Across all 22 reports, what was the range of evaluation focus areas?
- Were any evaluation focus areas used more frequently in different types of evaluation?
- Could these evaluation focus areas be organized into a series of sequential, necessary steps to accomplish user impacts, thereby providing a guiding framework for Website evaluation designs?

By "range of evaluation focus area," I mean the questions that the studies addressed and the areas in which the authors provided information and findings to the readers of the reports. By "evaluation types," I mean phases of evaluation that parallel a development process as generally used among the informal learning evaluation community: front-end, formative, remedial, and summative (Screven, 1990). I did not assume that all these evaluation focus areas would be appropriate for all phases of evaluation. In addition, each evaluation design must balance many factors, respond to many audiences, and be developed within the boundaries of a budget and timeline. My aim was to develop a useful heuristic, specifically suited to websites, to help identify the most useful evaluation questions.

Methodology and Method

I used grounded theory (Glaser & Strauss, 2009) as my overarching methodology. Unlike experimental design, which aims to test theory, a grounded theory researcher generates or discovers theory from the data. I used the constant comparative analysis method throughout my

analysis. I refined categories by identifying and examining similar categories in the literature on general evaluation practice, Web site evaluation, and marketing theory that appeared similar to the categories developed from the study data.

No one approaches the task of developing a grounded theory as a blank slate. In this instance, I am in the same community of practice as many of the authors whose reports I studied. Most of the authors come from the visitor studies field: I share experience in similar contexts, a specific literature, and a common language with these colleagues. One of my own reports is included in the sample for this study. Several other authors of the reports I analyzed came from the field of broadcast and media evaluation. I found their language and concepts familiar, too. Before transitioning to work as an evaluator in museums and science centers, I worked for 10 years as an instructional designer and media producer. Thus, while I did not begin data analysis with a pre-ordinate set of categories, I did bring to the study pre-existing concepts from my own experience.

During the process of identifying and developing categories across the reports, I realized that some of the categories could be connected as sequential, necessary steps that lead to outcomes. Exploring the relationship among categories is a step in developing a grounded theory. Upon reflection, I realized that my identification of and recognition of these relationships were based on experiences in applying Carol Weiss' Program Theory technique. As Weiss (1998) points out, "For evaluation purposes, it is useful to know not only what the program is expected to achieve but how it expects to achieve it" (p. 55).

> Much evaluation is done by investigating outcomes without much attention to the paths by which they were produced. But evaluation is increasingly being called upon not only to answer the question 'Did the program work?' but also 'What made it work? Why was it successful or unsuccessful?" (p. 55)

Weiss makes a clear distinction between formal theory (e.g. constructivism, behaviorism, relativity) that provides an overarching framework for a body of research or a field of study and program theory. Program theory is specific to each development project or entity.

> By theory, I don't mean anything highbrow or multi-syllabic. I mean the set of beliefs that underlie action. The theory doesn't have to be uniformly accepted. It doesn't have to

> be right. It is a set of hypotheses upon which people build their program plans. It is an explanation of the causal links that tie program inputs to expected program outputs, or as Bickman (1987) has put it, 'a plausible and sensible model of how the program is supposed to work." (p. 55)

Several years ago, on the recommendation of a colleague, I adopted Weiss' approach and found it a useful way to elicit assumptions from development teams and to visually display them for analysis. I also found it useful in identifying where to prioritize evaluation efforts.

I used constant comparative analysis to generate and verify categories. I developed categories across the 22 cases in the study. I began by reviewing the Findings and Conclusions sections of the reports and coding some categories into temporary groups. I found that my understanding of what had been the actual focus was not always clear. At this point, I revised my plan and reviewed each report in its entirety. As I continued, my next step was to add additional bits of data into temporary categories that apparently related to the same type of information (Lincoln & Guba, 1985). This meant that, when I discovered what appeared to be a new category in a report, I would return to all previous coded reports to see whether I had not identified data related to this content because I had not been explicitly looking for the category during previous rounds of coding. My next step was to "devise rules that describe category properties and that can, ultimately, be used to justify the inclusion of each data bit that remains assigned to the category as well as to provide a basis for later tests of replicability" (p. 347). This involved reviewing the temporary categories, writing explicit definitions, and then excluding any instances that failed to meet the definitions. In developing definitions, I looked for similar categories in relevant literature that connected directly to the practical areas involved. These sources are cited with the definitions in the Results section. As I continued, I looked for relationships among the categories, using the Program Theory technique to arrange them in a logical and sensible sequence ending in User Impacts.

Characteristics of the Sample of Reports

I reviewed all of the cases where the evaluands had been identified by coders as a Web site. This resulted in having a sample of 22 cases. When reports contained more than one evaluand, BISE coders split reports by evaluands type. Of the 22 cases in this sample, 10 were split reports, one was a combination of two reports, and 11 were full reports. (I refer to all these

cases as reports throughout the rest of this article.) Based on the BISE codes, the reports fell into only four categories: 6 were identified as Formative Evaluation studies, 1 as Remedial, 14 as Summative, and 1 as Don't Know.

