Usability Testing of an Academic Library Web Site: A Case Study

by Brenda Battleson, Austin Booth, and Jane Weintrop

Usability testing is an invaluable tool for evaluating the effectiveness and ease of use of academic library Web sites. This article reviews major usability principles and explores the application of formal usability testing to an existing site at the University at Buffalo libraries.

Brenda Battleson is Technical Services Librarian, University at Buffalo, Acquisitions Department, 134 Lockwood Library, Buffalo, New York 14260 <blb@acsu.buffalo.edu>; Austin Booth is Humanities Librarian, University at Buffalo, Reference Department, 521 Lockwood Library, Buffalo, New York 14260 <habooth@acsu.buffalo.edu>; Jane Weintrop is Reference Librarian, University at Buffalo, Business and Government Documents Department, Lockwood Library, Buffalo, New York 14260 <ulcjanew@acsu.buffalo.edu>.

Library Web sites are evolving into information gateways, unlocking access to library resources and services as well as electronic indexes and databases, primary research materials, and the Internet at large. Although intended to ease the process of information access, the staggering amount of information available via these sites can produce a kind of "information overload" that can bewilder, confuse, and even discourage users. There is a fundamental need for "usability" in library Web sites and usability testing is an invaluable tool for evaluating interfaces in terms of their effectiveness and ease of use. Designed and implemented properly, usability testing becomes an integral part of the Web site's development and evolution. This article reviews the concept of usability and explores the application of usability testing to library Web sites through a case study of usability testing at the University at Buffalo libraries.

Background on Usability Engineering and Human-Computer Interaction

Despite the explosion of information technology and a growing dependence on computers in all facets of society, only recently have "user needs" become part of software and interface development. During the 1990s software companies began to address their customer needs seriously and to design "usability" into their products rather than focusing solely on functionality.[1] Usability engineering involves studying and designing "ease of use" into a product. Its major component, Human-Computer Interaction (HCI), the study of how people interact with computer technology and how to make this interaction effective, provides the theoretical basis for applying usability concepts to software applications and computer interfaces.

Web sites are as well suited as software to the precepts of HCI; usability engineering can, therefore, impart a systemic approach to Web design. HCI dictates that interfaces should meet the following goals: (1) *provide task support*, that is, enable users to achieve their goals and meet their particular needs; (2) *be usable* by making it possible for users to work easily, efficiently and with few errors; and (3) provide an aesthetically *pleasant* interface design. A usable or "user-centered" interface is one that effectively meets these goals.

Task support is crucial to user-centered interface design. Elements of task support include knowing who a library's users are, what they wish to accomplish using the site, and what support they need to complete their tasks successfully. First and foremost, an interface's primary users must be identified. When a Web site caters to a variety of users performing different tasks, the needs of the defined primary user group take priority. Designers must ask, "Who are our users and what tasks do we want them to be able to accomplish using this site?" Clearly defined priorities in terms of the "who" and "what" of a Web site are the bases for assessing whether or not the site provides sufficient task support.

"Clearly defined priorities in terms of the 'who' and 'what' of a Web site are the bases for assessing whether or not the site provides sufficient task support."

The International Standards Organization (ISO) defines *usability* as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."[2] Applied to Web technology, this means simply that, for the tasks it is designed to support, an interface must be easy to learn, remember, and use, and must lead to few errors. The most effective means of assessing a site's usability is with usability testing, which will be discussed in more detail later in this article.

To a somewhat lesser degree, HCI conveys the need for aesthetics in interface design. Given the graphics-intensive nature of the present information environment, especially with regard to the World Wide Web, it is important that aesthetics be considered in usability testing since they can affect user satisfaction, as well as the ways in which users navigate a given interface. Aesthetics are, indeed, listed among the goals of HCI, but the impact of aesthetic variables on usability is extremely difficult to measure. While a formal mechanism for evaluating aesthetics was not designed into the case study described below, the comments of the test participants regarding the interface's visual appeal were certainly a factor in the overall evaluation of the site's usability.

Literature Review

Usability Engineering and HCI

There are numerous guides and handbooks available concerning the design and maintenance of Web sites. While these publications vary in degree and sophistication they tend to discuss the how-to's of Web site design solely from the site developer perspective. Few of these handbooks discuss usability engineering, HCI, or the concept of testing to see if users' needs are being met. This absence is not surprising since very few of these publications even acknowledge user needs as a factor in the Web site development process.

Jakob Nielsen's *Usability Engineering* is the most comprehensive and practical discussion of usability engineering and testing, covering the usability engineering life cycle from product conceptualization to design and evaluation.[3] Usability testing, a component of usability engineering involving the testing of an interface/system to determine whether or not it meets the precepts of HCI, is specifically targeted in Jeffrey Rubin's *The Handbook of Usability Testing* and Joseph Dumas and Janice Redish's *A Practical Guide to Usability Testing*.[4] Both of these books outline detailed guidelines for conducting usability tests, but *The Handbook of Usability Testing* provides especially useful information about steps to take before undertaking usability testing, including developing a purpose statement, task list, and test plan.

While these books give useful overviews of usability engineering and testing, they focus primarily on hardware and software development. Applying HCI and usability testing to the graphical user interface (GUI) has been a relatively recent concept, spurred on by the popularity of the Web. Nielsen addresses this in *Designing Web Usability*, as well as in a series of *Useit.com: Usable Information Technology* "Alertbox" articles.[5] Jared Spool's book, *Web Site Usability* documents a test of nine popular Web sites for usability.[6] Spool's work is of interest to library Web site developers since it was one of the first studies to evaluate the usability of sites designed, at least in part, to support information retrieval. A useful and informative source for applying usability testing to Web-based interfaces is Alison Head's *Design Wise*, which describes the application of task support, usability, and aesthetics to Web site design as well as the importance of integrating these HCI concepts into the evaluation of existing sites.[7]

Usability Testing

Usability testing can be divided into three categories: *inquiry*, *inspection*, and *formal usability testing*. While the first and last involve real users, the second does not. (The term "usability testing" encompasses numerous methods of evaluating site usability. This general term should not be confused with "formal usability testing" which, as described below, is a specific usability test method.) The most comprehensive and practical outline of usability test methodology is James Hom's *The Usability Methods Toolbox*.[8]

In the context of Web site and interface assessment, inquiry involves requesting information about a particular site from the users. Methods of inquiry include focus groups, interviews, questionnaires, and surveys. Interviews and focus groups are structured methods of inquiry which are used to gather information about users' experiences and preferences. While surveys and questionnaires may also be used to gather such information, interviews and focus groups allow for more interaction with the users and for immediate answers to questions raised during the interview or focus group. In addition, focus groups consist of multiple users whose interaction may raise additional issues of interest. Interviews and focus groups are generally conducted at early stages of product development, while surveys and questionnaires are generally used later in the product's life cycle.

