

Usability Literature Review

**Evaluation of State-Based Integrated
Health Information Systems
Contract: 200-96-0598, Task 23**

Submitted to:

**Centers for Disease Control and Prevention*
Epidemiology Program Office**

by:

**ORC Macro
and
QRC Division of Macro International Inc.
3 Corporate Square, Suite 370
Atlanta, Georgia 30329
404-321-3211**

September 15, 2000

***This report is provided solely as a resource and does not represent the official positions or policies of the CDC.**

CONTENTS

1. OVERVIEW: USABILITY EVALUATION.....	1
1.1 Usability Testing Methods.....	1
1.1.1 Choosing an Approach.....	2
1.2 Usability Engineering in the Product Development Lifecycle.....	3
1.2.1 Usability Methods.....	3
1.2.2 User Inquiry.....	3
2. USER INQUIRY.....	6
2.1 Gathering User Characteristics.....	6
2.2 Task Analysis.....	6
2.3 Methods of Gathering User and Task Data.....	7
2.3.1 Document Analysis.....	7
2.3.2 Feedback/Support Request Analysis.....	7
2.3.3 Questionnaires and Surveys.....	7
2.3.4 Online Surveys.....	8
2.3.5 Site Visits.....	9
3. USABILITY INSPECTION METHODS.....	10
3.1 Heuristic Evaluation.....	11
3.1.1 Personnel.....	11
3.1.2 Time Required to Conduct the Evaluation.....	11
3.1.3 When to Evaluate a Design.....	12
3.1.4 Heuristics, or Design Rules of Thumb.....	12
3.1.5 Conducting the Evaluation.....	12
3.2 Walkthroughs.....	12
3.2.1 Pluralistic Usability Walkthrough.....	12
3.2.2 Cognitive Walkthroughs.....	14
3.3 Web Log Analysis.....	15
3.4 Site Mapping.....	15
3.5 Automated Tools.....	16
4. EMPIRICAL OR USER-CENTERED EVALUATION METHODS.....	17
4.1 Usability Testing.....	17
4.2 Focus Groups.....	18
4.2.1 Preparation.....	18
4.2.2 Participants.....	19
4.3 Online Focus Groups.....	19
4.3.1 Advantages.....	20
4.3.2 Disadvantages.....	20
4.4 Prototyping.....	20
4.4.1 Types of Prototypes.....	21
4.4.2 Advantages.....	21
4.4.3 Disadvantages.....	21

Appendices:

- A-1. Literature Review: Federal Web Sites
- A-2. Lists of Heuristics
- A-3. Case Studies in Web Evaluation

1. OVERVIEW: USABILITY EVALUATION

Usability evaluation methods measure the probable effectiveness of a computer system or tool by looking at how learnable, efficient, memorable, safe, and satisfying it is for a given set of users. With many of the tools that we use every day to accomplish our tasks and goals—computers, wireless phones, Web pages—we communicate with complex computer systems. Because human beings don't speak in zeros and ones (binary code), and computers don't speak natural English, the software and the user must communicate via an intermediary—a layer of information that both user and system can interpret. We call this layer the “user interface.” So when we are evaluating a computer system or application's usability, we are evaluating how well it communicates with and supports the functional needs of its target users.

Usable interfaces possess five basic attributes:

Learnability.—The system is easy to learn. Novice users are able to complete basic tasks in a short period of time, with a minimum of training.

Efficiency of use.—Experienced users are able to reach a steady state of productivity.

Memorability.—The system is easy to remember. Users can return to it after an absence and complete tasks without retraining.

Error Prevention.—Users experience few errors while using the system, and recover quickly from them.

Satisfaction.—The system is pleasant to use.

1.1 USABILITY TESTING METHODS

Usability testing techniques are generally split into two basic types: inspection and empirical. Inspection methods are conducted by usability specialists and the product design team, and are useful for identifying major flaws in an interface before users test it. Empirical tests are conducted with actual members of the target user population.

Inspection methods are widely used because they are inexpensive and can be conducted quickly and informally. However, inspection methods are not a substitute for empirical usability testing. One study found that usability experts conducting a heuristic evaluation (an inspection technique) caught only about 29 percent of the most serious problems with an interface.¹ For this reason, many usability specialists recommend using a mixture of inspection and empirical methods throughout the usability lifecycle.

When selecting testing methods, tradeoffs are always made between scheduling, usability criteria, and cost issues. Each development process brings up new conflicts and issues, calling for different approaches to creating a usable interface. Design teams should be flexible enough to

¹ Desurvire, Heather. “Faster, Cheaper!! Are Usability Methods as Effective as Empirical Testing?” *Usability Inspection Methods*. Nielsen, Jakob and Robert L. Mack, Eds. John Wiley & Sons: New York, 1994.

choose the methods that best suit their budget and time frame. They should learn the benefits and tradeoffs to be made with each method, and select the combination that offers the best hope of valid results.

The most important thing is that testing occur early and often, so that interface problems aren't left for users to find during field tests, when they're most expensive to correct. Usability testing methods are intended to be part of an iterative design cycle, during which the interface design is tested and then redesigned to address issues raised in testing.

1.1.1 Choosing an Approach

Empirical usability testing is extremely flexible in that any given evaluation may have several objectives and still return reliable data. Inspection methods are not often able to address the same breadth of issues. It is easier to obtain specific data on user behavior (such as time spent on each task, how frequently tasks were completed without errors, and how often tasks were completed) when users are employed in the testing.

Inspection methods are often recommended early in the development cycle and when designers are deciding among competing solutions to usability problems. They are also frequently used when resources are limited. Inspection methods are not as good as user testing for identifying which features of the design are most important to users. Inspection methods cannot tell designers which problems users will have with the interface; they can only identify potential problem areas. Observed user behavior is often regarded as the final word in design; while experts may argue back and forth about the direction the interface will take, results from user tests are likely to be regarded as incontrovertible fact and accepted.

When the goal of the evaluation is to find solutions to complex usability issues, a team evaluation approach (such as a pluralistic walkthrough) is recommended; the team will be able to leverage its experience and skills to find solutions to the problems with the interface. In other words, the design team should work with individuals to discover problems, but work with teams to solve them.

A moderately sized team (6 to 8 people) is ideal for conducting usability walkthroughs, since this number brings together a sufficiently diverse group. When conducting empirical tests, testing user responses with individual subjects is often recommended, since the group dynamic can influence user responses. However, if the goal is to measure user satisfaction, group tests can be extremely useful.

Usability evaluation is most useful when customized to meet the design team's needs.

To design a test, the evaluator must:

- Understand the needs of key stakeholders (this information can be collected during face-to-face meetings between the evaluator and all stakeholders in the project)
- Document test objectives and circulate to stakeholders

- Design the test after consensus has been reached and make sure that the results are actionable (in other words, someone can make a decision and take action based on them)

1.2 USABILITY ENGINEERING IN THE PRODUCT DEVELOPMENT LIFECYCLE

Usability engineering activities take place throughout product development. Usability methodologies give developers the means to ensure ease of use across product families and in either legacy or future versions of a system. Usability engineering methods also allow developers to avoid costly redesigns after significant resources have been expended.

The most cost-effective way to incorporate usability methodologies into a system design process is to do as much as possible as early as possible, including the integration of usability into the first discussions of requirements. This ensures that the system will not have to be changed retroactively to meet usability recommendations. It also helps prevent “feature creep,” the random development of system features that don’t directly serve user needs. Although some of the pre-design methods recommended by usability specialists are often considered part of market research or product planning, traditional market research does not provide all the information required for usable design. However, an added benefit of usability studies can be an increased ability to market a product based on users’ perception of what the product can do for them, as opposed to the developer’s understanding of how it does what it does.

1.2.1 Usability Methods

The design team may choose from a variety of methods to support their efforts.² Some should be implemented early on, others should be employed later, and many may be used effectively at any stage in development. Listed below are several considerations and approaches to integrating usability into system design.

1.2.1.1 User Inquiry

In order to begin a usability process, we need to know who the users are and how they’ll use the system. The best way to accomplish this is to visit users in their working environments, to observe them as they go about their tasks (there are other ways to gather this information, but they’re less reliable than contact with actual users). In this context, the term “user” may refer to anyone whose work or tasks are affected by the system in question.