The 22 reports were written within a seven-year timeframe: 2003-2011. Figure 1 shows the distribution of reports by year. I do not consider timeframe of the reports surprising since the Internet did not become widely available for public use until the 1990's, and Informalscience.org, the online site to which the reports were submitted opened for contributions during the early part of this timeframe

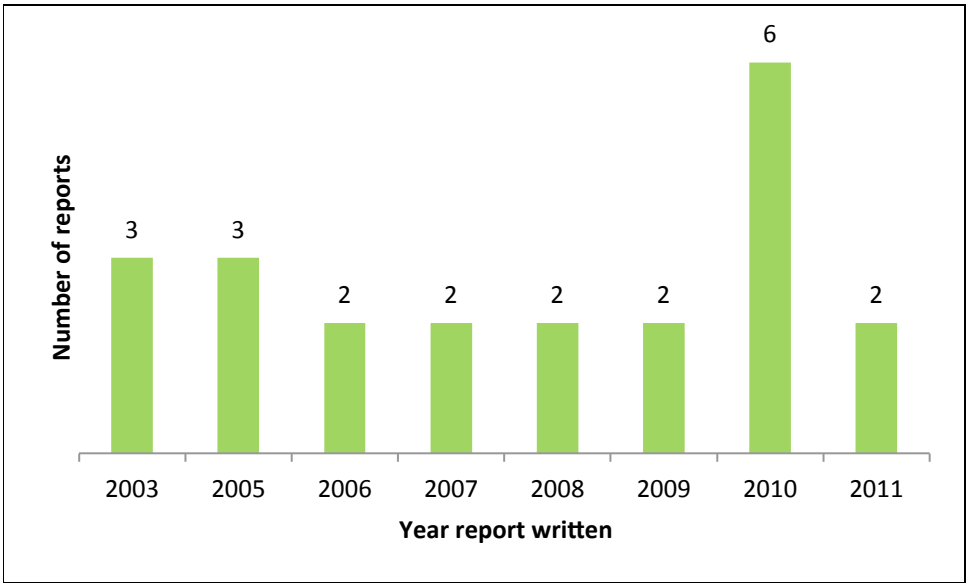

Figure 1. Number of reports by year written.

The small number of reports in this sample is surprising, and I find it highly unlikely that the sample is representative of all informal science learning websites evaluated during this timeframe. Other characteristics of the sample provide insight into which types of evaluation reports may be missing and which areas may be over-represented.

All 22 of the evaluation reports were part of grant-funded projects, including 20 (90.91%) projects funded by the National Science Foundation (NSF), one (4.55%) funded by the National Oceanic and Atmospheric Administration, and one (4.55%) funded by the National Institute of Health's Science Education Partner Awards. The sample does not include any studies conducted

with internal institutional funds, by other federal agencies such as the Institute for Museum and Library Services, or by or for foundations.

External evaluators wrote all reports with 50.00% contributed to the database from two evaluation organizations; one of these evaluation organizations wrote 7 reports and the other wrote 4. One organization wrote two reports and each of the other four contributed one report to the sample. The sample does not include any studies conducted by internal evaluators.

One commonality among the reports appears to be important, not just as to the nature of the sample but to the practice of evaluating websites: all but 2 of the reports focused on websites that were part of designed learning systems with other components. This points to a need to look at the effects of websites upon larger systems and the larger system's effects on the websites. I found that several reports with only one evaluand were associated with other system elements such as broadcast media (i.e., television series, television programs, and radio programs) and museum exhibits and exhibitions. In general, the sample had a comparatively high concentration of reports associated with broadcast media and fewer reports on websites associated with museum exhibitions and program, large format films, or professional development for museum staff and scientists. The types of associated system across all 22 reports, in order of frequency, included:

- 59.09% (n=13) broadcast media including televisions series, programs and radio
- 36.36% (n=8) print curriculum and activities for K-12 school setting
- 27.27% (n=6) social media including Facebook, Twitter, YouTube and Delicious
- 18.18% (n=4) community programs including community events, science festivals, and Science Cafes
- 18.18% (n=4) exhibits and exhibitions including permanent exhibitions, special exhibitions, lobby exhibits, and kiosks
- 13.64% (n=3) museum programs including floor programs, lecture series, museum-based programs for school groups and plays
- 4.55% (n=2) professional development for museum staff and scientists
- 4.55% (n=1) an audio CD set
- 4.55% (n=1) collaboration

- 4.55% (n=1) large format film

This group of reports would not be suitable as a representative sample for quantitative analysis involving inferential statistics. It is suitable to use the studies as cases for grounded theory development.

Discussion of Results and Findings

In this section, I present results and discuss findings related to the three questions the study aimed to address. First, I define nine evaluation focus areas identified across the 22 reports and then provide exemplars of evidence included in the nine major areas and sub-categories within them. Next, I compare the overall frequency of evaluation focus areas across types of evaluation in order to identify patterns. Finally, I present the nine major evaluation focus areas that create necessary, sequential steps leading to outcomes. Discussion of findings is included in each section.

Evaluation Focus Areas across Reports

I identified nine major evaluation focus areas that appear in one or more of the 22 reports. These included Target Audience and User Characteristics, Awareness, Motivation, Access, Usability, Use, and User Impact. In each area, I define the evaluation focus area and sub-sets of the area, discuss their sequence in a chain of connections, and present connections to other sources in various literatures. I include the percentage of the categories across the 22 reports and sample excerpts from the reports as examples of the categories.

