Reprint of paper from *SIGDOC 97 Proceedings*,
(Salt Lake City, UT), published by Association
for Computing Machinery, Inc., 1997.

Usability Studies of WWW Sites: Heuristic Evaluation vs. Laboratory Testing

Laurie Kantner
Tec-Ed, Inc.
P.O. Box 1905
Ann Arbor, MI 48106
734-995-1010
734-995-1025 fax
laurie@teced.com

Stephanie Rosenbaum
Tec-Ed, Inc.
P.O. Box 2351
Palo Alto, CA 94309
650-493-1010
650-494-1010 fax
stephanie@teced.com

ABSTRACT

This paper describes the strengths and weaknesses of two usability assessment methods frequently applied to web sites. It uses case histories of WWW usability studies conducted by the authors to illustrate issues of special interest to designers of web sites. The discussion not only compares the two methods, but also discusses how an effective usability process can combine them, applying the methods at different times during site development.

PREREQUISITES FOR ASSESSING WEB SITE USABILITY

The two methods discussed in this paper for assessing the usability of web sites both require the usability specialist to have three vital pieces of background information: the purpose of the web site; profiles of its intended users; and typical scenarios for users accessing the site. These elements are equally important in evaluating the usability of any product or service, but here is how they apply especially to Web site evaluation.

- When discussing the purpose of a web site, it's helpful to consider three categories. Web sites that supply descriptions of companies (or other organizations) and their products, services, informational offerings, or events can be described as informational sites. Web sites that provide explicit links to extensive databases are called search sites. Web sites that behave like products, where users perform other tasks in addition to reading or retrieving information, are referred to as transactional sites. Multi-purpose sites blur these boundaries.

- Unfortunately, the definition of *user* in our increasingly Web-centric environment is becoming more vague, because "anyone can access the site." However, we must keep in mind which site visitors are the most likely—or the most welcome—and focus usability efforts on those subgroups.

- Finally, the scenario for accessing a web site might be a straightforward URL to a home page, or a more roundabout path through a link in search-engine results to a page deep in the bowels of a site. Evaluators should keep in mind that any web page might be the user's door to that web site. Although users may perform more complex tasks in transactional sites, the free-form nature of navigation in any type of web site makes ensuring (and measuring) success more complex in the Web environment.

Because web sites are becoming comparable in functionality depth to many computer-based products, it's not surprising that the methods discussed in this paper can be similarly used for other software and hardware products and systems. The authors encourage readers to apply our findings to other development efforts.

HEURISTIC EVALUATION IDENTIFIES MANY PROBLEMS

Heuristic evaluations are expert evaluations of products or systems, including information systems and documentation. They're conducted by usability specialists, domain experts, or—preferably—by "double experts" with both usability and domain experience.

Evaluators use industry-accepted guidelines for usability ("heuristics"), their own experience from prior usability studies, their domain knowledge, and their ability to "put on the user's hat" when identifying problems and recommending solutions. The most effective heuristic evaluations bring all of these skills to the evaluation effort.

Heuristic evaluations by two or more usability specialists can identify a majority of the usability problems in a web site or other product, with the problem-identification percentage increasing as you add evaluators [3]. Two evaluators can identify over 50% of the problems, and three can identify about 60%. The curve flattens after five evaluators; it would take 15 evaluators to identify 90% of usability problems.

More evaluators not only find more problems, but also provide a better indication of their severity. However, more evaluators also require more resources to perform the evaluation, as well as more schedule time to coordinate their findings and agree on recommendations.

Strengths of Heuristic Evaluation

Heuristic evaluation is especially valuable when time and resources are short. Skilled evaluators can produce high-quality results in a limited time—usually two or three weeks, including a report of findings and recommendations—because the method doesn't involve detailed scripting or time-consuming participant recruiting.

As the case histories in this paper illustrate, heuristic evaluation can enable many usability improvements to take place before a release deadline that would not permit formal laboratory testing. If the development team is open to new ideas, heuristic evaluation can be an excellent investment of usability resources.

Also, heuristic evaluation as the first phase of a two-phase usability effort can greatly increase the value of laboratory testing. By identifying obvious or clear-cut usability problems, heuristic evaluation "harvests the low-hanging fruit" and provides a focus for laboratory testing.

Without prior heuristic evaluation, ten test participants may spend half their sessions struggling with the same obvious usability problem. Meanwhile, other, equally important usability problems can be "masked" by the first problem and not be found during laboratory testing.