- **Identify User Characteristics.**—Understanding individual users’ computer skill sets, work experience, educational levels, ages, and other identifying information allows developers to anticipate their ability to learn the interface, and modify its complexity to meet their needs. Knowing users’ work and social contexts can help developers anticipate possible irritants and control issues. This type of information may come from market research or observing users directly. It can also be gathered using interviews or questionnaires.

² Nielsen, Jakob. *Usability Engineering*. Morgan Kaufmann Publishers, Inc., 1993.

- **Identify User Tasks.**—Users’ goals, their approaches to tasks, information needs, and means of dealing with emergencies should also be addressed. Evaluators should observe particularly effective users and their strategies, since workarounds and other user-developed strategies could inform future development. Outcomes of task analysis could include: lists of user goals, preconditions for achieving those goals, steps and actions to be performed, criteria for acceptable results, and communication needs of user groups while performing system tasks.
- **Identify User Goals.**—Analysis should cover not only the task being performed, but also the reason for the task. This will prevent creating a system that supports inadequate methods caused by limitations of current or past technologies. The evaluator should assess what truly needs to be accomplished.

1.2.1.2 Competitive Analysis

By conducting evaluations of competing products or Web sites, the design team can anticipate potential problems or consistency issues in their own development efforts. Since the existing competitor sites or systems will already be functional, they can serve as high-fidelity prototypes for the current development effort. Comparisons of multiple products allow design teams to address competing design solutions to usability problems.

1.2.1.3 Setting Usability Goals

Usability components can sometimes be in conflict with one another. Goals should be set based on analysis of target users and their desired tasks, and may vary widely from product to product. For each usability attribute, levels of acceptable performance should be part of the initial goal-setting process. For example, in a new system with no legacy versions, learnability would be the development team’s primary concern. In fact, one of the most common areas of tension is between learnability and efficiency.

1.2.1.4 Financial Impact Analysis

Conducting a financial analysis of the benefits of usability involves estimating the number of users, cost of users’ loaded salaries (including administrative costs, taxes, overhead, and benefits), and the estimated time that those users will be operating the system. Additional costs may include training costs, as well system errors, help, and support.

1.2.1.5 Parallel Design

In a parallel design process, several designers come up with different preliminary designs. They work separately, exploring design alternatives before the team settles on a single approach that will undergo more extensive usability analysis. The goal of parallel design is to generate rough drafts, not complete designs, and to come up with as many ideas as possible to solve specific interface issues. It’s a way of exploring multiple directions early in the process. Parallel design may not solve the system’s usability problems, however; more designs don’t necessarily mean better designs. The multiple designer approach simply guarantees that the design team will have several approaches to choose from.

1.2.1.6 Guidelines and Heuristic Analysis

Guidelines are lists of user-interface design principles that should be applied in every project. In a heuristic evaluation, an expert evaluator or usability professional examines the user interface in accordance with established usability guidelines. There are many different sets of interface guidelines, many of which contain hundreds of design principles. The development team should agree on the guidelines to be followed in designing the interface.

1.2.1.7 Coordinating the Total Interface

Consistency is extremely important for usability; the screens, documentation, and media that make up the interface must use individual elements consistently. Interface standards should be established as a means of ensuring consistency across families of systems and products. Code sharing and development constraints can also help ensure consistency.

1.2.1.8 Paper Prototyping

Use of paper prototypes can facilitate usability testing at an early stage in the design process, since paper prototypes can be developed more quickly and changed more easily than a fully developed system or Web site. Prototypes can also be useful in communicating design considerations to developers. Prototypes do not need to be fully functional; paper mockups of an interface design can elicit extremely useful user feedback.

1.2.1.9 Empirical Testing

In empirical usability testing methods, usability is assessed by testing the interface with real users. Empirical methods are the primary means by which user interfaces are evaluated, and user testing is the most commonly used method. In a usability evaluation process, empirical methods and inspection methods are often used together in order to test all aspects of all versions of an evolving design, as real users can be expensive or difficult to recruit.

1.2.1.10 Iterative Design

Based on the results of empirical tests and inspection methods, the user interface can be redesigned. By testing multiple iterations of a design with users and against selected guidelines, and making quick changes to a prototype or paper mockup, designers can identify weaknesses in the overall design of the interface. Solutions to some interface issues may create problems for other users. In those cases, designers should evaluate the benefits of the redesign against the number of users likely to find the new iteration problematic.

1.2.1.11 Field Testing

Usability tests can be conducted after the release of a product or launch of a Web site to gather data for the design of the next release or to anticipate new product designs. Some of this information can be gathered by analyzing user complaints or email queries.

2. USER INQUIRY

User inquiry, also called user and task analysis, is a group of methods designed to help us understand how people accomplish tasks. In the product development lifecycle, this type of data is gathered before design begins. User inquiry may yield information about:

- User goals (what they are attempting to accomplish with the task)
- What they do to accomplish their goals
- The characteristics they bring to the task (who they are and where they came from)
- The environment in which they accomplish goals and tasks
- Their skills and experience
- Their needs

2.1 GATHERING USER CHARACTERISTICS

A system's users are the individuals who will actually use it to accomplish tasks. Before beginning a design or evaluation process, system users must be identified. Representative users may include technical support staff, administrators, managers, and customers—the entire community of individuals who will use the system or its products.

Steps in user inquiry can include:

- Assembling a team from within the developing organization that regularly interacts with system users (if the system predates the design effort), or is involved in the design or marketing effort (if it is a new system or Web site)
- Brainstorming techniques to identify known and potential users
- Creating task and user characteristic matrices to model the anticipated user community. Field study and other user and task analysis techniques should either support or refute these models
- Listing characteristics of individual users and groups of users
- Testing the design team's assumptions

2.2 TASK ANALYSIS

Task analysis is a means of identifying user goals and tasks. Tasks should be analyzed before the interface is designed in order to allow the designers to make critical decisions about what actions the system should support.

There are several types of task analysis:

- **Workflow analysis** evaluates how work is done when several individuals work together in a group.
- **Job analysis** studies what an individual does for a set period of time.
- **Task inventories** list all the tasks performed by individuals using the system or Web site.

- **Task sequences** indicate the order in which tasks are accomplished.
- **Task hierarchies** study subtasks within a task.
- **Procedural analysis** follows the steps people take while doing the task or subtask.
- **Essential use cases** model tasks independent of any technology. They are an important starting point.

2.3 METHODS OF GATHERING USER AND TASK DATA

2.3.1 Document Analysis

A review of existing documents can help a Web site evaluator understand the purpose and original conceptualization of the site; stakeholders who should be interviewed or surveyed; documents that will be published on the site; and the target audience(s) for the site.

2.3.2 Feedback/Support Request Analysis

Email and phone messages that contain comments, questions, and/or feedback on a Web site can indicate users' informational needs and the types of problems they encounter. Support request analysis categorizes and analyzes textual information, such as that found in email messages.³

2.3.3 Questionnaires and Surveys

Questionnaires are an excellent way of obtaining either quantitative or qualitative data (depending on the questionnaire design), since user responses are written and can be tallied to illustrate user preferences. Questionnaires can only evaluate users' opinions about the user interface, not their behavior while using it. User testing data that illustrate actual behavior must be weighed more heavily than users' statements on questionnaires or in interviews.

2.3.3.1 Benefits

One of the benefits of questionnaires is that they can be administered without an evaluator present; forms can be distributed to users. Another benefit is that questionnaires can be distributed to large groups or geographically dispersed populations. In fact, a questionnaire could be distributed to every user of a particular system. This comprehensive coverage increases the opportunity to find differences between user groups and identify specialized needs of smaller user groups. Questionnaires are often limited to a randomly selected sample of 50–1,000 users.

2.3.3.2 Questionnaire Content

It is often effective to ask users to write down specific, critical incidents that occurred while using the system. Recording when the system performed poorly can help designers avoid worst-case issues in future redesigns.

³ Standard references on content analysis methodology include: Holsti, O.R. *Content Analysis for the Social Sciences and the Humanities*. Addison-Wesley, 1969, and Krippendorff, K. *Content Analysis: An Introduction to its Methodology*. Sage Publications, 1980.