Target Audience and User Characteristics

I identified the Target Audience and User Characteristics evaluation focus areas from reports and connected them to a variety of sources in the literature. For example, in guidelines for program producers, the Corporation for Public Broadcast (2002) set forth this important question for summative evaluation: “Have you reached your target audience?” In the sequence of steps, reaching the target audience is a prerequisite for that audience experiencing or using a technology, which is, in turn, a prerequisite for audience impact.

In the NSF-funded report *Framework for Evaluating Impacts of Informal Science Education Projects* (2008), Dierking (2008) presents a hierarchy of anticipated outcomes adapted from Well & Butler (2002). At its base are two types of variables related to target audiences:

> **Psychographic data** – motivations, interests, existing knowledge, expectations, perceptions, etc.
>
> **Descriptive data** – ages, ethnicity, gender, income, residence, distance traveled, group size, etc. (p. 28)

In the reports I analyzed, I found questions, evidence, or conclusions that focused on two general types of target audience or user characteristics:

- Demographics — descriptive information typical of population statistics such as age, income, educational level, and residence.
- Psychographics — psychological characteristics of the website users such as motivations, interests, and habits.

Among all 22 reports, 18 (81.82%) included some information about target audience or user characteristics, with 17 (77.3%) including some form of demographic information and 36.36% (8) including one or more types of psychographics. Demographics included age, gender, income, educational background, and residence. Psychographics included ability level judged by teachers, and frequency of behaviors (e.g. game play, television viewing, museum going, travel, Internet use, and interest in science).

While demographic and psychographic information appeared as an evaluation focus area in numerous studies, the implicit question behind the collection of this information was not always “Did the Web site reach its target audience?” In 9 of the 17 studies, demographic information was presented as part of the discussion of methods used by evaluators. In these instances, demographics were used to frame formative evaluation questions. Summative reports often focused on information about actual users.

Awareness

For websites, Awareness, along with Motivation, form important first steps in any assumed set of steps leading to user impact. Members of the target population have to know about and gain access in order to use the Web site or they may be motivated to search for a Web site to find some specific information or experience or accomplish a task. A variety practical fields use the idea of awareness as the first step in a causal chain leading to some desired action. The way authors of these reports use the term Motivation appeared to be similar to the idea of brand awareness used in business and marketing (Aaker & Mcloughlin, 2010, pp. 176-177; Gelder, S., 2005, p 149) that is, the extent to which a brand is recognized as available to potential users and known to be available. Yet, since users may also find websites through search engines using key search terms other than the Web site name, prior awareness is only one way users may end up using a Web site. I included information about search engines and terms in the evaluation focus area, Access. Both Awareness and Access are potential steps in a chain leading to Web site use and impact.

Across the 22 reports, I identified information and two categories related to awareness,

- Level of Awareness — the extent to which the target population, or part of the target population, knew about the Web site's existence.
- Source of Awareness — the source or channel through which respondents found out about the Web site's existence.

I found only one report in which Level of Awareness was an evaluation focus area (334). I identified Source of Awareness as an evaluation focus area in 6 (27.27%) reports.

The two following excepts provide samples of the type of information from the reports that was included in this evaluation focus area.

> The majority of the visitors (58%) found out about the Web site from the *Strange Days on Planet Earth* television series. A PBS television promotion was the next most frequently mentioned source, cited by one-quarter (24%) of the visitors. Nearly one-fifth each (18%) reported learning about the Web site from the PBS Web site or from a friend, family member, or colleague. Just under one-tenth (8%) said they found out about the

> Web site through a *Strange Days on Planet Earth* email announcement (epostcard). (76, p. 6)
>
> . . . just over half of the respondents reached the NsN [*Nova science NOW*] site by visiting the NOVA Web site first, and an additional 11% came directly to the site via a bookmark or by typing the URL themselves. Those who described an "other" means of finding the NsN Web site primarily noted they found about the site from "a TV program," with a few mentioning NOVA or NsN by name. (233_2, p. 13)

Motivation

Motivation, in the sense it was used by the authors of the report, means the reasons people used or visited the Web site. I grouped two types of information under Motivation:

- Motivation for — reasons people reported they would use or did use a Web site.
- Visit Prompts — cues that prompt a return visit to the site.

Across all evaluation types, 8 (36.36%) reports provided information about Motivation for Use. The following excerpt provides an example of the type of report information included in a formative study. These respondents were prospective Web site users.

> Before specific discussion about the *Design Squad* Web site, GRG conducted a Mind Map exercise where children described what they thought of, or would expect from, a Web site associated with a TV show. Responses focused on children's interest in being able to watch episodes of the TV show and play games related to the show, as well as interest in seeing bright and contrasting colors and shapes and seeing an overall look that corresponds with the TV program. (124, p. 3)

In contrast, this excerpt provides an example from a summative evaluation with information obtained from actual Web site users.

> A substantial percentage (42.4%) said they visited when they needed inspiration, when as Chart 1 shows, most of the visitors visited the Ice Stories Web site to find information about scientists featured on the site (67%). The second most frequently stated reason was to find information about a topic covered on the site (51%). Some visitors went on the site to find information about a specific research method discussed on the site (21%).