With inspection methods like heuristic evaluation and cognitive walkthrough, a site's designers and information specialists serve as testers and subjects, often putting themselves in the place of the user to perform various tasks using the site. Unlike inquiry and formal usability testing, these forms of assessment do not enlist the participation of actual users. In cognitive walkthrough, experts attempt to accomplish typical user tasks with a given interface. Heuristic evaluation involves usability experts checking elements of an interface against a checklist of heuristics, or design principles.[9] While evaluations based on inspection are relatively inexpensive to conduct, they are less useful in identifying usability errors than tests with actual users.[10]

"In formal usability testing, users are observed using a site, or prototype, to perform given tasks or achieve a set of defined goals."

In formal usability testing, users are observed using a site, or prototype, to perform given tasks or achieve a set of defined goals. This method involves employing experiments to gather specific information about a design. Formal usability testing was first used in experimental psychology, and originally involved the gathering and analysis of large quantities of data. Today, however, it is more concerned with interpretation and rapid, useful results rather than amassing large bodies of quantitative data. Dumas and Redish discuss five facets of formal usability testing: (1) the goal is to improve the usability of the interface; (2) testers represent real users; (3) testers perform real tasks; (4) user behavior and commentary are observed and recorded; and (5) data are analyzed to recognize problems and suggest solutions.[11] Applied to Web

site interfaces, this test method not only results in a more usable site, but also allows the site design team to function more efficiently, since it replaces opinion with user-centered data.

In terms of test design and implementation, Rubin's *Handbook of Usability Testing* and Hom's *Usability Methods Toolbox* provide the most detailed discussion of the actual methodology of formal usability testing. A formal usability test typically involves introducing the interface, asking the user to attempt a set of tasks, observing the human-computer interaction that takes place and evaluating the results to identify design problems exposed by the interaction.

There is a consensus in the literature that usability testing be an iterative process, preferably one built into a Web site's initial design. HCI concepts lay out interface design considerations, while proper testing techniques advocate testing for these design concepts as part of a Web site's development process. Site developers should test for usability, redesign, and test again—these steps create a cycle for maintaining, evaluating and continually improving a site.[12]

Usability Testing in Libraries

Although libraries have a long history of studying and responding to user behavior through end-user studies, the relative lack of literature on the topic reveals that libraries are only beginning to apply usability testing to their Web sites. Karen Eliasen, Jill McKinstry, and Beth Mabel Fraser tested students' ability to navigate online menus to select correctly databases from the library home page at the University of Washington. This test used a relatively simple testing mechanism where participants were observed as they attempted tasks using three alternate versions of a low-fidelity prototype. The testers planned to do further testing using think-aloud protocol, timed-task analysis, and a post-test debriefing.[13] Janet Chisman, Karen Diller, and Sharon Walbridge conducted formal usability testing of the Washington State University Libraries' online public access catalog and indexes.[14] A recent article by Jerilyn Veldof, Michael Prasse, and Victoria Mills described case studies and outlined much useful information about usability testing in general. The testing methods described include heuristic and formal usability testing with paper and online prototypes at the University of Arizona and OCLC.[15]

The case study that follows is a description of the use of HCI concepts to design and implement formal usability testing of an academic library's Web site. This study is notable in that the testing procedures were extremely effective, utilizing techniques that were neither expensive nor technically difficult to implement. More importantly, this testing was applied to a Web-based environment typical of most libraries, that is, the Web site was not new, but had evolved over a number of years.

A Case Study of Formal Usability Testing in an Academic Library

At the onset of testing, in May 1999, the University at Buffalo Libraries Web site was well beyond the design stage. It had been launched four years prior and over the course of that period had expanded and changed dramatically in scope and purpose. Because this site was becoming an integral part of the library and university communities, the Web manager called for an "overall assessment" of the site as the initial step in addressing the need for design enhancements. To this end, a committee comprised of librarians working in different areas of the university libraries (including a library school student) was charged to develop a comprehensive evaluation program that would apply some method of testing. This testing would allow for the identification of problem areas and provide the data upon which to base design changes.

Upon a literature review of testing methods and principles, it was decided that an "overall assessment of the library Web site" was far too broad in scope. The Web manager and committee developed a more refined statement; "Determining whether or not the libraries' users could effectively use the Web site to perform specific tasks" clarified the purpose for testing and dictated that the test method would be *formal usability testing*—watching real users perform real tasks.

Setting the Goals

To design and implement an effective test, the usability testing committee needed to apply HCI design concepts to the library Web site. A framework was developed based on the key elements of *task support*: who are the users; what must they accomplish; and what support should the site provide?[16]

Users of a library Web site typically comprise a very heterogeneous population. The University of Buffalo Libraries centralized Web site services a large comprehensive public university with a student population of approximately 23,500 enrolled in a full range of undergraduate, graduate, and professional programs. Besides students, users of the libraries' Web site include faculty and staff, as well as members of the general public. Because a single usability test measuring task support for all of these user groups was neither practical nor feasible, the testing committee had to target the primary users of the libraries' Web site. Ideally, this user group would have been defined during the site's initial design stages. In fact, if there is no clear statement defining a target group, the Web manager, library administrators, or whatever decision-making body administers the Web site must identify the site's primary users before any usability testing can proceed. In this case study, the Web manager defined the primary users as "undergraduate students with little or no experience using the libraries' site."

The libraries' Web site accommodates numerous tasks. Among other things, undergraduates can search the library catalog; access course readings and numerous electronic indexes; ask reference questions via e-mail; renew books; and look up hours of operation for the various campus libraries. For the purposes of this study, the committee and Web manager focused on tasks that would be key to evaluating whether or not the site was "working for the users." What primary tasks could undergraduates expect to accomplish using the site and what kind of support must the interface provide so that they might effectively complete those tasks? To ensure a user-centered approach, site functionality was defined in terms of what the *user* needed to do, rather than all of the possible tasks the site could support, and "conducting library research" was defined as the most important activity for which the Web site should provide this task support.