This two-phase approach is consistent with current iterative software development practices. For example, heuristic evaluation can take place on an early prototype, while laboratory testing can follow at the alpha stage.

Concerns about Heuristic Evaluation

The major drawback of heuristic evaluation is that, regardless of the evaluators' skill and experience, they remain surrogate users (expert evaluators who emulate users) and not typical users of the web site. The results of heuristic evaluation are not actual ("primary") user data and thus are slightly suspect.

Real users always surprise us: they often have problems we don't expect, and they sometimes breeze through where we expect them to bog down. Other reasons why heuristic evaluation shouldn't replace studying actual users are that it rarely emulates all the key audience groups for the site, and it doesn't necessarily indicate which problems users will encounter most frequently.

In addition, heuristic evaluation is highly dependent on the skills and experience of the evaluators. Usability specialists may lack domain expertise, and domain specialists are rarely trained or experienced in usability methodology.

The authors find it best to concentrate on usability expertise, because the web site publishers or product developers usually can fill gaps in domain knowledge. Another approach, especially when the site is designed for users with specialized backgrounds, is to combine heuristic evaluation with a few user interviews. These interviews inform the evaluation by giving the usability specialists insight into the specific needs and concerns of the target users. The results report normally summarizes the interviews as well as the evaluators' findings and recommendations.

Another concern about heuristic evaluation is political rather than relating to human factors methodology, but it's no less real. For every new web site (or other product in progress), the product developers and marketers often have strong design opinions. The results of a heuristic evaluation can sound like just another opinion, and why should the developers accept the usability specialists' opinion over their own?

In organizations with ongoing usability programs, it's easier to educate new project teams about the value usability specialists bring to the development effort. We can describe the research basis for heuristic evaluation, and we can manage expectations and write usable, explicit results reports.

However, the authors rarely recommend a heuristic evaluation as the first usability project for an organization. Giving software developers the experience of watching real users, while not always economical, may be needed to build the credibility of the usability specialists.

Methodology for Heuristic Evaluation

The authors' methodology for performing heuristic evaluations is to create a team of at least two usability specialists, who perform independent evaluations of the user interface and take notes on their findings. The evaluators then discuss their separate findings and find common ground for communicating the findings to the developers.

The findings are usability problems and concerns about the site, as well as notes of successful features that shouldn't be changed. Often we can recommend specific UI improvements; sometimes we only suggest design directions to follow.

We generally organize our findings into four categories: user task support, UI behavior, presentation, and terminology. Although there tends to be overlap in findings among these categories, using the categories ensures that we give full attention to each aspect of a usability problem.

The evaluation team always delivers a written report of findings and recommendations. When practical, we give an oral results presentation as well, to discuss the findings with the developers.

CASE HISTORIES OF WEB SITE HEURISTIC EVALUATIONS

The authors recently performed heuristic evaluations of two quite different web sites, one providing end-user access to organizations' internal documents through a catalog server, and another, to be published later this year, that will be used to search for industrial product information within a proprietary database. The latter project was the first of a series of iterative usability studies; this paper also describes the usability testing performed on the same site.

Web Site for Document Access

A major company developing Web-centric software was working on a new release of a catalog server product. The catalog server locates and categorizes information from throughout an intranet, and creates a specialized database (called a catalog) of information about documents and other URLs that contain desired information. An end-user web site enables users to browse and search the catalog, and to link to the desired documents.

Tec-Ed was commissioned to conduct a heuristic evaluation of the end-user and the catalog administration portions of the product. The client company chose the heuristic evaluation method both to limit their investment and to obtain results in time for a fast-approaching release.

Fortunately, the company has an active and experienced user-interface design group, and the company culture supports their recommendations. Thus the evaluation team obtained good support from the developers, and our recommendations were welcomed. For example, the administration software was so complex that the evaluators required several hours of demonstrations and explanations from the developers before we could be confident of identifying usability problems, despite our assignment of a "double expert" to the evaluation team to provide domain knowledge.

Some of the usability problems identified in the end-user web site included insufficient location indicators on the pages to tell how far down the user had probed in the category hierarchy, lack of a smooth transition between searching and browsing processes, and a variety of terminology problems inherited from the administration software. These problems were fairly similar to other the authors have noted in search sites.

In the oral presentation of our recommendations, the evaluators noted which improvements were good candidates for implementing before the next product release. The audience, consisting of about 15 developers, UI designers, and documentation staff members (and their managers), participated actively in the discussion and pointed out dependencies between this site and other products the evaluators had not studied.