> Smaller groups of visitors went on the site to find information for a class they teach (12%) or to find information for kid activities (12%). (340, p. 101)

Only 2 (9.09%) of the 22 reports asked respondents what cued a repeat visit to the site. One of these reports was remedial and the other was summative.

> We asked online survey respondents what prompted a visit to *ExhibitFiles.* Respondents could select multiple reasons for visits. Of the 245 respondents to this item, 57.6% said that they visited in response to an email from the site. This response is consistent with the receipt of a monthly email from the site managers. (334, p. 42)

> As might be expected, when listserv respondents were asked what prompted their most recent visit to the site, the most frequent answer was that they followed a link in an email newsletter (34%), followed by coming to find specific information (18%) and checking in/browsing for new content (17%). Conference participants were more likely to be seeking specific information (32%) or browsing (32%)… The most frequent responses from both sets of respondents for the Projects and Toolkit sections were increased content, and notification about new content. (263, p. 4)

Access

I defined Access as the means by which users connect to the site, or the pathways through which they reached it. Means included the software or hardware used by respondents, and pathways included the traffic sources by which they entered the site. This category, and those associated with it, were identified from the reports; however, I recognized the role of Access as a potential step in a sequential, necessary set of steps from my previous work in exploring K-12 teacher access to professional development and youth access to out-of-school programs. Easy access is a prerequisite to wide-spread use by the target audience. I identified two types of access,

- Technology — the hardware, software, and technology used to link to the Web site.
- Traffic Sources — the number of entrances to the Web site from search engines, links on other sites, or direct entry as measured by Web site analytics.

Of the total 22 reports, 5 (22.73%) included Access as a focus of evaluation. Evaluators focused on the technology to access websites in 2 (9.09%) reports.

Software and hardware used to access a Web site was the focus in the formative study of a Web site that featured video stream and downloads of National Public Radio programs with an international target audience. The evaluators looked at the following areas:

> Respondents answered a variety of questions about the systems and technology they used to access the *Ganga* Web site, including the type of computer and operating system used, the type of Internet connection used, and the location from which they viewed the site. (163, p. 5)

In the summative report, evaluators cited technology as a barrier to access: " Several — including a few who used the Web site — said they do not like using the Internet other than e-mail, and would never have looked at the Web site" (370_4, p. 55).

The rationale for placing Traffic Sources as a measure of access deserves some explanation. Web site analytic packages commonly include the number of entrances to the site from search engines (including counts by popular search engines and terms used to search), from links on other websites, and from direct entry (e.g. bookmarks and typing in the URL). From the discussions in the reports, it is difficult to determine if these were being used as indicators of use, awareness, or the most literal sense, access. I chose to place Traffic Sources under access using the following rational: if a Web site is difficult to get to through search engines based on user generated search terms, it is not easily accessible for users. Problems in locating the site through search engines can be an obstacle to use in much the same way a Web site designed for hardware and software users don't have or don't like is not accessible. In this sense, Traffic Sources are an indicator of one type of Access.

Authors provided information about traffic sources in 4 (18.18%) reports. All of these reports included information about how users entered the Web site. For example,

> Nearly a half of the site traffic (49%) came from referring sites, such as *exploratorium.edu* (29,016 visits) and *images.google.com* (25,959 visits). About a third (34%) of the traffic came from search engines, such as Google (74,349 visits) and Yahoo!

> (5,076 visits). The rest (17%) of the site traffic came directly to the site through its URL. As noted in the previous section, two of the top five peaks in web traffic involved the referral site *StumbleUpon.com*, followed by *Bing*, and then the *exploratorium.edu* site. (340_2, p. 90)

Two evaluators also reported what search engines people used to reach the site and what key search terms they had used to find it. For example,

> During the 16-month period beginning September 2008, 42,700 site visitors were referred to InformalScience.org via 17 different search engines. Google was the most common search engine used, accounting for 93% of these visits. Of the visitors who accessed the site using search terms, the most popular terms were variations of the site URL (people searching for the site specifically) or the term "informal science," accounting for 8 of the top 10 search terms (see Table 1). The remaining two search terms are the name of a project and the name of an individual member. (495, p. 5)

Appeal

I defined Appeal as the extent to which users found the site as a whole, or an aspect or feature on the site, attractive, pleasing, or interesting. I identified four aspects of appeal featured in these Web site evaluations.

- Means of Engagement — the attractiveness of the specific instances and types of interaction on the site and the overall range of ways users interact on the site (e.g. reading, playing games, watching streaming video, viewing photographs, commenting, listening to audio).
- Visual Appeal — the attractiveness of the graphic elements and their arrangement (e.g. color, photos, layout, fonts).
- Content Appeal — user interest and attraction to specific topics on the site and the attractiveness of the range of information on the site.
- New Content and Features — user interest in finding new content or features on repeat visits.

Content appeal also included questions about the additional content needed, depth of content, currency of present content, and content that was lacking. Each of these evaluation focus

areas was measured through a range of indicators, including how much users liked an aspect such as visual appeal, the level of user interest in content, and how often users selected a means of engagement.