Thus, the goal of the usability test was set: to determine how effectively the libraries' Web site "worked" when used for library research by undergraduates with little or no experience using the site. To further address this, the process of library research was broken into the broad tasks that typical undergraduates would be expected to complete:

- Using the site to identify an item/title that is part of the Libraries' collections;

- Using the site to locate the most appropriate resource for finding journal articles on a specific topic; and
- Using the site to find an appropriate starting point for researching a topic without necessarily knowing the format or sources of information.

Designing the Test

With the goal clearly stated and the test method identified, specific test questions representative of the broad tasks outlined above could be developed. Again, formal usability testing involves observing users as they perform given tasks or achieve a set of defined goals using a site. The challenge lay in presenting these tasks to the users in a way in which testers could learn from their responses and subsequent actions.

The number of test questions developed was limited to 11. In researching other usability tests,[17] the committee found that most were comprised of only 10 to 12 questions. This number was deemed sufficient since properly designed questions would provide the information needed without becoming tiresome to the students being tested. The committee sought questions that would be representative of the broader tasks outlined above and still reflect *real* questions that an undergraduate might be faced with. To reduce potential tester bias, the work of devising the questions was spread among the members of the committee, with each suggesting a number of test questions addressing each task. All questions were then reviewed, selected and edited by the committee as a whole.

Developing the questions presented a significant challenge. They had to be worded in a way that would test the site's design rather than the student's library skills. Also, there was concern that the multi-linked, cross-referenced structure of Web design would provide too many options as students addressed the test's assigned tasks, which would, in turn, complicate attempts to measure and identify the site's major usability problems. Thus, for this initial test, questions with only one or two "correct" answers were selected.[18] Developing such questions was not an easy chore, given the size and scope of the libraries' Web site and its multiple internal and external links. Committee members tested each of the questions, trying every possible approach and identifying as many ways as possible to find each "correct" answer. Questions with too many potentially "correct" answers were thrown out while those with more readily apparent "appropriate" and "inappropriate" approaches were retained. By quickly determining whether or not a student could effectively navigate the site to complete a task, the committee was able to reduce not only the amount of time spent on each question but also the potential for fatigue and confusion.

Finally, the committee considered the order in which the test questions would be presented. To determine whether or not the site was easy to learn and easy to remember, a number of questions were designed to build on tasks learned earlier in the test. The questions as they were presented were:

1. Do the University Libraries own the book *Alias Grace*?
2. Is the journal *Northeast Anthropology* available in the UB Libraries?
3. Can you find a journal article on gospel music?
4. How would you find a journal article on soap operas?
5. I am interested in investing in what are referred to as "callable securities." Where can I find recent articles about them? (You need to see the full article.)
6. Use the database *SocioFile* to look for an article about nursing homes and mental illness. What would you have to do to obtain this article?
7. How would you go about finding the author of this quotation: "And we meet, with champagne and a chicken, at last?"
8. Assume you are taking a class in a subject you know very little about, for example, psychology, literature, environmental science, architecture or film. If you were assigned a research paper, how would you find information resources on that subject?
9. Find an encyclopedia article about French wine.
10. Obtain background information about the "ethnic cleansing" that has taken place in Kosovo.
11. Does the Libraries' Web site have a guide to doing research in computer science?

The same questions are organized below in terms of the broad tasks they were designed to test. Each question is followed by the user action deemed most appropriate for successfully completing the task:

Task 1: Identify an item/title that is part of the Libraries' collections;

- Question 1 → link to library catalog for a given book
- Question 2 → link to library catalog for a given journal
- Question 6 → link to library catalog for a journal title given in a citation.

Task 2: Locate the most appropriate resource for finding journal articles on a topic;

- Question 3 → link to subject-specific or general index to full-text articles
- Question 4 → link to a general index to full-text articles
- Question 5 → link to a subject-specific index to full-text articles
- Question 6 → link to a title index to citations

Task 3: Find an appropriate starting point for research on a topic;

- Question 7 → link to "Reference Sources" [a listing of electronic reference resources organized by subject]
- Question 8 → link to "Quick Start" page [a listing of basic tools designed for novice researchers.]
- Question 9 → link to "Reference Sources" or "Quick Start"
- Question 10 → link to "Quick Start"
- Question 11 → link to "Need More?" [a series of subject specific pathfinders.]

After the set of test questions was finalized, the committee devised a method of presenting the questions to users and compiling the test results that would not only give insight into how effectively the Web site actually worked, but that was also within the committee's budget and level of expertise. Because the test emphasized identifying gross usability problems rather than making a detailed analysis of the site, videotaping, eye tracking, and other elaborate techniques for gathering data were deemed unnecessary. Instead, the committee utilized "think-aloud protocol," which involves the user articu-

lating his/her thought processes and opinions while directly interacting with technology to accomplish a given task.[19] Nielsen states that think-aloud protocol "may be the single most valuable usability engineering method. . . . One gets a very direct understanding of what parts of the [interface/user] dialog cause the most problems, because the thinking aloud method shows how users interpret each individual interface item."[20] Given the scope of the University at Buffalo Libraries study and the wide support of this technique in the literature, think-aloud protocol was the obvious choice.

Keith Instone refers to formal usability testing simply as "watch and learn."[21] The "watching" phase involves direct observation of the users being tested. In the study two testers observed each participant, but only one, the moderator, interacted with the student. The other, the scribe, kept a written log of what was taking place during the test. To promote consistency, the moderator worked from a script that suggested what and how much to say to the student. The script contained an introduction, during which time the moderator obtained consent and other information about the student (experience with the site, year enrolled, etc.). As the test procedure was outlined, the moderator tried to impress upon students that the committee was testing the Web site, not them. It was also explained that there were no "right" or "wrong" answers and that students would need to describe aloud what they were doing and why.

For each question, the script stipulated the ideal response, the minimum requirements for a satisfactory response, when to intervene and prompt the student to move on to the next question and when to end the test. For example, in Question 1, "Do the libraries own the book *Alias Grace*?" the script stated that the ideal response would be for the user to select "Libraries Catalog" from the main screen and then to search for the title. The moderator was instructed to end the test when the user reached the "results screen" in the catalog. The script indicated how many attempts the student should make at successfully navigating from one page to the next. It also described when to let the subjects depend entirely on the site for direction and what instructions to give when intervention was necessary. Again, with Question 1, the moderator was directed to intervene "when the user makes three incorrect selections from the main screen without reaching the catalog." In this case the moderator was to clarify the term "Library Catalog" if needed and then observe the results. Although the script was not designed to be used verbatim, it did suggest the wording of appropriate interaction, such as "what do you understand by the phrase. . . " or "I have learned enough from this question—let's move on to the next."