One of the drawbacks of using questionnaires is that questions cannot be rephrased as they can during verbal interviews. Questionnaire forms should therefore be subjected to pilot testing and iterative design before they are distributed to users. As a questionnaire is really a user interface in its own right, usability principles should be in force. Questionnaires that are too long, hard to understand, or unprofessional will often get a low response rate. Revised questionnaires can be used later in the system's evolution to measure changes in user response.

Users frequently have difficulty responding to open-ended questions (for example, how using a site or product made them feel), and may simply ignore the question or answer it cryptically. Since these types of responses are difficult to interpret, most questionnaires use closed-ended questions where users simply respond with a single, easily quantifiable fact (for example, the number of times they visited a Web site). Questionnaires may also use checklists or ratings scales to obtain specific, easily tabulated responses.

Users are more likely to respond to a short questionnaire than a long one. Questionnaire or survey designers should limit questions to those directly related to the success of the project at hand.

2.3.3.3 Response Bias

With tools like questionnaires, surveys, and interviews, users may say they do one thing but in fact do another. This tendency is more evident when interviews are conducted in person. Online surveys or questionnaires administered via email may garner more accurate responses than verbal interviews when the topical matter is sensitive and users might avoid responding out of embarrassment.

2.3.4 Online Surveys

Online surveys are a cost-effective way to gather Web site information from a large and/or dispersed population.

Placement of the online survey is likely to affect response rates and the types of participants. Sample selection of questionnaire respondents is non-random (self-selected), regardless of where a questionnaire is positioned on a Web site. Demographic information collected by the survey helps to characterize respondents, but care should be taken in generalizing responses to all Web site visitors.

The questionnaire results can provide information about the variety and characteristics of users participating in the survey, how they learned about the Web site, and how frequently they have used it. Questions can be designed to elicit user responses to content and design, and to identify valuable new services and features users may want.

The process for conducting an online survey includes the following steps:

Development.—Survey development is based on research questions. The questions may be formatted using a Web-enabled survey software package. Placement of the survey is based on level of traffic through the site (which can be determined through use of Web server log analysis software), as well as the objectives of the survey.

Implementation.—A link inviting participation in the survey is placed on the Web site and responses are monitored daily. Placement of the link may be adjusted based on the number of responses received from a given location. Pilot testing can help determine the best placement. The link and the data collection instrument should remain available until the desired number of responses is obtained.

Reporting.—Following the completion of data collection and removal of the survey from the Web site, data are stored in database records for analysis. Closed-ended questions (e.g., scales and pick lists) are tabulated and summarized by category of respondent, while open-ended questions are coded and categorized.

Off-the-shelf survey software can be used to format and conduct the pilot test survey. To set up a questionnaire and conduct a survey, software should allow the user to do the following:

- Easily format and set up a survey comprised of several types of questions
- Link the survey from the Internet or make it accessible via an email package
- Easily create reports and/or export the data to other software for analysis and reporting

2.3.5 Site Visits

It is extremely important that design teams observe users in their working environments. Site visits allow developers to understand all the ways users are interacting with the system, many of which cannot be anticipated during the design process.

In a site visit, the designer or evaluator visits a user or user group on-site. There, he or she watches the users work and takes notes or videotapes them at work (if the users agree).

The observer should be as unobtrusive as possible. If he or she requires an explanation of user behavior, it is generally best to note the action and wait to see if it recurs. At the end of the site visit, the users can be debriefed and questions answered.

Users often have questions for observers, especially if the observer is there as a representative of the system design group. But the goal of the visit is to gather information, not to provide instruction, so observers should politely decline to answer questions until they have as much information as possible about the system's use.

3. USABILITY INSPECTION METHODS

Usability inspection is a generic name for a set of evaluation methods in which skilled evaluators examine a user interface for usability. Usability inspection is a way of evaluating user interface designs inexpensively, since testing with users is sometimes costly in terms of time and resources. To achieve the best results, empirical tests should be combined with inspection methods.

Usability inspection methods include:

- **Heuristic Evaluation:** Heuristic evaluations are the most informal method of usability inspection. Usability specialists determine whether the interface conforms to established usability principles, called heuristics.
- **Guideline Review:** In a guideline review the interface is tested for conformance with a comprehensive list of usability guidelines. Due to their complexity (guideline documents frequently contain hundreds of specifications), guideline reviews require a high degree of expertise.
- **Pluralistic Walkthrough:** During pluralistic walkthroughs, representative users, developers, and human factors professionals follow a scenario and talk through potential usability issues.
- **Consistency Inspections:** In a consistency inspection, system designers from a group of projects meet to see whether an interface's behavior is consistent with their designs. Such inspections evaluate consistency across a group of products.
- **Standards Inspections:** In a standards inspection, an expert inspects an interface for compliance with industry standards. These evaluations are designed to increase the usability of an interface in comparison with other systems on the market that follow the same set of standards.
- **Cognitive Walkthroughs:** Cognitive walkthroughs simulate users' problem-solving processes; the test evaluates whether the simulated user's goals lead from one action to the next correctly.
- **Formal Usability Inspections:** In formal usability inspections, a moderator is appointed to manage inspections and the inspection meeting; a design owner is responsible for design and redesigns; inspectors find problems with the interface; and a scribe records all issues identified during the meeting. Formal inspections use the following process: planning, a kickoff meeting, a preparation phase where inspectors review the interface, an inspection review where lists of usability problems are merged, and a followup phase where the effectiveness of the inspection process itself is assessed.
- **Feature Inspections:** These evaluations test the functionality delivered in a computer system or Web site and assess whether it meets the needs of the users.

For purposes of this review, we will be discussing heuristic evaluations and walkthrough methods, as these inspection methods will most likely be implemented on an informal basis.

3.1 HEURISTIC EVALUATION

In a heuristic evaluation (also called an “expert review” or “usability audit”), one or more usability professionals evaluate an application based on recognized rules of thumb (also called “heuristics”). Typically the emphasis is not on comprehensively examining the functionality of the site. More often the review is conducted in the context of use cases (typical user tasks), to provide feedback to the site’s developers on the extent to which the interface is likely to be compatible with the intended users’ needs and preferences.

3.1.1 Personnel

Number of Reviewers.—Heuristic evaluations are typically conducted by one or a small number of reviewers. Studies that have examined the number of usability problems identified in a user interface in relation to the number of reviewers⁴ have shown the advantages of involving more than one reviewer. It is difficult for any one reviewer, no matter how knowledgeable, to anticipate the full range of usability issues that a system’s users may encounter. On the other hand, there may be diminishing returns as additional reviewers are added. Having three to five reviewers examine an interface is advisable, but meaningful reviews can be accomplished with fewer.

Qualifications of Reviewers.—Because heuristic evaluations focus on the user interface design and likely user concerns, it is best if they are conducted by reviewers who are knowledgeable about industry best practices and current thinking in designing for ease of use. Experience in performing such evaluations is probably a better predictor of competence than any academic credentials.

Heuristic evaluations are best accomplished by individuals other than those who created the interface that is under review. While prior domain knowledge about the content of the Web site is helpful, it is not critical. It is useful for the reviewer to consider the business goals of the Web site, the nature of the competition, and the constraints under which the organization responsible for the Web site is operating. It is critical, however, for the reviewer to examine the Web site from the perspective of a user who may not have prior domain knowledge about the site.

3.1.2 Time Required to Conduct the Evaluation

Most heuristic evaluations can be accomplished in a matter of days. The time required varies with the size of the Web site, its complexity, the purpose of the review, the nature of the usability issues that arise in the review, and the competence of the reviewers.

The time requirements include not only a visual inspection of the site, but also an understanding of the design objectives, the range of users the site is intended to accommodate, and typical use cases. Time involved also includes documenting usability concerns and formulating design change recommendations, if required.

⁴ Nielsen, Jakob, and Robert L. Mack, Eds. *Usability Inspection Methods*. John Wiley & Sons: New York, 1994.

The Web site's stage of development may also affect the time required. A cursory review of an early stage prototype to assure the developers that they are on the right track can be done quickly, while a more comprehensive review of a fully developed site may take longer.