All 22 reports focused on some aspect of appeal with 20 (90.0%) of the studies exploring engagement appeal, 13 (65.0%) exploring visual attractiveness, and 21 (95.1%) accessing the content appeal.

The following conclusion provides an example of information included in the Engagement Appeal evaluation focus area.

> Visitors who like the NsN [*Nova scienceNow*] site are particularly interested in being able to watch the program online, and would like more of these features added to the site. Watching the program online was listed as the favorite aspect of the site, the biggest benefit of the site, and the primary reason people will visit again. Over half of the Follow-Up Visitors had been back to the site to watch additional segments since completing their initial survey. The most often cited request for changing the Web site focused on adding more video segments. (129_3, p. vi)

Evaluators focused on Visual Appeal in 12 (54.54%) reports. In this example from a formative study, the evaluators identified differences in visual appeal based on the type of Internet connection.

> The fourth area rated was the overall appearance of the site, including the graphics, pictures, and colors used. Ratings for this area varied based on the type of Internet connection being used. Respondents with a fast connection, who were most likely to be able to view the appearance-based features, rated the appearance of the Web site higher than those with a slow connection (mean ratings of 4.57 and 3.44, respectively). Those with a broad-band connection or better rated the appearance of the Web site as *very good* to *excellent* on average, while those with dial-up rated the appearance of the site as *good* to *very good* ($p<.01$). (163_3, p. 11)

I identified Content Appeal as an evaluation focus area in all 22 reports. In this example from a formative study, evaluators tested children's interest in several topics on one specific page of the Web site.

The children expressed interest in all the topics on the Find It Landing page, as follows:

• 60% (15) liked Science and Engineering

• 56% (14) like Math and Sports

• 52% (13) liked Money

• 44% (11) like Math and Weather

• 40% (10) liked Holidays

• 36% (9) liked Problem Solving

• 20% (5) liked Geometry, Pre-Algebra, and Using Numbers

• 16% (4) liked Measurement

• 12% (3) liked Using Data (428, p. 17)

I identified another feature of Web site appeal, appearing in 3 studies (13.63%), that encompassed both Engagement Appeal and Content Appeal, that is, the desire of users for current and new information when they return to the site. Instances included responses from both children and adults.

> Children are interested in games on a Web site, but will lose interest if they complete all the games available and then do not see new activities offered. If they see that new information, games, and interactive opportunities are on the Web site when they visit, they will be interested in checking back often. (124, p. 12)

> When asked what would prompt them to visit the various sections of the Web site more often, the most frequent responses from both sets of respondents for the Projects and Toolkit sections were increased content, and notification about new content. (263, p. 3)

Usability

The father of usability testing is Jacob Nielsen. Based on the examples in a recent book he co-authored (Nielsen & Loranger, 2006), most of his studies have focused on websites in business and government. He defines usability as,

> a quality attribute relating to how easy something is to use. More specifically, it refers to how quickly people can learn to use something, how efficient they are while using it, how memorable it is, how error prone it is, and how much users like using it. If people can't or won't use a feature, it might as well not exist. (Nielsen & Loranger, 2006, p. 21)

Classic usability testing involves observing representative users trying to complete a specific task and often asks them to think aloud as they use the Web site. While numerous authors referred to usability, I found that only 2 (9.09%) of the studies in this sample (110, 450) had tested websites in one or more ways consistent with Nielsen's usability technique, that is, through direct observation of users using mock-ups or prototypes, using talk aloud techniques, or documenting errors.

However, many of the reports focus on aspects of usability as measured by self-report through surveys and interviews. Web designers make distinctions among several design features that affect usability including navigation, search adequacy and site architecture (Nielsen & Loranger, 2006, p. 178). In most of the reports in this study, evaluators did not define and break out issues in such a precise manner. For this reason, I grouped all of these issues into one focus area, Finding Information.

- Finding Information — the ease with which users can locate specific content or features on single pages across a site

I identified three other focus areas that are aspects of Usability.

- Technical Reliability — the extent to which the site functions without technical problems (e.g. active links, playing video, playing audio, downloading PDFs, playing games).
- Age appropriateness — the ease with which users of various ages (including children) could use the site.
- Satisfaction — the degree to which users like or enjoyed using the Web site.

Overall, 18 (81.82%) studies focused on some aspect of Usability. The studies focus on the following aspects of Usability:

- 68.18% reported on Finding Information
- 59.09% reported on Satisfaction
- 31.82% reported on issues related to Technical Reliability
- 18.18% explored questions on Age Appropriateness

In summary, while only 2 studies tested Usability by observing users undertaking specific tasks, most focused on some aspect of Usability including Finding Information, Satisfaction, Technical Reliability, and Age Appropriateness.

Use

The nature and extent of use of a Web site provides the most immediate link in the steps leading to user impacts. Within this sample, evaluators focused on several different aspects of use:

- Frequency of Use — how often users visited the Website, sometimes expressed as numbers of previous visits.
- Level of Use — the level of engagement across the site or in specific sections, including time spent and the number of pages visited.
- Range of Engagement — ways users engaged with content on the site (e.g., reading, game playing, watching video) and comparisons of frequency among the ways.