The scribe recorded the user's movement through the Web site as well as his/her comments. A pre-printed log specific to each question aided this process. The design of the log enabled the scribe to check off quickly the sequence of selections made by the user and still have time to write down any additional observations or comments.

Upon completing the actual test, students were asked to fill out a post-test evaluation asking for comments and opinions on both the Web site and the test. As soon as each student left the room, testers independently completed their own post-test evaluations, for which they reflected on what they had observed and commented on the obvious problems of the Web site, as well as features that appeared to work well. These comments proved to be an integral part of the "learning" phase of the watch and learn process.

"With careful planning and design, implementing the tests were inexpensive, straightforward, and informative."

Conducting the Test

With careful planning and design, implementing the tests were inexpensive, straightforward, and informative. Eleven tests were conducted, which is more than Nielsen advocates as necessary when testing a homogeneous group of users. (This is discussed below in depth.) To keep the time commitment to a minimum, each pair of testers was drawn from a pool of six committee members. Not only did this distribute the workload, but also reduced tester bias by allowing seven different people to observe the users and analyze the usability of the site.

Testing required the use of a standard personal computer, with a Web browser and Internet access. The only limitation on the setting was that it be private enough so that the students would be comfortable talking aloud. Testers thought the privacy of an office would be intimidating and instead, chose a hands-on classroom. Volunteers were solicited by instructors of a few sections of undergraduate English 101 classes, who handed out fliers and encouraged participation. Although individual tests were conducted by appointment, it should be noted that despite agreed-upon meeting times, no-shows were inevitable. To further encourage participation and to help convey the importance of the test, participants were therefore compensated with gift certificates to a local music/video/book store.

Conducting the test was a relatively straightforward exercise. The committee tested the study's design by running two mock-tests using volunteers. This "testing the test" was well worth the effort because it resulted in the addition of instructions that greatly improved the effectiveness of the assessment process for both students and testers. For instance, results of the mock-test suggested that the students be instructed to read each question out loud, an exercise that actually encouraged the "think aloud" process. In addition, testers discovered that, without specifically instructing students to return to the home page before beginning each question, they would start navigating from wherever they were on the Web site, thus masking what we were trying to learn. The lack of this simple instruction, "click 'home' before beginning the next question," would have seriously skewed the final test results.

Evaluating the Results

Just as the HCI element of "task support" served as the framework for designing the test, "measures of usability" were used to evaluate the results. Did students find that the Web site was easy to learn, easy to remember, pleasant to use, and caused few errors?[22] Much of this analysis was, of course, based on comments made by users during the "think aloud process." However, as each student attempted to perform specific tasks using the site, testers based additional assessment on a number of observations: how quickly did the user work? Did choices seem obvious? How carefully did students read the information on the screens? Was there an increasing level of certainty and success as the test progressed?

Testers carefully noted body language, facial expressions, and reaction times in addition to recording each user's comments. Both the scribe and the moderator, in their immediate post-test evaluations, assessed all observations and comments noted in the test log. These evaluations were then collected and reviewed by the larger committee.

Assessing the site with regard to the first two usability measures—is it easy to learn and easy to remember?—was heavily based on observation and proved difficult to measure in certain respects. In a few instances, as students performed the assigned tasks, there was not always a clear distinction between a response resulting from "learning" and one that was simply a "guess." Also, even though a homogeneous user group was tested, individual students obviously had different levels of technical proficiency and varying learning abilities. Yet, the nature of the study allowed for these individual differences while quickly revealing the major usability problems with the site.

Ascertaining whether or not students found the site "pleasant to use" was challenging given the subjective nature of such variables as aesthetics and the idiosyncrasies of individual users. The ideal method of measuring this element would have been to ask each student to complete a post-test questionnaire designed to measure satisfaction with the site. Because the committee was as interested in the application of usability testing to the site as in the actual usability of the site, the committee chose instead to design the post-test survey to measure satisfaction with the test *process*. Fortunately, tester observations and user comments compiled during the test process were more than sufficient in revealing the extent of user satisfaction with the site. Of course, not all users are as "talkative" or "expressive" as the ones we tested. A post-test questionnaire designed to measure satisfaction would have been useful and will be incorporated into future test design.

Assessing the fourth usability measure—"does the site cause errors?—was a much more objective process, since for each question, there were pre-defined "ideal" and "satisfactory" responses against which testers could measure performance. Using the pre-printed log, testers tracked a user's navigation through the site and tallied the number of attempts made when moving from one page to the next. These quantifiable measures of outcome in addition to observations relating the ease with which each student worked, provided more than just anecdotal evidence of usability. Because the students were encouraged to talk aloud as they were working, testers saw not only what they did and how easily they worked, but learned why they made the specific choices they did. Testers developed an almost immediate understanding of what site features were ineffective as well as why those features did not work. As testing progressed and different students encountered the same problem areas, the committee was confident that these usability tests would successfully identify the major usability problems of the site.

Test Results

Main Screen

Formal usability testing focused on those main screen links directly related to library research:

- Online Resources (links to electronic information products to which the Library subscribes);
- UB Libraries Catalog (links to the online public access catalog);
- Web Search (links to various aids for identifying information on the site and on the Internet); and
- Need Help (links to guides for starting research).

Emphasis was placed on assessing the students' experiences with the "UB Libraries Catalog" and "Online Resources" links, since "Web Search" and "Need Help" were viewed as secondary tools for supporting library research. However, the test revealed surprising data on the latter two links. From the main screen (see Figure 1), the "Web Search" and "Need Help" links both failed the "easy to learn" criterion, creating enough confusion with users that they were rarely selected. According to comments, students erroneously assumed that the "Web Search" link led only to Web search engines and the Internet, when, in fact, it included links to site-specific search features designed to help users. Also, while there was little direct feedback as to why "Need Help" was not regularly selected, it was apparent that when this link was chosen, students were not at all satisfied with the information obtained. Testing clearly revealed a need to address usability problems of the main screen with regard to these two links.