3.1.3 When to Evaluate a Design

Heuristic evaluations can be conducted on very early stage prototypes, including paper mockups, as well as later-stage electronic prototypes, with or without all of the back-end functionality implemented.

3.1.4 Heuristics, or Design Rules of Thumb

There are many sets of usability design heuristics. These are not mutually exclusive and cover many of the same aspects of interface design. The most commonly used is a set of interface design principles collected by Jakob Nielsen. (For a list of commonly used heuristics, see appendix .)

3.1.5 Conducting the Evaluation

Planning for a heuristic evaluation involves acquainting the reviewers with the Web site or application, specifying usability objectives, identifying the characteristics of typical users, and delineating use cases (i.e., typical task scenarios). Information on problems that may have surfaced from help-desk inquiries, user email comments, or professional critiques by media or industry reviewers should be incorporated into preparation for the evaluation.

After gathering background information on site objectives, user characteristics, and user tasks, the reviewer can proceed with a systematic examination of the site. If more than one reviewer is involved, each should work independently. The reviewer should make two passes: one to become acquainted with the flow of the interface screens, and another to focus on individual design elements or functionality.

3.2 WALKTHROUGHS

3.2.1 Pluralistic Usability Walkthrough

Pluralistic walkthroughs include three types of participants: representative users, product developers, and human factors professionals/usability evaluators. In a pluralistic walkthrough, interface screens are presented in the same order in which they would be viewed in a Web or computer interface. An alternative method would be to present an online scenario, or hypertext mockup of a specific flow of interface screens, and have users take notes on each screen as the group moves through them. Participants are asked to act the role of the user, as the user population has been defined by user and task analysis.

Participants write down the action they would take in pursuing the designated task, before any discussion takes place (the representative users are asked to speak first, to prevent their being influenced by the opinions of “professionals”). They write their responses in as much detail as possible, down to the keystroke, waiting for everyone to finish their written accounts. Then the group discusses individual results and suggestions for improvement.

The walkthrough is generally conducted early in the development process. Documentation and other materials are frequently not available; to combat the deficit in accompanying information, product designers and developers are often asked to communicate this information by acting as “living publications.”⁵

3.2.1.1 Procedures

At the beginning of the walkthrough, the participants are given written instructions and ground rules. The ground rules simply ask participants to assume the user’s role, to write down the actions they would take to complete the task at hand with any comments, not to move ahead of the group, and to hold discussion until the walkthrough is complete. A hard copy of the interface scenario is given to each participant. A different scenario is provided for each task evaluated. (An online version of an interface scenario may also be used.)

3.2.1.2 Objectives

In addition to the goal of creating a usable interface, this type of group testing has the added objective of increasing developers’ understanding of user concerns and difficulties. The method is especially effective in early assessment of usability issues, as it combines the advantages of both heuristic evaluation and focus group testing.

3.2.1.3 Benefits

Pluralistic walkthroughs give usability evaluators information on tasks where no prototype or previous version of an interface design exists. Walkthroughs are a flexible technique for examining an interface’s flow early in the development cycle; they can provide information on user satisfaction and performance before a prototype is available. A unique benefit of a pluralistic walkthrough is that it saves time by bringing together disparate groups (developers, users, usability professionals) involved in development to share information. Pluralistic walkthroughs can help win the support of developers, who often have limited opportunity to interact with users.

The use of paper mockups can allow a test group to evaluate new designs and work together to improve them without spending time creating a working prototype. The use of paper mockups is especially effective in helping the development team to identify and discuss instances where

⁵ Bias, Randolph. “The Pluralistic Usability Walkthrough: Coordinated Empathies.” *Usability Inspection Methods*. Nielsen, Jakob and Robert L. Mack, Eds. John Wiley & Sons, New York: 1994.

users are uncertain of the correct response. The inclusion of developers in the group may also allow participants to identify requirements for accompanying documentation.

3.2.1.4 Limitations

The pluralistic walkthrough method has some limitations. The group can only move as quickly as its slowest member. Since the entire group discusses the screens, the full group cannot finish testing one screen until all participants have written their responses. As a result, participants do not always clearly conceptualize the flow of the entire interface design.

When using a paper mockup to conduct a pluralistic walkthrough, evaluators may find that the functionality of the interface cannot be completely communicated. Hard copies prevent users from exploring the flow of the interface by browsing through it, which is a commonly employed behavior when learning to use a new interface.

3.2.2 Cognitive Walkthroughs

Cognitive walkthroughs were developed to test interfaces that can be learned by browsing (such as Web sites). Multiple walkthroughs may be conducted during a single design process. Cognitive walkthroughs can be conducted either with an individual or with members of the design team.

In a group walkthrough, an aspect of the design is presented to a group of peers (not including users and usability professionals, unlike pluralistic walkthroughs) to initiate discussion. The feedback from this discussion is then used to improve the next version of the design. The design can be presented as a paper mockup, as a minimal prototype, or as a fully operational prototype.

In an individual walkthrough, an expert simulates user activity in carrying out tasks. Individual walkthrough can be used very early in the design process to gain feedback from designers and developers. These individual walkthroughs are generally followed by a group walkthrough with the design team.

3.2.2.1 Procedures

Before a group cognitive walkthrough, evaluators agree on representative user groups, tasks to be accomplished by the design, and actions required for accomplishing those tasks. Tasks should be selected based on assessment of the target users (defined during market research, user analysis, and/or requirements analysis).

At a group walkthrough, the presenter should include:

- A detailed description of the design or design element (demonstrated by a paper mockup or functional prototype)
- A task scenario

- Descriptions of the users and the context in which the interface element will be used
- Description of the flow of actions or steps users take to accomplish given tasks

The design team then analyzes the actions the user would need to take to complete the task. The interface should clearly lead the user through the required actions.

3.2.2.2 Benefits

Cognitive walkthroughs are excellent for pinpointing where the design team's understanding of a task may be incorrect or incomplete, such as poor labeling of menu titles, icons, and buttons; and where the team has poor or incomplete feedback about the user's progress through the task. Because this method is conducted with members of the design team, it is a low-cost evaluative tool. Cognitive walkthroughs can be beneficial when conducted in any phase of the development process.

3.2.2.3 Limitations

The cognitive walkthrough method is designed to expose design problems that might get in the way of the users' ability to learn while exploring the interface; solutions suggested during these evaluations will be focused toward improving users' overall ease of learning. This could create problems where design tradeoffs must be made. For example, a feature designed to increase productivity for advanced users may not test well using this method.

3.3 WEB LOG ANALYSIS

Analysis of Web server transaction logs provides comprehensive information about Web server traffic. Knowledge of server traffic can provide information about who is accessing a current Web site, what site they are coming from, and when and where they are visiting. This type of data can be beneficial in assessing what pages on the site receive the most frequent traffic and who is using them. This can help the design team to further identify target user groups for their current site or for redesigns.

Log analysis is typically conducted as an automated procedure with log analyzer software. During log analysis all Web server activity is recorded. This includes data such as the IP address and/or domain of the individual requesting a Web page from the server, the date and time the request was made, the filename of the page accessed, and the number of bytes of data served.

3.4 SITE MAPPING

A graphical representation of the document structure of a Web site provides an excellent overview of the site content and the relationships between the pages. Site mapping can help the design team visualize the structure of a current Web site (especially helpful in cases where a site's growth may not have been well planned or organized) and find any weaknesses in the structure.

There are a number of automated site-mapping tools on the market. Common errors detected include broken links, HTML errors, server errors (based on timing out due to a broken CGI script or similar problem), and problems with page titles (e.g., duplicated page titles).

3.5 AUTOMATED TOOLS

Several tools are being developed to automate testing and assessment of Web site usability and accessibility. Usability testing cannot be fully automated, however; when possible, interfaces should be tested with users. Current automated usability evaluation tools include:

- Web Metrics, developed by the National Institute of Standards and Technology (NIST) as a suite of automated usability assessment tools. One of these tools, WebSAT, checks the HTML of Web pages against a set of usability guidelines to identify potential problems to be investigated in manual usability testing. Another tool, WebVIP, is a usability testing tool that can be used with a given set of tasks.