A majority of the 22 studies (13, 59.1%) reported on some aspect of Use for the entire Web site or specific features or pages. In order of frequency, the reports contain evidence of the following aspects of use;

- 45.45% focused on the Level of Use
- 31.82% focused on Frequency of Use
- 31.82% reported Range of Engagement

Evaluators often explored Frequency of Use through surveys. For example,

> The second set of questions focused on how often and how recently respondents visit the *Citizen Science Central* site . . . . Of the listserv members, about 18% of the respondents

> had never previously visited the *Citizen Science Central* site, but most (70%) had visited at least a few times. (263, p. 3)

Many of the reports contain information on Level of Use from software such Google Analytics.

> On average, visitors viewed 2.15 pages per visit and spent 1.35 minutes on the site during one visit. (340_2, p. 93)

> The depth of the average users visit indicated by the number of pages visited varied across the different referring sites, possibly reflecting differences in users' interest in informal science. For example, visitors who came to InformalScience.org from the LRDC, UPCLOSE and CAISE web Center for Technology in Learning 6 sites, which are specifically about informal science, visited more pages than average (over 13, 11 and 7 pages, respectively), while visitors who came to the site from facebook.com visited fewer pages (less than 3) on average 2. (495, pp. 6-7)

Most evaluators explored the Range of Engagement through surveys and interviews.

> Visitors were asked to identify any activities they did while at the website, choosing from 11 possible response options . . . . . The visitors reported engaging in a variety of activities while at the website and most respondents indicated they engaged in at least four different activities while at the site. The four most frequently mentioned activities were done by more than half of the visitor group and included: read more in-depth information about an episode (66%), read "What do experts say?" (60%), read "What can I do?" (60%), and read "Why should I care?"(54%). (76, p. 18)

In summary, over half of the sample (59.08%) of Web site evaluation studies included Use as an evaluation focus area. Some studies reported frequency, level, and range of engagement collected through analytic software and others used self-report methods such as surveys and interviews.

User Impacts

The effect of an intervention is the final destination in the sequence of steps. For websites, this is the Nature and level of User Impact. Evaluators discussed impacts using a

variety of language and presented information about this evaluation focus area in a variety of ways. I found only two explicit statements that described intended audience goals or user outcomes (61_2, p ii; 334, p. 3). One report used a learning goals approach and the other used the impact categories form set forth by Friedman, et. al (2008). Yet, I did find numerous instances of information, findings, and conclusions that appear to be related to impact. Among the 22 cases, 11 (50.00%) included reports of information, findings or conclusions that appeared related to user impacts or audience goals.

The following excerpts provide examples of the type of information included in this focus area,

> While less than half of the users said that the web site changed the way they think about race or human variation, those who did say their ideas were changed said that the site clarified their understanding, made them more aware, dispelled myths, enlightened them, and heightened their awareness of assumptions. (146 & 148, p. 1).

> There were mixed results from interviewees about the science content on the website. Some said they did learn a few new interesting things, like about volcanology and alpine flora and fauna. But several said that they did not learn anything new, as they were familiar with the topics presented due to their science backgrounds. (259_3, p. 16)

Effectiveness of Web site within the Larger System

As I reported earlier, 20 (90.91%) of the 22 websites were part of a multi-platform system. Several studies (11, 50.00%) presented conclusions about the effectiveness of the entire system of which the Web site was a part.

For example, in a study of a public television series, authors concluded:

> Finally, this evaluation has demonstrated how different offerings from the same initiative can bolster one another. Visitors to the Web site, for example, visited primarily to watch program segments or to learn more about a topic they had heard about through the program. The bioscience classroom activities and a subset of Science Cafés also reinforced the program by building on NsN [*NOVA scienceNOW*] segment content. This intersection of initiative offerings provides the general public with multiple ways to

> continue engaging with science, in general, and with NSN topics, in particular. (129_3, p. viii)

In another instance, evaluators studying the contribution of web-based media to changing the perception of a museum concluded,

> The Tennessee Aquarium strengthened use of web-based media to communicate ocean literacy and stewardship messages by adding extensive ocean-related educational content to the main Aquarium website and by using social media outlets to pique web visitors' interest in the ocean and ocean life. . . . Many members of the Aquarium's internet audience had come to regard the Aquarium as a source of ocean-related educational content, with over two-thirds of survey respondents saying they would visit the Tennessee Aquarium website for ocean-related educational content. (309_2, p. 3)

Discussion of Evaluation Focus Areas

These 9 major evaluation focus areas provide a guiding framework for evaluators to use when designing Web site evaluation. Along with definitions and examples, there were some apparent connections among these elements, which I will discuss in relation to a guiding framework for Web site evaluation design.

Evaluation Focus Areas by Evaluation Type

The second question I asked was if any focus area were used more frequently in different types of evaluation. The purpose of this comparison was to look for patterns to explore whether these evaluators had identified any of the areas as more or less useful in different types of evaluation. There were the small number of cases for two of the evaluation types (Remedial and Don't Know); therefore, I combined these types and Formative evaluation reports into one category labeled Formative & Other. This made logical sense based on Scriven's (1967) original definitions of formative and summative evaluation as well as my own judgment that the one case classified as Don't Know was probably a remedial evaluation. Using Scriven's typology, both the Don't Know study and the Remedial study would be classified as formative evaluation. Table 1 presents the percent of each focus area by evaluation type.