Web Site for Industrial Product Information, Phase 1

A major publisher of industrial product information is developing a web site for engineers, managers, purchasing personnel, and other audiences to look up product information. The target audiences may or may not already use web information resources, so web site success depends on successful first use. In addition, the site needs to be easy to use for ongoing use.

The publisher commissioned a series of studies of the web site user interface. The first study was a heuristic evaluation to identify first-tier problems that did not require collection of user data to identify—problems such as inconvenient placement of screen elements, unfamiliar terminology, and cross-platform readability issues.

The software engineers developing the site were already receptive to the value added by usability assessments. Many of the issues the evaluators identified had already emerged in development discussions and informal UI walkthroughs. In addition, although the prototype user interface had not yet undergone graphic redesign, the heuristic evaluation results gave the design firm more insight into how users approached their search tasks.

Meanwhile, the developers worked from the evaluators' suggestions to create a more usable interface for the next prototype, on which we conducted exploratory scenario-based usability testing.

LABORATORY TESTING COLLECTS ACTUAL USER EXPERIENCES

In laboratory-based usability testing, people whose characteristics (or "profiles") match those of the web site's target audience perform a sequence of typical tasks using the site. These test participants, usually working one at a time, all perform the same tasks under controlled conditions.

A detailed description of formal laboratory testing methodology is beyond the scope of this paper. Several recent books and papers discuss laboratory testing in detail [1, 2]. A previous paper by the authors also compares laboratory testing to several other usability methods [4].

Laboratory testing of web sites can explore questions with measurable answers, confirm or challenge the assumptions of developers, and help choose between design alternatives. The issues and questions to be answered and the characteristics of the desired participants are usually described in a test plan or test design document.

Based on the test plan, the usability team creates a script for the test administrator, so that all participants receive the same instructions and error remediation. We also create a participant screening questionnaire and recruiting script.

Using the test script, the administrator facilitates the usability testing sessions, while a second usability specialist observes, taking detailed notes of the participants' behavior and comments. Participants also complete questionnaires or have debriefing interviews about their experiences and opinions. The entire process usually takes four to six weeks, including results reporting, although it can be accelerated.

Strengths of Laboratory Testing

Laboratory testing is valuable when making clear-cut design decisions about web sites, especially if measurable data will help in the decision-making. For example, laboratory testing can answer questions like:

- Which of two alternative designs for a home page do users prefer, and why?
- What problems do users encounter performing product registration on a web site? How long does the registration process take? How long do users want it to take?
- How long does it take users to find desired information on a search site? How many and what kind of errors do users make in specifying the desired information?
- What problems do people encounter when downloading software from a web site? How long does a typical download process take?

Because it can collect measurable, quantitative data, laboratory testing builds credibility for usability research, especially in technical or engineering-driven organizations. Corporate managers accustomed to numerical data also find laboratory testing reassuring.

In addition, laboratory testing has a strong psychological benefit for the observers. If web site developers can watch actual test participants having problems using the site, this experience is often more convincing than the opinions of usability specialists, however similar. (A dedicated laboratory facility isn't required; developers can observe at a remote monitor through a video-camera feed, or watch videotapes after the test sessions.)

Concerns about Laboratory Testing

Web sites are revised more quickly and more often than software that will reside on a user's computer. While usability feedback on web sites can be more immediately implemented, there may be less motivation to conduct formal usability testing because the version that was evaluated has already undergone revision.

Especially when navigation from the home page is an issue, a changing web site can degrade the script developed to explore the issues identified for a study. Cooperation is needed from web-site developers to resist modifying a particular web-page version while laboratory testing takes place, and from usability specialists to be willing to adjust the script right up to the day before the test, if needed.

Laboratory testing, even scaled-back small-sample usability testing with tightly focused issues and 4 to 6 participants per audience group [5], usually requires more resources and takes longer than heuristic evaluation. Because of the need to recruit participants with profiles that match the target audience for the site, it's very difficult to gain reliable data from a laboratory test in less than three weeks from the project start date, and many laboratory tests take considerably longer.

Methodology for Laboratory Testing

The authors' methodology for collecting data during usability tests of web sites is to work from a script that provides specific prompts for note-taking about user activities. We also have a printout of the web pages themselves on which to jot down where users visited and in what page order.