<http://www.nist.gov/webmetrics>

- Bobby, developed by the Center for Applied Special Technology (CAST), is an automated tool that tells users whether their Web pages are accessible to people with disabilities. Bobby reports are based on a set of accessibility guidelines issued by the World Wide Web Consortium. Bobby provides users with a page-by-page analysis of their HTML code; in Bobby reports, possible accessibility problems are identified and solutions are suggested.

All Federal Web sites must be made accessible under Section 508 of the Federal Rehabilitation Act. Private Web sites may also face accessibility challenges under Title III of the Americans with Disabilities Act (ADA).

<http://www.cast.org/bobby>

- Web site Analysis and Measurement Inventory (WAMMI) is a standardized Web questionnaire developed in Europe. It consists of 20 questions selected to collect subjective ratings of a Web site's ease of use on a series of design aspects. WAMMI can be used for monitoring users' experiences and benchmarking a Web site relative to other sites. An international database of results has been compiled, and test results can be analyzed and compared with how users rate Web sites generally.

<http://www.nomos.se/wammi>

4. EMPIRICAL OR USER-CENTERED EVALUATION METHODS

Usability testing reveals the extent to which the Web site meets the users' needs and expectations, and the extent to which it can readily be used. It can also assess the Web site's accessibility for users with disabilities. Individuals recruited to participate in usability testing should be selected from among members of the intended audience of the Web site.

4.1 USABILITY TESTING

Unlike expert critiques, usability tests are conducted using representative Web site users. Users are systematically observed as they perform tasks. The tests are conducted by a human factors engineer or other usability professional using a test scenario script. Such tests produce high-quality data, and can reveal the extent to which a Web site or application meets users' needs and the extent to which it can be used and/or learned readily. The test often is conducted in a laboratory setting with audio-video equipment to record and measure performance. Typical measures include:

- Incidence of various usability problems (derived from observations of performance or user comments)
- Time required to accomplish specific tasks or subtasks
- Nature and incidence of various user errors or failures to accomplish tasks
- Subjective ratings of user satisfaction along various design dimensions

A typical process for conducting a usability test would include the following steps:

Planning the Test.—The test administrator becomes acquainted with the Web site or application and identifies specific usability issues. With the assistance of the design team, the test administrator defines the test objectives and clarifies performance measures. The test administrator then develops the experimental model, and determines both the number and characteristics of participants required for the test and the appropriate configuration of recording equipment to be used, if any.

Preparing the Test.—The necessary equipment, both that which the participant will use and any observational recording equipment, is set up, the materials to be used are readied, and the participants are recruited and scheduled. Materials typically needed include task scenarios (i.e., the tasks to be accomplished by the test participants), notes for briefing and debriefing the participants, and any questionnaires for gathering demographic information from participants or quantifying their perceptions of the site. Often a coding scheme is devised to facilitate the collection of observational data with regard to specific behaviors, events, or expected participant comments.

Data Collection.—Test participants are observed individually as they attempt the predefined tasks. Sessions are usually videotaped; a real-time, scan-converted image of the user's computer screen can also be informative. Depending upon the purposes of the test and the anticipated usability challenges, the test administrator may observe relatively unobtrusively or may carry on

a running dialog with the test participant to obtain user feedback on various design issues. Of interest may be the participants' performances, how they go about accomplishing tasks, and/or their comments as they proceed. The data collected can consist of notes, the time to accomplish various tasks (e.g., search for specific information), and participant questionnaire responses. A data-logging software package may be used to facilitate the collection of time-stamped observational notes.

Analysis.—The analysis phase may involve compiling and categorizing usability problems observed, transferring data logs to a database package or spreadsheet in order to better summarize the coded observations, or calculating summary statistics on the subjective ratings data collected. Audio and video recordings of test sessions can be reviewed as needed. Typically an attempt is made to categorize the severity of the usability problems that emerged, taking into account the effect on user task performance, incidence of problems, and the frequency of each problem's occurrence.

Reporting.—The usability test objectives, methods, results, and any design recommendations are documented in a written report. Design change recommendations for improving the Web site are offered as needed. Often the suggested design changes involve screen design or information architecture. If sufficient cost estimates and return on investment data are available, a cost-benefit analysis of alternative means for dealing with design deficiencies may be helpful in deciding how to fix the usability problems observed.

4.2 FOCUS GROUPS

The flexible environment of a focus group provides subjective feedback. Focus groups are a source of information on why people make certain decisions, how they arrive at decisions, and how they might respond in proposed situations. Focus groups can gather information about user needs and responses either before the interface has been designed or after its release.

In a focus group discussion, 6 to 9 representative users are brought together for a period of not more than 2 hours. Either a member of the design team or an outside evaluator can moderate the group to keep it focused on issues relevant to the interface design. The moderator generally follows a preplanned guide or script, which lists topics for discussion. Because focus groups allow users to interact, they provide observers with an opportunity to observe group dynamics. These dynamics can also create difficulties, however, as individual opinions may be unduly influenced by group thought. Moderators should take care that all members of the group participate fully and have an opportunity to respond to questions.

4.2.1 Preparation

When preparing for a focus group, a moderator creates a list of discussion issues and establishes goals for the types of data to be gathered. When the session has come to a close, the moderator writes a short report summarizing the group's overall impressions that includes several quotes. More detailed analyses can be conducted, but these can be time-consuming due to the free-flowing nature of the discussion.

4.2.2 Participants

In order to maintain a flow of conversation and to have diverse points of view represented, focus groups should be run with at least six participants. In order to maximize the effectiveness of the tests, it is better to run more than one group. Although a number of firms specialize in recruiting focus group participants, the recruiting process can be costly and time-consuming, and the firms may not have experience recruiting appropriate subjects for Web site evaluation.

Because the method is based on asking users what they want rather than measuring or observing how they actually use things, focus group participants may think they want one thing but really need another. This problem can be minimized by exposing the users to concrete examples of the technology being discussed in the group.

4.3 ONLINE FOCUS GROUPS

Conducting focus groups online obtains the same information as face-to-face focus groups, and has other benefits. Anyone in the world with a computer, Web browser, and Internet access can participate, moderate, or administer an online focus group. State, regional, and national boundaries are eliminated, whereas in a typical focus group, participation is limited to an immediate area. A transcript can be automatically produced, eliminating many hours of transcription labor. On the downside, online focus group participants must have computer access and a basic level of computer literacy.

A typical process for conducting an online customer satisfaction focus group would include the following steps:

Recruitment.—The moderator works with the client to determine the user population from which to recruit the online focus group members. Text for a recruitment screener (a screener is a list of criteria that participants must meet) is developed. A link to the screener may be placed on the client's Web site. Responses are monitored, and a followup email is sent to those individuals qualified and interested in participating. A test login site is set up; potential participants attempt to log in ahead of time to be sure they can connect. A reminder email is sent the day before the scheduled focus group to those participants who have successfully completed the test login. Potential focus group members would be oversampled to account for "no-shows" on the day of the focus group.

Preparation.—The moderator develops a guide, online presentation materials, and an online feedback survey for the online focus group. Within one week of the online focus group, a walk-through is conducted with the administrator, moderator, and a few "dummy" participants. This ensures that the system is functioning, and gives the moderator a chance to review the materials and/or procedures one last time. During this time test logins may also be conducted by anyone who wants to observe the online focus group "silently."

Implementing the Online Focus Group.—The moderator welcomes participants and observers, and describes the ground rules. The online focus group is conducted in the same manner as a face-to-face focus group (i.e., the discussion is moderated, and everyone is given a chance to

respond to questions). Once the online focus group is underway, questions and answers occur in “real time.” Following the online focus group, a complete transcript is produced and edited. Results from the online feedback survey administered at the conclusion of the focus group session are tabulated and reported, and a final written analytical report is produced.

4.3.1 Advantages

In an online focus group, distance and travel costs are eliminated for both participants and moderator/administrators. Online focus groups can easily recruit participants from across state, region, and even national boundaries. In addition, online focus groups are particularly appropriate for topics that relate to technology itself such as online simulations, company Web sites, online advertising, and online databases and information sources.

Comments may be more thoughtful and useful when participants are required to put them into writing. Transcripts can be automatically produced, eliminating many hours of labor to transcribe recorded conversations. Online focus groups may provide more objective information than face-to-face focus groups, since participants are not as easily influenced by group dynamics or moderator appearance and personality.