Table 1. *Percent of focus areas by evaluation type*

Focus Areas	% Formative & Other*	% Summative**
Target Audience	87.50	78.57
Awareness	12.50	35.71
Motivation	37.50	35.71
Access	12.50	28.57
Appeal	100.00	92.86
Usability	50.00	92.86
Use	50.00	64.29
Intended Impacts	25.00	64.29
System effectiveness	50.00	50.00
Total *(N*=22)	100.00	100.00

Note: Formative & Other Column includes reports identified as Formative, Remedial, and Don't Know.

* *N*=8

** *N*=14

Strong conclusions should not be drawn from this comparison because the sample is too small and has such a high concentration of reports from two organizations and systems including broadcast media. Even so, patterns are apparent within this specific group of 22 reports. Usability was an evaluation focus area in 50.00% (*N* =8) of the Formative & Other reports and 92.86% (*N* =14) of Summative reports. In addition, I found evidence of a focus on Intended Impacts in only 25.00% of the Formative reports compared to 92.86% of Summative reports. Overall, both Formative and Summative studies show a relatively high percent of focus on

Target Audiences and Appeal. In addition, both Formative and Summative studies show relatively low percentages of focus on Awareness, Motivation, and Access. About half of each type of evaluation report included evidence of a focus on Use and System Effectiveness.

I identified Appeal as an evaluation focus area in all of the formative evaluation reports; other steps such as Awareness, Motivation, Access, and Usability were included much less frequently. One explanation is that, in the realm of broadcast media, Awareness, Motivation, and Access may be the responsibility of marketing specialists; reports on these topics were not shared in this database which focused on informal science learning. Yet another explanation is that testing Appeal may be a more familiar area to evaluators and developers than the other three. As I reported earlier, only two of the reports feature Neilson- type usability testing. An explanation for a lower level of this focus in formative reports may be little cross-over between techniques developed for testing of business and education websites. Further, reports written after the timeframe in this sample may include these techniques. The comparatively high level of frequency of Appeal in the formative reports compared to all other areas is striking.

The set of summative reports (N =14) also had the comparatively low levels of focus on Awareness, Motivation, and Access, which may not be problematic since none of these reports have explicit user impacts such as those described in Friedman (2008). Many of these studies were written prior to the wide-spread availability of that resource. My analysis showed only one report with impacts written using the Friedman framework. It will be interesting to analyze more recent reports to look at the level of adoption, and to see if explanations for impacts are include a full range of sequential necessary steps for the program to accomplish user impacts or if only a few are studied.

In summary, due to several factors, including the similarity of reports in the sample and the timeframe in which they were written, this comparison did not yield useful results for the development of a guiding framework.

Guiding Framework for Web site Evaluation Design

In this section, I present a guiding framework that has sequential, necessary steps leading to User Impact and System Effectiveness. Figure 1 shows a guiding framework for planning and designing Web site evaluation. In general, the framework presents evaluation focus areas from

left to right as sequential, necessary steps to accomplish User Impacts and System Effectiveness. This framework does not prescribe evaluation questions, rather, it functions as a heuristic to allow evaluators and clients to consider what specific questions are important at different stages of development and evaluation and in different contexts. The framework could also serve as a template for the development of an explicit program theory for a Web site.

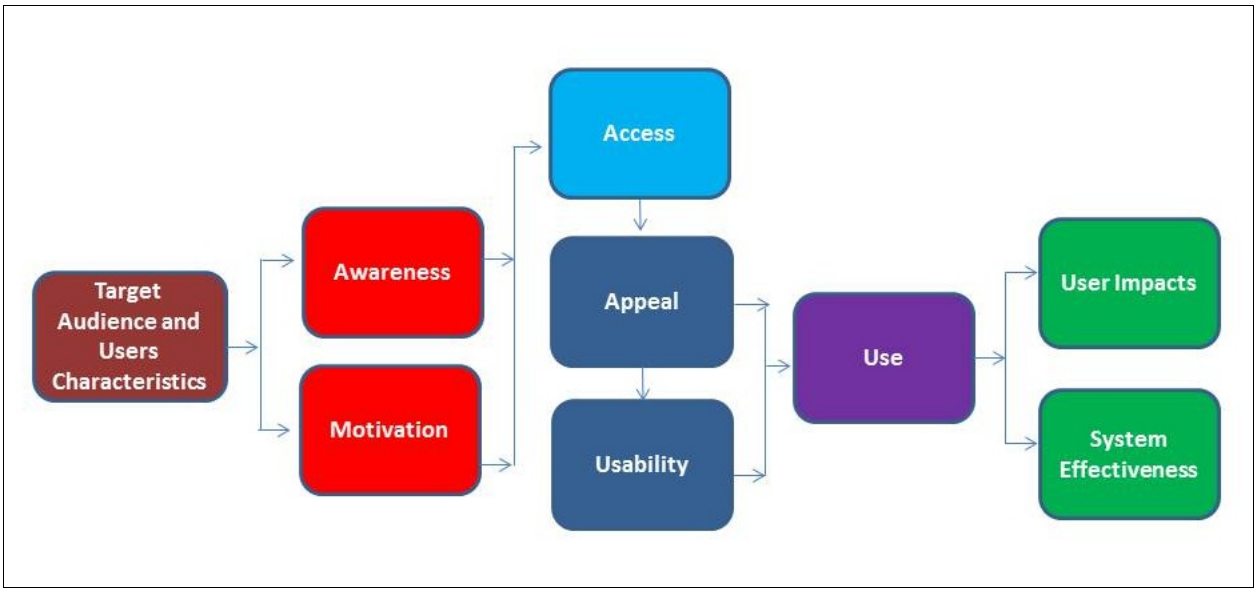

Figure 2. Framework for planning and designing Web site evaluation: evaluation focus areas presented as a causal chain leading to user impact and system effectiveness