Students easily identified the link for the "Libraries Catalog" and for tasks that involved identifying an item in the library's collection, the main page met usability criteria. For the non-catalog research tasks (finding journal articles, researching a topic), the ideal choice was the "Online Resources" link, yet most students initially selected "Libraries Catalog." There was obvious confusion with terminology[23] as well as a clear misunderstanding of what the term "Online Resources" implied, so the screen initially failed the "causes few errors" measure. Students quickly realized the error and upon returning to the main screen, usually selected "Online Resources" as the appropriate choice. User comments indicated that, although they did not clearly understand the term "Online Resources," most students selected it on their second attempts because they were able to eliminate the other choices as obviously inappropriate. In short, they learned that "Libraries Catalog" was inappropriate and used process of elimination to select "Online Resources." Once they understood the distinction between "Libraries Catalog" and "Online Resources," the main screen was easy to use, easy to remember and few errors were made.

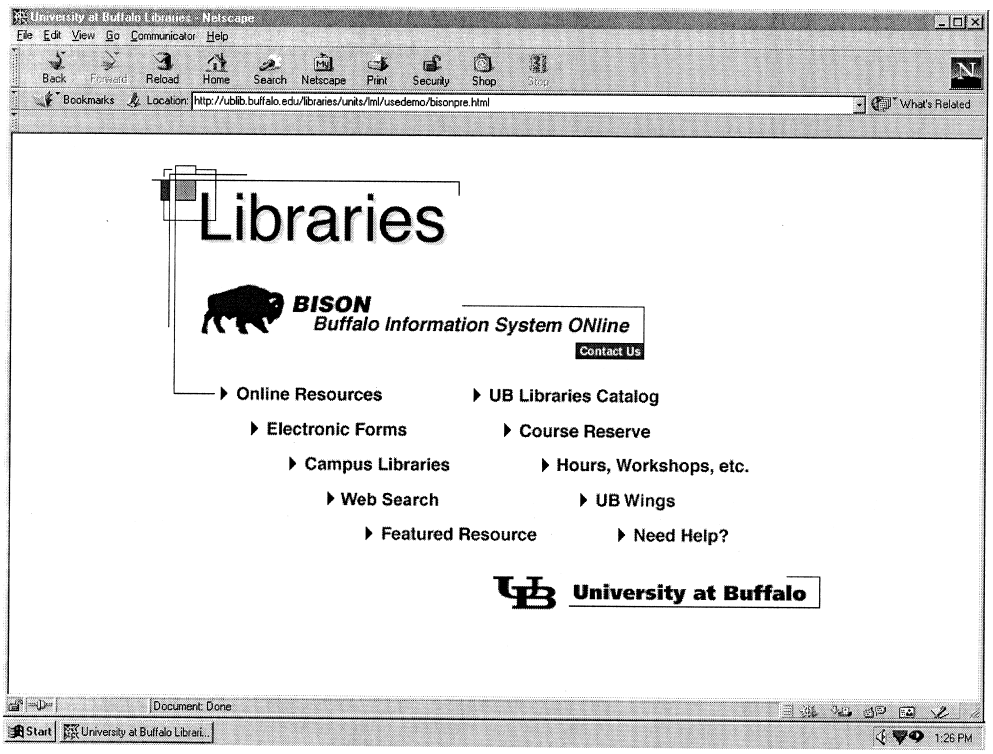
Second-level Screens

The "Libraries Catalog" link leads to a secondary page listing three options for connecting to the online catalog: one link allows users to connect to a Web-based graphical user interface, whereas the other two links provide access to the text version (see Figure 2). While users wishing to access the catalog were successful in connecting via the preferred Web-interface link, the process seemed unnecessarily unpleasant. The text-heavy presentation of the connection options resulted in confusion and hesitation among the users and certainly detracted from the usability of the page.

Students may have had difficulty in selecting the "Online Resources" link from the main screen, but they did find it to be an appropriate choice once they accessed the actual page (see Figure 3). Deciphering this page was another matter entirely, however. This page was very usable in some areas with regard to library research tasks, but it proved to be surprisingly problematic in others.

When looking for journal articles on a topic that could be easily associated with an academic discipline (test examples were gospel music and callable securi-

Figure 1
Main Page of the UB Libraries Web Site



ties), the "Online Resources" page met all usability criteria. Students easily identified the "Databases by Subject" link as the correct choice and used it repeatedly to successfully complete a number of tasks. However, in the Question 6, which required that users find an article from the named database *SocioFile*, they did not find the "Online Resources" page easy to use since most were unable to readily identify the most appropriate choice, "Databases by Title." Interestingly, students who did find the link to *SocioFile* often did so through indirect, often "creative" routes.

For finding articles on general topics (e.g., soap operas), the "Online Resources" page failed all usability criteria. Students did not identify "Quick Start," which was the only link that would have easily led them to such articles. Because they were unwilling to select this link, even as a second or third choice, students failed to learn that this link was useful. Indeed, observations of facial expressions confirmed that, when attempting to research a general topic, the site was not very "pleasant to use."

The "Online Resources" page was not usable as a starting point for research. In addition to the users' unwillingness to select "Quick Start," they did not readily identify "Reference Resources" as a logical link to begin exploration of a topic. Although not always appropriate, "Databases by Subject" was selected most often, while virtually every other link on the "Online Resources" screen was ignored.