4.3.2 Disadvantages

Online focus groups face the disadvantages created by the technology: Participants must have computer access and a basic level of computer literacy; typing on a keyboard can be slower than talking aloud; Internet-based conferencing—especially using dial-up modem connections—is not a perfect process. Power surges, phone line traffic, computer lockups, server malfunctions, and other maladies present occasional challenges. On the personal level, online focus group moderators give up the ability to observe facial expressions, body language, side conversations, and other group dynamics.

4.4 PROTOTYPING

Systems do not have to be complete for usability testing to occur. Early tests can be performed on simple prototypes. Prototypes are easily changed drafts of all or part of a user interface. Prototypes can be built faster, and redesigned more cheaply, than a finished interface. In an iterative design process, prototypes can undergo multiple revisions while the design team perfects the interface elements.

Prototypes can be extremely low-tech paper mockups or high-fidelity interactive models of the finished system interface. Techniques for prototyping include:

- Traditional artistic media like chalk, markers, pens on whiteboard, paper, overhead projector, or sticky notes
- Word processing programs like Microsoft Word, with screens printed for simple testing
- Programs like Microsoft PowerPoint to simulate the interactivity of the interface screens

- Advanced tools like Macromedia Dreamweaver or another HTML editor to create low fidelity simulations of the completed Web interface

4.4.1 Types of Prototypes

The design team can choose to use several different types of prototypes:

- **Vertical prototypes:** In a vertical prototype, the designer limits the features in the model. This results in a prototype that includes in-depth functionality, but only for a few specific tasks. A vertical prototype will enable the design team to test only part of the system, but the test will closely mimic real user situations.
- **Horizontal prototyping:** In a horizontal prototype, the system's functionality is reduced. Representative users will see a model of the interface where real tasks cannot be performed. The primary benefit of a horizontal prototype is that it can be created quickly and it can help the design team analyze how well the entire interface works together.
- **Scenarios:** A scenario simulates the user experience along a single planned path. Scenarios are useful early in the design process for getting user feedback without the cost of producing a working prototype.

4.4.2 Advantages

The goal of prototyping is to save costs and time in relation to the overall development lifecycle. Prototypes have other advantages as well. Because they can be changed quickly, prototypes encourage discussion and commentary by representative users. Building a prototype doesn't require advanced development tools; it's a simple technique that almost anyone can implement. Alternative designs can be created quickly; no money or time is lost by throwing out inadequate designs since the prototype does not require a significant investment.

Prototypes can be an excellent way of communicating design issues to developers. The pitfall of this method is that the prototype can include details that may not be part of the final design; the design team should be aware of which elements of the prototype may not be included in the final design and inform developers accordingly.

4.4.3 Disadvantages

There are some disadvantages to prototyping. For example, a prototype only simulates or shows some of the finished system's functionality. Several prototypes may be thrown away, which can be perceived by some as a waste of effort. When paper mockups are used, users may not react as they would to a working model.

APPENDIX A-1
LITERATURE REVIEW: FEDERAL WEB SITES

A review of the current literature on Web site evaluation as well as reviews of sites with similar missions can help identify successes from which designers can learn. Several pieces of Federal legislation and executive orders are relevant to the development and management of government Web sites:

- Clinger-Cohen Act. The Information Technology Management Reform Act (ITMRA, also known as Clinger-Cohen) is intended to address the management of information technology in the Federal Government. <http://www.itpolicy.gsa.gov/mke/capplan/cohen.htm>
- Government Performance and Results Act of 1993 (GPRA). GPRA seeks to improve the effectiveness, efficiency, and accountability of Federal programs by mandating that Federal agencies set strategic goals, measure performance, and report on the degree to which those goals are met. <http://www.npr.gov/initiati/mfr/>
- Executive Order #13011, Federal Information Technology. This Executive order links the ITMRA, the Paperwork Reduction Act (PRA), and GPRA. It formalizes the OMB's oversight of information technology (IT) management and stresses the importance of accountability, mission- and performance-based planning, and implementation of Federal IT. <http://www.whitehouse.gov/search/executive-orders.html>
- Executive Order #12862, Setting Customer Service Standards. This Executive Order defines the standard of quality for services provided to the public as "customer service equal to the best in business" and requires all executive departments and agencies that provide significant services directly to the public to develop and meet service standards. <http://www.whitehouse.gov/search/executive-orders.html>
- Section 508 of the Federal Rehabilitation Act of 1973, as amended (1998). This section of the Act requires that when Federal agencies develop, procure, maintain, or use electronic and information technology, they must ensure that it is accessible to people with disabilities, unless doing so would impose an undue burden. Federal agencies that provide information to the public or to their employees through Web sites must ensure that such sites are available to all persons with Internet or Intranet access, including persons with disabilities.
<http://www.access-board.gov/eitaac/section-508-q&a.htm>
<http://www.usdoj.gov/crt/508/508home.html>

A Web site evaluation study should also include a review of current standards and evaluation criteria. At the present time, there have been several efforts to establish evaluation criteria and development standards for public health information Web sites. The most notable is the Health on the Net Foundation Code of Conduct (HONcode) for medical and health Web sites. <http://www.hon.ch/HONcode/Conduct.html>

Comprehensive Web site reviews are occasionally published. One of the most useful is the Hert and Marchionini (1997) study of the BLS Web site.⁶ Hert and Marchionini evaluated the BLS Web site, Current Population Survey (CPS) Web site (co-sponsored by BLS and the Bureau of the Census), and the FedStats Web site (sponsored by the Interagency Council on Statistical Policy). The objectives of the study were to “determine who uses these services, what types of tasks they bring to the sites, what strategies they use for finding statistical information, and to make recommendations for design improvements.” The BLS study used a variety of evaluation methodologies to address different aspects of the Web sites from both internal (developer) and external (user) perspectives. The research was divided into two phases: investigative and analysis of user activities. During the investigative phase, Hert and Marchionini sought to clarify the objectives of the site and understand the organizational perspective of the site’s history and development processes. They used five data gathering methods: literature and Web site reviews, expert critiques, site mapping, document analysis, and interviews. During the user activities phase, the investigators focused on collecting data specifically related to how users accessed the site. This included online interviews and focus groups, content analysis of email requests, impressionistic analysis of online comments, usability tests, and transaction log analyses.

Undoubtedly, additional studies will continue to be conducted and published. An online bibliography of selected resources related to Web site evaluation can be found at <http://istweb.syr.edu/~mcclure/Web.Eval.Bibl.May20.html>.

⁶ Hert, C. A., and Marchionini, G., *Seeking Statistical Information in Federal Web sites: Users, Tasks, Strategies, and Design Recommendations*. Final Report to the Bureau of Labor Statistics, 1997. <http://ils.unc.edu/~march/blsreport/mainbls.html>.

**APPENDIX A-2
LISTS OF HEURISTICS**

Heuristics are also known as “design rules of thumb.” Usability specialists use these lists of design standards to identify potential problems with a user interface. Although there are several published lists of usability heuristics, two of the best-known were published by Jakob Nielsen and Larry Constantine. These can be summarized as follows:

Nielsen’s usability heuristics:

- **Visibility of system status:** The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
- **Match between system and real world:** The system should speak the user’s language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms. It should follow real-world conventions, making information appear in a natural and logical order.
- **User control and freedom:** Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. The system should support undo and redo.
- **Consistency and standards:** The system should follow platform conventions. Users should not have to wonder whether different words, situations, or actions mean the same thing.
- **Error prevention:** Even better than a good error message is a careful design that prevents a problem from occurring in the first place.
- **Recognition rather than recall:** Make objects, actions, and options visible, so the user does not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
- **Flexibility and efficiency of use:** Accelerators—unseen by the novice user—may often speed up the interaction for the expert user to such an extent that the system can cater to both inexperienced and experienced users and allow users to tailor frequent actions.
- **Aesthetic and minimalist design:** Dialogues should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
- **Help users recognize, diagnose, and recover from errors:** Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
- **Help and documentation:** The ideal system can be used without documentation, but it may often be necessary to provide help and documentation. Any such information should be easy to search, task focused, list concrete steps to be carried out, and not be too large.