The chain begins with a focus on Target Audience and User Characteristics. In Front-end, this area suggests evaluation questions to confirm initial assumptions about the target audience and to collect additional information about their demographic and psychographic characteristics. In summative evaluation, it supports evaluation questions such as, "Did the Web site reach its target user group?" Questions can also be generated to frame issues surrounding System Effectiveness, particularly when different components of a system have different audiences and user groups.

Next, I assumed that user Awareness and Motivation are pre-requisites for Access. Clearly, based on the analysis of the reports, some users become aware of the Web site and go directly to it; others may be motivated to look for specific information or experiences and access the Web site through search engines. After accessing the Web site, Appeal and Usability

influence the nature and level of Use. In turn, the nature and level of use (along with other factors) influence Impacts and System Effectiveness.

By generating evaluation questions based on a framework such as this, evaluators can support Web specialists in developing comprehensive strategies for user impact and testing them through front-end and formative evaluation studies.

In summative studies, using such a framework works against interpretations of what influenced impact based on guesswork and assumptions. Lack of impact may not be due to lack of appeal but to lack of usability. The effectiveness of a Web site may need to be considered in the context of its effectiveness within a larger system.

Conclusions

Through the analysis and refinement of categories, I identified and defined 9 major evaluation focus areas appearing in at least one of the 22 reports included in the sample. Major areas included Target Audience and User Characteristics, Awareness, Motivation, Access, Usability, Use, and User Impact. In addition, I identified possible connections among the major evaluation focus areas. Finally, I organized 7 of the major areas as a set of sequential, necessary steps leading to User Impact and System Effectiveness. This framework does not prescribe evaluation questions rather, it functions as a heuristic to allow evaluators and clients to consider what specific questions may be important at different stages of development and evaluation and in different contexts. The framework could also serve as a template for the development of an explicit program theory for a Web site.

Designing evaluation studies to assess user impact is only one among a number of approaches. Another approach I often use involves identifying the issues and concerns among various stakeholding groups as set forth by Guba & Lincoln in *Fourth Generation Evaluation* (1989). Designers and developers as well as funders are, often, most familiar with evaluations organized around the audience and user impact. Funders, in particular, are increasing their requirements for measuring these items, and funders are a very important stakeholding group in any evaluation of a grant-funded project. Learners in informal environments choose their own experience freely; without levels of awareness and access, that free choice is not possible. Direct cause-and-effect relationships are rare, particularly in the complex systems that make up

informal learning environments. I am not suggesting this framework as a theoretical model for experimental design measuring effects at only one level of a system. I doubt it is possible to completely explain what causes most Web site user impacts in any single study even with mixed-methods to capture different parts of the system. As this analysis indicates, many websites are embedded in larger and more complex learning systems where impacts are influenced by the Web site, other components in the learning system, and even other systems within the user's learning environment. Yet, by providing as complete a picture as possible of the influences on user impact and system effectiveness, and identifying gaps, evaluators can provide more contextually authentic and sophisticated information for decisions that improve products during development and draw better informed conclusions about their value.

References

Aaker, D. A., & Mcloughlin, D. (2010). *Strategic market management*. Chichester, West Sussex, United Kingdom: John Wiley & Sons, pp. 176-177 .

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation*, *1987*(33), 5–17.

Corporation for Public Broadcasting. (2002). Evaluating your efforts. *WGBH enhancing education*. Retrieved March 20, 2013, from http://enhancinged.wgbh.org/process/evaluation/index.html.

Dierking, L.D. (2008). Chapter 3 evidence and categories of ISE impact. In Friedman, A. (Ed.) *Framework for evaluating impacts of informal science education projects* [On-line]. Retrieved from http://caise.insci.org/uploads/docs/Eval_Framework.pdf.

Gelder, S. V. (2005). *Global brand Strategy: Unlocking brand potential across countries, cultures & markets*. Sterling, VA: Kogan Page Publishers.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine Publishing Company.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Beverly Hills, CA: Sage Publications

Nielsen, J., & Loranger, H. (2006). *Prioritizing web usability*. Berkley, CA: Pearson Education.

Screven, C. G. (1990). Uses of evaluation before, during, and after exhibit design. *International Laboratory for Visitor Studies Review*, *1*(2), 36–66.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*, 39-83. Chicago, IL: Rand McNally.

Weiss, C. H. (1998). Evaluation: Methods for studying programs and policies (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Wells, M. & Butler, B. (2002). A visitor-centered evaluation hierarchy. *Visitor Studies Today*, *5*(1), 5-11.