Usability Testing: What We Learned

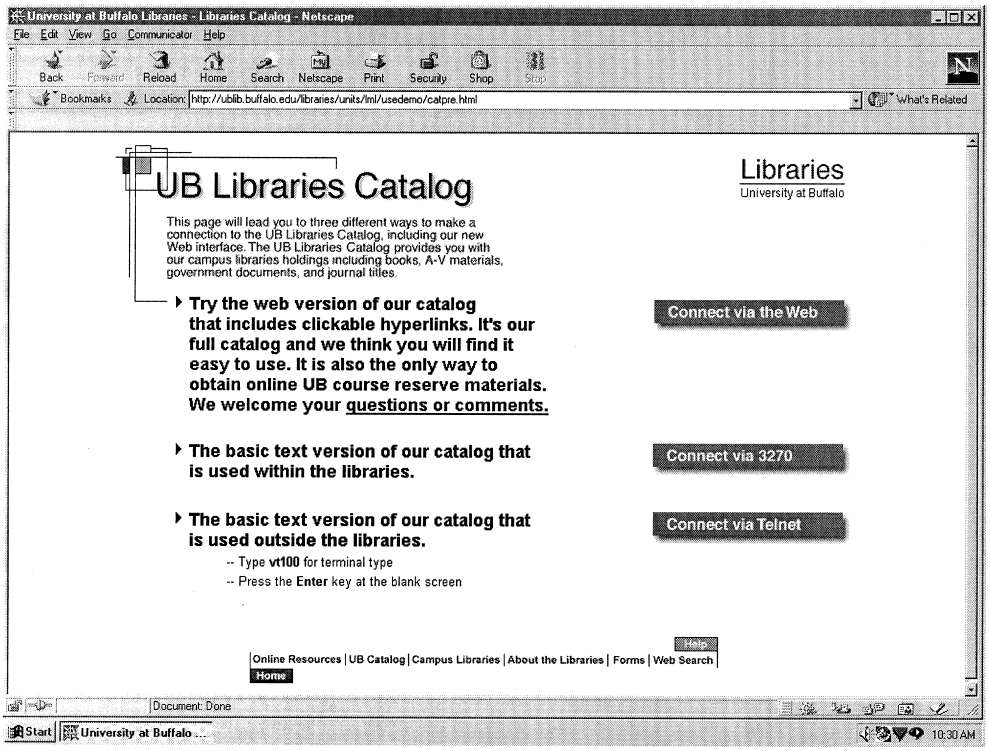
A great deal of time, effort and money had been invested in the "Online Resources" region of the library's Web site. That testing revealed a relative ineffectiveness of the presentation of information in this area was significant and surprising. Overall, it is fair to state that the usability problems discussed above would have not been considered, much less identified, had formal usability testing of the Web site not been undertaken. Testers' observations and the comments of the students participating in the test were invaluable in revealing where and why the site failed and helped evaluators to identify and prioritize the gross usability problems to be addressed. It is not an exaggeration to say that all parties involved were as impressed by the effectiveness of usability testing as they were by the results.

Although the usability test supported the literature with regard to test design and implementation, the committee was surprised to find that the test experience also supported the use of a relatively small test group. The committee had initially been skeptical of Nielsen's idea that when dealing with a homogeneous user group, watching as few as five users can identify a high percentage of the most critical errors:

> The most striking truth. . . is that zero users give zero insights. . . . As soon as you collect data from a single test user, your insights shoot up and you have already learned almost a third of all there is to know about the usability of the design. . . . As you add more and more users, you learn less and less because you will keep seeing the same things again and again. . . . After the fifth user, you are wasting your time by observing the same findings repeatedly but not learning much new."[24]

Nielsen outlines a mathematical formula developed with Tom Landauer, which supports his use of relatively small usability test groups:

Figure 2
"UB Libraries Catalog" Page



> . . . the number of usability problems found in a usability test with *n* users is *N(1-(1-L)n)*, where *N* is the total number of usability problems in the design and *L* is the proportion of usability problems discovered while testing a single user. The typical value of *L* is 31%, averaged across a large number of projects we studied. Plotting the curve for *L*=31% gives the following result:[25]

As Nielsen stresses, this formula is only valid with a homogeneous group of users: "You need to test additional users when a Website has several highly distinct groups of users. The formula only holds for comparable users who will be using the site in fairly similar ways."[26] Even with a homogeneous test group, however, Nielsen further challenges his own formula:

> "The curve (see Figure 4) clearly shows that you need to test with at least 15 users to discover all the usability problems in the design. So why do I recommend testing with a much smaller number of users?
> The main reason is that it is better to distribute your budget for user testing across many small tests instead of blowing everything on a single, elaborate study. . . . "[27]

As novice, cautious planners, the committee tested 11 users only to find that Nielsen was indeed correct. The students in our test group committed the same mistakes, with the most obvious errors and major usability problems apparent after the first few tests. Nielsen states that "[e]laborate usability tests are a waste of resources. The best results come from testing no more than 5 users and running as many small tests as you can afford."[28] Because there was very little new information obtained from our last few tests, they were essentially unnecessary and possibly wasteful, since the time and money invested may have been better spent on assessing the design changes we hoped to recommend as a result of this test.

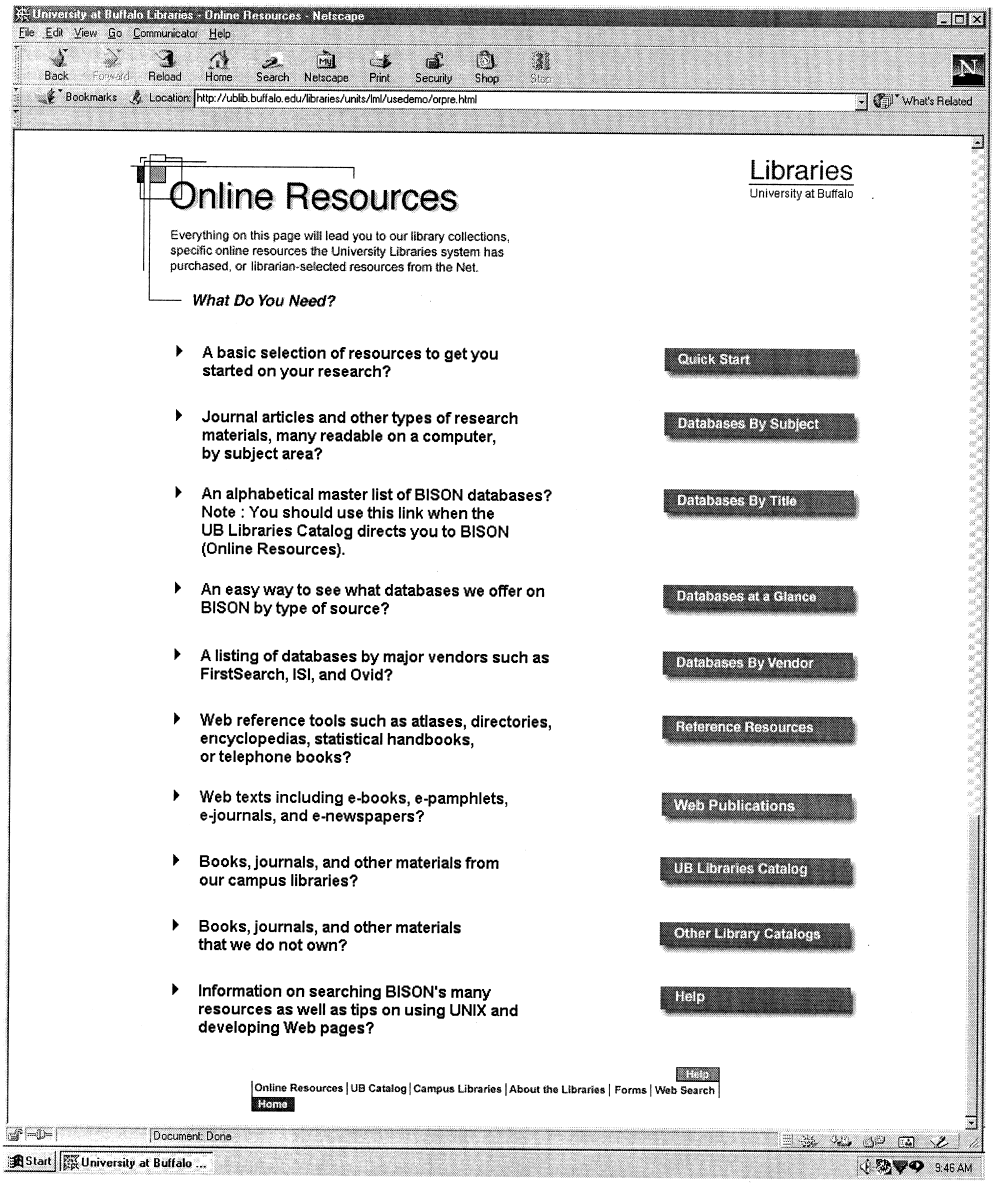
"The test process revealed a great deal about the site without being terribly complicated or expensive."