Constantine’s usability principles:

- **Structure Principle.** Organize the user interface purposefully, in meaningful and useful ways that put related things together and separate unrelated things based on clear, consistent models that are apparent and recognizable to others.
- **Simplicity Principle.** Make simple, common tasks easy to do, communicating simply in the user’s own language and providing good shortcuts that are meaningfully related to longer procedures.
- **Visibility Principle.** Keep visible needed options and materials for a given task without distracting the user with extraneous or redundant information.

- **Feedback Principle.** Keep users informed of actions or interpretations, changes of state or condition, and errors or exceptions using clear, concise, and unambiguous language familiar to users.
- **Tolerance Principle.** Be flexible and tolerant, reducing the cost of mistakes and misuse by allowing undoing and redoing while preventing errors wherever possible by tolerating varied inputs and sequences and by interpreting all reasonable actions reasonably.
- **Reuse Principle.** Reduce the need for users to rethink and remember by reusing internal and external components and behaviors, maintaining consistency with purpose rather than merely arbitrary consistency.

APPENDIX A-3
CASE STUDIES IN WEB SITE EVALUATION

This document illustrates how usability methods can be modified to fit different evaluation processes and goals, and how a variety of methods can be used in a single evaluation to gather a wide range of data. Each case study identifies the type of data gathered by each method and the method's placement in the chronology of the overall evaluation.

1. MEDLINEPLUS

“MEDLINEplus Interface Evaluation: Final Report”

Date: August 1999

Reviewer: Keith Cogdill, Ph.D., University of Maryland Human-Computer Interaction Laboratory

Site URL: <http://www.nlm.nih.gov/medlineplus>

Report URL: <http://www.clis.umd.edu/megasite/toc.html>

1.1 PROJECT SUMMARY

The MEDLINEplus system provides members of the public with Web access to sources of health information, helping them make informed health-care decisions. These information sources include hyperlinks to health-related content on the World Wide Web and records of recent journal articles.

1.2 OVERVIEW OF TESTING METHODS

The evaluation of the MEDLINEplus interface incorporated two types of evaluation, heuristic evaluation and usability testing. Heuristic evaluations, or expert reviews, identify general interface and design issues based on a predetermined series of evaluation criteria. Usability evaluations gather more specific feedback from users about interface elements and site tasks.

1.3 HEURISTIC EVALUATION

1.3.1 procedures

An expert panel of three usability professionals was convened for analysis of the MEDLINEplus system. Panel participants were Human-Computer Interaction faculty at the University of Maryland. Panelists conducted independent analyses of the interface based on widely used lists of heuristics (design rules of thumb). When they completed their evaluations, they met to discuss their findings, reach consensus on usability issues, and recommend enhancements.

1.3.2 results

The panel praised the MEDLINEplus interface for its uncomplicated layout, which showed a high degree of consistency and good overall organization. System pages were compact and could be printed easily. Graphics were used sparingly, so pages downloaded quickly. Page and text

formatting was consistent throughout the site, and important keywords were used consistently. Reviewers felt that users who were not health professionals would easily understand the content in narrative portions of the site. Dates of site updates were readily available, helping users identify timely information.

Reviewers found that the site was organized around categories of information, but that a system of organization based on user goals might be more helpful. The emphasis on information was also evident in the site's name and banner logo graphic.

Lists of health links were arranged alphabetically within the site. Panelists thought those links could also be broken out under topical headings (e.g., listing types of cancer under a "Cancer" heading). They also thought that the site's global navigation menu could be "fanned out" to provide access to subcategories of information on top-level pages.

Many reviewer comments focused on the site's search feature. The search interface was inconsistent with similar interfaces in other NIH sites. Panelists believed that common structure and terminology would increase the usability of search features across related Web sites. Reviewers also found that the MEDLINE*plus* search feature did not allow users to conduct phrase searches.

1.4 USABILITY TESTING

1.4.1 Procedures

To recruit participants for usability testing, questionnaires were distributed and completed at primary care practices in Maryland and the District of Columbia. Several respondents were disqualified due to predefined screening criteria such as inexperience with the Web or medical expertise (one woman was a nurse). Nine volunteers eventually participated in testing.

Prior to testing with volunteer users, testing procedures were pilot tested to identify ambiguities in task statements and time limits for completion.

During testing participants were given 10 minutes to explore MEDLINE*plus* before attempting tasks. (Each task had a 20-minute time limit.) A video camera recorded activity on the computer monitor, but participants' faces were not recorded. Tests were based on users' success in completing five tasks, which were developed with staff at the National Library of Medicine (NLM). (Criteria for task completion were also developed with NLM staff.) Tasks included:

- Find information about whether a dark bump on your shoulder might be skin cancer.
- Find information about whether it's safe to take Prozac during pregnancy.
- Find information about whether there is a vaccine for Hepatitis C.
- Find recommendations about the treatment of breast cancer—specifically, use of mastectomies.
- Find information about dangers associated with drinking alcohol during pregnancy.

After completing all tasks, participants were asked to comment on:

- Their impression of MEDLINE*plus*
- Problems that arose during tasks
- What they felt was “best” about the interface
- What they felt was “worst” about the interface

After this debriefing, participants completed a post-test questionnaire that measured their degree of satisfaction with the interface.

1.4.2 Quantitative data

Quantitative data collected during usability testing of MEDLINE*plus* measured users’ performance in completing tasks. The reviewer collected information on completion, including instances when the task was voluntarily terminated and the participant exceeded the 20-minute time limit. Collected data also included rates of completion (finding a vaccine for hepatitis C took the least amount of time, at an average rate of 5 minutes). The reviewer recorded browser data, such as how many pages were downloaded for each task.

1.4.3 Qualitative Data

Qualitative results focused on user responses, both positive (“It’s got a lot of information. You could find just about anything you wanted if you looked long enough.”) and negative (“Why would they have two different kinds of cancer, but not one for breast cancer? That’s confusing.”).

Reviewers used qualitative responses to identify patterns of usability issues. For example, several participants had difficulty with alphabetical lists of health topics when searching for specific types of cancer. (“I looked for a list of types of cancers.”) These responses supported issues identified in the earlier heuristic evaluation of the site.

1.5 RECOMMENDED ENHANCEMENTS

Reviewers concluded that the most fundamental improvement to the usability of the interface could be accomplished by increasing consistency between the organization of the Web site and users’ goals. Reviewers also recommended improving browsing, the site’s search feature, and boundaries between MEDLINE*plus* and external sites.

2. BUREAU OF LABOR STATISTICS

“Seeking Statistical Information in Federal Web Sites: Users, Tasks, Strategies, and Design Recommendations”

Date: July 1997

Reviewers: Carol A. Hert, Ph.D., and Gary Marchionini, Ph.D.

Site URLs:

BLS: <http://stats.bls.gov/blshome.html>

CPS: <http://www.bls.census.gov/cps/cpsmain.htm>

FedStats: <http://www.FedStats.gov>

Report URL: <http://ils.unc.edu/~march/blsreport/mainbls.html>

2.1 PROJECT SUMMARY

This 9-month study of three government statistical Web sites (the Bureau of Labor Statistics (BLS) Web site, the Current Population Survey Web site, and the FedStats Web site) gathered data on who used the services, what tasks they accomplished, and what search strategies for statistical information they used. The final report included design recommendations.

2.2 OVERVIEW OF METHODS

The broad range of users and services tested for this study required a correspondingly broad range of usability methods. Methods employed in the study included:

- Literature and Web site reviews
- Site mapping, document analysis
- Individual interviews and email questionnaires with staff at government agencies
- Focus groups
- Content analyses of user email requests
- Usability tests
- Transaction log analyses

2.3 EXPERT CRITIQUES (HEURISTIC EVALUATIONS)

In expert reviews conducted by Hert and Marchionini, Web sites were independently evaluated to identify content, structures, and areas for improvement. Examinations were based on task scenarios constructed by evaluators (e.g., “explain what the geometric mean formula means for the CPI”) and reconstructed session patterns recorded in server logs.

2.4 INTERVIEWS

Reviewers conducted interviews with analysis and help desk personnel at BLS and the Census Bureau to gather data about users. Interviews were conducted in person and by telephone and generally lasted an hour. After the discussion, notes were emailed to participants for clarification. Interview topics included content and context of the service, user information and feedback, and strategies used by staff and others.

A second round of interviews was conducted through email questionnaires sent to people responsible for the FedStats site. Few usable responses were returned. A third questionnaire was sent to the FedStats Task Force on an electronic mailing list. A fourth and final round of questionnaires was emailed to help-desk staff at various government agencies.

2.5 FOCUS GROUPS

Three focus groups were held in Bloomington and Indianapolis, IN. Nineteen participants were identified through contacts with the FedStats task force. Focus groups were asked:

- What is your role in helping people find Federal statistical information?

- What types of questions do the public ask about Federal statistics?
- Which tasks lead the public to ask for Federal statistical information?
- What types of information or data do you provide to the public in response to those questions or tasks?
- How do you help people find statistical data now?
- How do you think people go about finding statistical data on their own?
- How might a Web-based service (such as FedStats) affect how people find statistical information?
- What else is important for us to understand about the public's use of Federal statistical information?

2.6 EMAIL CONTENT ANALYSIS

To find patterns in textual data, the reviewers performed content analysis on 2 months' worth of the hundreds of email requests BLS receives each month. These requests provided insights into the types of problems users had with the system and statistical problems that they encountered on Web sites in general.

As messages were read, an analyst identified categories of queries and requests. When no additional categories could be identified, messages were coded based on the list of categories. The coding scheme included content dimensions (what type of information was requested) and strategy/question dimensions (what the user wanted to know and what form the question took).

2.7 USABILITY TESTS

Usability tests were conducted with groups in Bloomington, IN, and Washington, DC. Testing groups included between four and seven users. Each user sat at a computer terminal and explored the system for 1 hour. The research team designed scenarios that explored functionality on the FedStats site in terms of tasks related to statistics location rather than statistics use.

During tests participants were asked to write their responses on the sheets of paper that described the scenarios. They were also asked to use the Bookmarks feature in their browsers to note pages that were helpful (the use of frames on the FedStats site made this tool unusable).

After exploring, users filled out a demographic and user satisfaction questionnaire. Sessions concluded with 1-hour group debriefing interviews to gather additional responses.

2.8 TRANSACTION LOG ANALYSIS

The system's Web server software logs every request for information it receives. BLS receives more than one million requests for information ("hits") per month. The data gathered by the server logs represent the entire activity of the site's user population for that month.

Server log analysis can be an inaccurate measure of user behavior, since users may return to the same site multiple times, and browser caching can interfere with the log data. To accomplish

server log analysis for the BLS site, reviewers examined summary reports that BLS produces each month. They also analyzed overall usage patterns and specific patterns of use.

3. U.S. DEPARTMENT OF EDUCATION

“Evaluation of Selected Web Sites at the U.S. Department of Education: Increasing Access to Web-Related Resources”

Reviewers: Carol A. Hert, Ph.D., and Charles McClure, Ph.D.

Date: January 1999

Site URL: <http://www.ed.gov>

Report URL: <http://iis.syr.edu/webeval/>

3.1 SUMMARY

The U.S. Department of Education (DOEd) evaluated its Web sites in 1999. Goals for the evaluation project included:

- Identifying factors that affected success and usability of selected DOEd Web sites
- Examining processes by which Web sites are managed and coordinated
- Reviewing navigation issues to discover how, and how easily, users could locate information and services through DOEd Web sites
- Assessing DOEd’s policy system for operating Web sites and determining whether the system recognizes government-wide Web management and development policy
- Providing sample evaluation techniques that DOEd could use in future efforts

3.2 OVERVIEW OF METHODS

Reviewers used a number of evaluation and data gathering techniques over a 4-month period, including:

- Focus groups
- Interviews
- Document analysis
- Server log analysis
- Usability testing
- Surveys and group discussions

Analysis focused primarily on the DOEd home site. The Office of Elementary and Secondary Education (OESE) Web site was also evaluated to demonstrate assessment techniques that could be applied to other offices’ Web sites.

3.3 INTERVIEWS AND SURVEYS

In-person and telephone interviews and surveys were conducted with key stakeholders. Members of the study team also conducted on-site interviews with DOEd personnel to gather user data, learn about the Web site’s management structure, and identify relationships between the Web site and other channels for information dissemination.

Interviewers used two discussion guides. One was targeted toward identifying user characteristics, and the other toward identifying management issues. Reviewers decided which guide to use based on interviewees' job responsibilities. These on-site interviews provided the research team with a clear picture of the DOEd Web site management issues and organizational relationships.

A self-administered survey was also disseminated during an Internet Working Group meeting. Meeting participants were encouraged to complete the survey during the meeting and take a copy to distribute among their units. Findings suggested that participants had extensive background with DOEd but little in Web evaluation or development. Most respondents felt strongly that they did not possess the requisite technical skills and knowledge to support their Web development or management efforts.

3.4 DOCUMENT ANALYSIS

During the research team's analysis of DOEd's policy system, the team analyzed government Web policy documents to identify information policies that could effect Web site management and assessment. A few representative policies were:

- The Freedom of Information Act (FOIA) and the Electronic Freedom of Information Act
- The Privacy Act and the Computer Marching and Privacy Act
- The Computer Security Act
- The Government Printing Office's Depository Library Program and Federal printing laws
- The Copyright Act

3.5 TRANSACTION LOG ANALYSIS

The study team used log file analysis to determine Web usage data and identify issues like broken links, coding problems, user paths through site content, and consistency of design practices. WebTrends Log Analyzer was used to analyze server log files.

WebTrends found that the site was accessed more than 3.5 million times per month. It also found thousands of instances where the DOEd Web site linked to unavailable resources ("404 errors"). Numerous pages were identified with HTML coding errors (for example, incorrect height and width attributes of the IMG SRC element). Large images (more than 90KB) were used in several cases; the Web industry standard is to keep Web page size less than 60KB to minimize download times. Team members also used WebTrends reports to track user paths through the site by identifying page requests.

3.6 USER ASSESSMENT

The evaluation team used a varied approach to user assessment that included interviews with customer service personnel, task-based tests with novice users, task-based tests with expert users, user evaluation of interface elements and buttons, and an evaluation by interface experts.

3.6.1 Interviews

In interviews with customer service personnel, the study team gathered data on perceptions of users of DOEd resources, users' information needs, key information resources, Web site usability problems, and user expectations. The team also conducted on-site field interviews with employees; the interviews lasted from 40 to 60 minutes and focused on personnel's understanding of users, user needs, and user problems.

3.6.2 Task-Based Assessment

For a task-based assessment, the team established four tasks to structure users' exploration of the DOEd Web site. Participants were allowed to proceed through the tasks at their own pace, provided they completed the evaluation in 50 minutes. After the time limit elapsed, users were asked to complete a questionnaire on their perceptions. A group debriefing followed.

Web site exploration was conducted in two group sessions—one for novice users, and one for expert users. All participants had bachelor or master's degrees in education. All participants had some experience using the Internet.

3.6.3 User Evaluation of buttons

Evaluators created a testing method to illustrate how well users understood patterns and groups of links provided by home page help buttons. Seven novice participants were recruited. All had some computer and Internet experience and all were educators.

Evaluation team members printed out copies of the pages linked from the eight left-margin buttons on the home page. Users were asked to read the contents of each page and type a reaction to each page, including their own description of the page contents, a reaction to the page including their emotional response, and questions raised. Participants were asked to edit printed pages by circling parts of each page they had difficulty understanding and crossing out content they felt was inappropriate.

3.6.4 Expert Evaluation

The study team recruited four experienced Web site designers to conduct expert evaluations of the OESE Web site. Evaluators were given sets of instructions with suggested criteria for evaluation. Criteria were:

- **Orientation:** overview, scope, liability, copyright
- **Design:** esthetics, consistency, appropriateness
- **Navigation:** learnability, clarity, memorability
- **Quality:** updated links, timely information, complete pages
- **Customer service:** easy-to-use feedback mechanisms, usable help features

Evaluators examined the OESE site simultaneously for 1 hour and then shared their experiences and opinions in a group discussion.