The test process revealed a great deal about the site without being terribly complicated or expensive. Careful test design was crucial however. In addition to the scripting instructions, the wording and order of the questions were key to obtaining accurate and useful data about the site as well as contributing to the efficient implementation of the test. While administering the test was labor intensive, even with only 11 test subjects, breaking up the duties among numerous testers helped to minimized the work of each and provided numerous opportunities to observe different students using the site.

This test revealed the validity and usefulness of qualitative analysis in Web site evaluation. While the University at Buffalo Libraries study was not elaborate, it did not need to be, given the goal of the test—to determine whether or not undergraduate students could use the site effectively. This is not to say that more quantifiable data could not have been gathered using more sophisticated testing processes. Such techniques are very expensive, however, and it is doubtful that they would have been any more revealing.

The informal comments and opinions of the students tested were as enlightening as their test performances. For instance, asking users their overall impressions of the Web site as they were preparing to leave resulted in very useful opinions and observations on their part. Because they

Figure 3
"Online Resources" Page



assumed that the test had concluded, they seemed more relaxed and willing to give honest impressions regarding those difficult-to-measure factors like aesthetics. In the future, the usability committee plans to build these questions into the test.

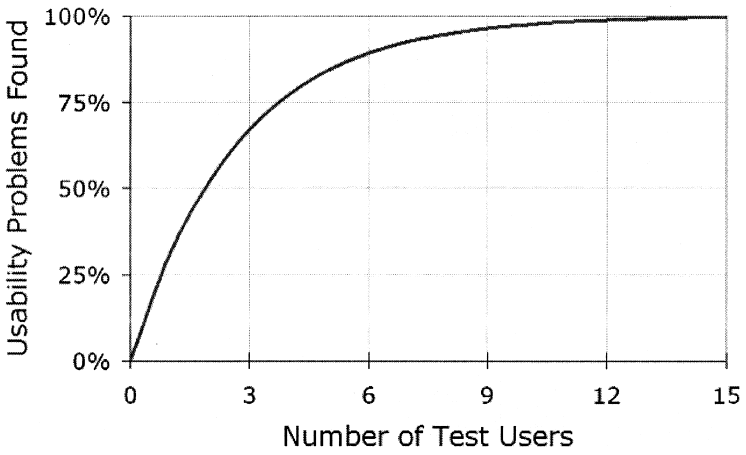
Finally, this study reiterated the fact that usability testing be a continuous and integral part of Web site development. Initial testing is only the beginning. "You want to run multiple tests because the real goal of usability engineering is to improve the design and not just to document its weaknesses."[29] Results of testing lead to recommended changes, which will, in turn, need to be evaluated, implemented and tested again. As Nielsen states, usability testing is an *iterative* process.[30]

Conclusion

Web sites have become an integral part of the information access mission of academic libraries. Although developing a usable and effective site is challenging in its own right, maintaining and redesigning that site to meet the constantly changing needs of users is a seemingly impossible task. Usability testing is perfectly geared to this task since it not only assists in identifying interface problems, but also in developing ways in which to attack those problems.

The data compiled as a result of the usability test described above were justification for changes recently made in the overall design of the University at Buffalo Libraries' Web site. While formal usability testing was the method of choice in this case, it is not the only means of evaluating interfaces. Indeed, this test was supplemented by a card-sort (inquiry method) test to evaluate the terminology of the library's site and the changes recently implemented will be evaluated using an appropriate usability test method. As the site continues to grow and change, testing will no doubt, be an important part of its evolution.

Figure 4
Nielsen and Landauer's Curve Showing the Relationship between the Number of Users Tested and the Number of Problems Found in a Usability Test (From Nielsen, "Why You Only Need to Test with Five Users")



"The importance of usability testing and the applicability of usability testing to library Web sites cannot be understated."

The importance of usability testing and the applicability of usability testing to library Web sites cannot be understated. Whether through inspection, inquiry or formal usability testing, usability test methods serve as both catalysts for design changes and as tools for evaluating those changes, especially as library sites strive to meet the increasing information demands of users. When developing a new site or evaluating an existing one, librarians, site designers and administrators must ask the following questions: Does (or will) the site provide task support? Is it usable? Is the interface pleasing in terms of ease of use and aesthetics? These questions are necessary elements of successful and usable Web site design. However, these questions must be directed at real users, who are the key to successful usability testing. The case study above illustrates that carefully planned usability testing is not difficult to implement and the returns in terms of the information revealed about the effectiveness and usability of a site are well worth the time, effort, and money invested.

Acknowledgment: The authors wish to express our appreciation to the additional members of the Usability Testing Committee for their time, insights and dedication to this study: Gayle Hardy, Lori Widzinski and Christopher Badurek. We also wish to thank the University Libraries Web Manager, Dr. Gemma DeVinney, who charged the committee and provided the administrative support that made this study possible.

Notes and References

1. Jakob Nielsen uses increasing corporate budget allocations for usability engineering as well as "the increasing number of personal computer trade press magazines that include usability measures in their reviews," as evidence of this increase. See Jakob Nielsen, "Usability Laboratories: A 1994 Survey," *Useit.com: Usable Information Technology*, (n.d.) available online: http://www.useit.com/papers/uselabs.html (accessed January 6, 2000).
2. ISO DIS 9241-11, *Ergonomic Requirements for Office Work with Visual Display Terminals. Part 11: Guidance on Usability* (London: International Standards Organization, 1994), p. 10.
3. Jakob Nielsen, *Usability Engineering* (Boston: Academic Press, 1993). Other useful books and Web resources on HCI and usability engineering in general include: Keith Instone, *Usable Web*, available online at http://usableWeb.com/(accessed July 15, 2000); Ben Shneiderman, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3rd ed. (Reading, MA: Addison Wesley Longman, 1998); Bryce L. Allen, *Information Task: Toward a User-Centered Approach to Information Systems* (San Diego, CA: Academic Press, 1996).
4. Jeffrey Rubin, *The Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests* (New York: Wiley, 1994); Joseph C. Dumas & Janice C. Redish, *A Practical Guide to Usability Testing* (Norwood, NJ: Ablex Publishing Co., 1993). Since our testing, a 1999 edition of this title has been published by Intellect, Exeter.
5. Jakob Nielsen, *Designing Web Usability: The Practice of Simplicity* (Indianapolis, IN: New Riders Publishing, 1999); Jakob Nielsen, "Alertbox," column in *Useit.com: Usable Information Technology*, available online: http://www.useit.com/alertbox/(accessed January 6, 2000).
6. Jared Spool, *Web Site Usability: A Designers' Guide* (North Andover, MA: User Interface Engineering, 1997). This book documents a test of nine popular Web sites for usability. A 1999 edition of this title has been published by Morgan Kaufmann Publishers, San Francisco.
7. Alison Head, *Design Wise: A Guide to Evaluating the Interface Design of Information Resources* (Medford, NJ: CyberAge Books, 1999).
8. James Hom, The Usability Methods Toolbox, available online: http://www.best.com/~jthom/usability (accessed January 6, 2000).
9. Nielsen lists usability principles that should be followed by all interface designers: (1) use simple and natural dialog; (2) speak the user's language; (3) ensure that instructions are easily visible or retrievable; (4) design consistency; (5) give user appropriate system feedback; (6) provide clearly marked exits; (7) provide shortcuts; (8) display easily interpreted error messages; (9) design to prevent errors; and (10) provide help and documentation (*Usability Engineering*, p. 20).
10. Useful sources for the discussion of these inspection methods include Jakob Nielsen & R. Mack, eds., *Usability Inspection Methods* (New York: Wiley, 1994), which introduces inspection methods, including heuristic evaluations and cognitive walkthroughs; and Keith Instone, "Usability Heuristics for the Web," *webreview.com*, available online: http://webreview.com/wr/pub/97/10/10/usability/sidebar.html (accessed January 6, 2000), which shows how Nielsen's list of heuristics can be adapted for the Web environment.
11. Dumas & Redish, *A Practical Guide to Usability Testing*, p. 22.
12. In addition to the sources cited above, there are numerous online periodicals and trade publications that discuss usability testing and it applicability to Web sites, The most useful are *webreview.com*, available at http://webreview.com and *Useit.com: Usable Information Technol-*

ogy, available at http://www.useit.com/. *Useit.com* includes Nielsen's "Alertbox" column that discusses various usability topics and well as practical methods by which to implement usability testing. Both publications devote significant space to the discussion of usability.

13. Karen Eliasen, Jill McKinstry, & Beth Mabel Fraser, "Navigating Online Menus: A Quantitative Experiment," *College & Research Libraries* 58 (November 1997): 509–516.
14. Janet Chisman, Karen Diller, & Sharon Walbridge, "Usability Testing: A Case Study," *College & Research Libraries* 60 (November 1999): 552–569.
15. Jerilyn Veldof, Michael Prasse, & Victoria Mills, "Chauffeured by the User: Usability in the Electronic Library," *Journal of Library Administration* 26 1999;115–140. It should be noted that more and more academic libraries have begun to apply usability testing to their Web sites since the writing of this article. When the University at Buffalo Libraries began to design a usability study in May 1999, only a handful of libraries had actually implemented some form of testing. In July 2000, however, a quick, superficial search on the terms "'usability testing' AND library" using the search engine *Google*SM, listed links to recent and ongoing usability testing of academic library Web sites at, for instance, the University of Washington, Indiana University-Bloomington, Yale, MIT, University of Nevada–Reno, Roger Williams University, and Washington State University.
16. Head, *Design Wise*, p. 48.
17. We modeled our test procedures after those used in a similar test conducted at the University of Arizona Library. See Michelle Clairmont, Ruth Dickstein, & Vicki Mills, "Living the Future 2: Testing For Usability in the Design of a New Information Gateway," available: http://www.library.arizona.edu/library/teams/access9798/lft2paper.htm (accessed January 6, 2000.) The information outlined in this paper was invaluable to us in that it conveyed the simplicity and effectiveness of conducting usability testing of a library Web site.
18. The student need not necessarily have answered the question to be considered "correct." For the purposes of this test, a "correct" response was one in which the user could navigate to the appropriate index, database, resource, etc. in which the question could then be answered.
19. For more on think-aloud protocol, see Dumas and Redish, *A Practical Guide to Usability Testing* and Rubin, *The Handbook of Usability Testing.*
20. Nielsen, *Usability Engineering*, p. 185.
21. Keith Instone, "User Test Your Web Site," *webreview.com*, available: http://webreview.com/97/04/25/usability/index.html (January 6, 2000.).
22. Head, *Design Wise*, p.53.
23. Students did not immediately recognize the term, "Online Resources" which suggests that terminology also be subjected to usability testing.
24. Nielsen, *Usability Engineering*, pp.172–174. See also Nielsen's "Why You Only Need to Test With 5 Users," *Useit.com: Usable Information Technology*, (March 19, 2000) available: http://www.useit.com/alertbox/20000319.html (January 6, 2000).
25. Jakob Nielsen, and Thomas K Landauer, "A Mathematical Model of the Finding of Usability Problems," Proceedings of ACM INTERCHI'93 Conference (Amsterdam, The Netherlands, 24–29 April 1993), pp. 206–213. Nielsen and Landauer's formula and curve reproduced in Nielsen, "Why You Only Need to Test With 5 Users."
26. Nielsen, "Why You Only Need to Test With 5 Users."
27. Ibid.
28. Ibid.
29. Ibid.
30. Nielsen, *Usability Engineering*, p. 21.