The vast number of user path alternatives at a web site, especially a large informational web site, makes usability test task scenarios trickier to scope. Rather than directing users to specific paths, our approach has been to allow users to go wherever they please to perform a task; we track where they go and their stated reasons. The greater the number of users recruited, the more we can assess which pathways are more frequently traveled and why.

The browser history list does not adequately record the order of pages visited, the links selected, or how much time users spent on each page. Server logs provide vast amounts of data that requires time-consuming analysis, and even then one does not know *why* a user spent a lot of time on a page. Our note-taking method captures these types of information, which we believe are critical to understanding the scope of usability problems at a web site. Of course, we also videotape the test sessions, but our clients usually want the results more quickly than we can deliver if we need to watch all the videos.

CASE HISTORIES OF WEB SITE LABORATORY TESTING

The following three case histories describe usability testing of three very different web sites: a site for downloading software, a site for industrial product information, and a company home page.

Web Site for Downloading Software

A software development company had learned through server log analysis that users were not succeeding in downloading a newly available application from its web site. The company commissioned a study to identify the problem areas and to compare its downloading process with that of a competitor.

Tec-Ed designed a study with three tasks for each web site: find the web page from which to download the software (we provided a generic software description, not the actual product name), download a trial version of the software, and purchase the software. A team of two usability specialists developed a script that contained simple task sheets for the user; a script for the administrator with places to note high-priority observations for quick-results reporting; and a copy of each possible web page users might visit, on which to note how the user got there, what the user did, and how the user left the page, for the detail in the final report.

The data collected enabled Tec-Ed to compare time spent on the web pages for each site, the number of pages visited for each site, the amount of scrolling performed to find links, the level of comfort users expressed with using the web pages for each site, users' stated preferences for which site was easier to use, users' ability to recognize the product name they were looking for, and users' success rates in filling in the registration form, downloading, and purchasing the software. Tec-Ed also collected users' opinions about the appeal of the home page for each site.

Tec-Ed presented the findings in a quick-results session and then in a formal report. Within days after the quick-results session, improvements in the web pages began to appear.

Web Site for Industrial Product Information, Phase 2

Concurrent with performing the heuristic evaluation of the industrial-products web site, the usability team planned the first round of usability testing, which was then performed with a version of the product that reflected the recommendations arising from the heuristic evaluation. The task scenarios for the usability test were based on the preliminary product information that the web site would access and the concerns identified during heuristic evaluation. The usability test informed the graphic design revision already underway.

Tec-Ed assigned a team of two usability specialists to administer and observe 12 test sessions, using participants who met the screening criteria for people who would be likely users of the web site. Tec-Ed collected both qualitative and quantitative data, including which choices users made to conduct their searches, how satisfied they were with the search results, and improvements they wanted in the final product.

Tec-Ed's recommendations were the basis for improving the product before demonstrating it at a national trade show. Still to come are usability studies of the newly redesigned user interface, the pre-alpha version, and the beta-test version of the product.

Two Designs for a Company Home Page

A software company with a large, heterogeneous audience for its web site was planning a major redesign of its home page. They created prototypes of two alternative designs for usability testing. However, the prototypes were very preliminary (containing limited information and having some functionality problems), and the usability project was under tight time and budget constraints.

Tec-Ed therefore conducted exploratory usability testing, with facilitation by the test administrator to compensate for missing information and aberrant system behavior. Although the schedule and budget only permitted testing eight participants (four each in two audience groups), we counterbalanced the study; two participants from each audience group saw each design alternative first.

The goals of the usability test were to explore:

- Users' understanding of how the various screen features work.
- Which features and functions users find difficult/easy to use.
- How well the screen behavior matches users' expectations.
- Which features and functions users like and don't like.
- Which design allows easiest access to information.
- Which design users prefer.

Since this site has not yet been published, this paper can't include specific design details. However, the results were quite consistent with other usability tests Tec-Ed has performed comparing two product designs.

Most of the participants had similar experiences and opinions of various UI behavioral features, regardless of which alternative design they were from. The usability team was able to make clear recommendations about which UI behavioral features to adopt, improve, or discard.

Participants' opinions were more divided with respect to the graphic or visual design of the two alternatives. For example, some participants liked the color or the typography of one alternative, while other participants didn't.

Although the company wanted to know which design users preferred, the study couldn't answer that question. Five participants preferred one alternative and three the other. This preference is not statistically significant; you can't reliably predict from this data that more users will prefer one design.

The usability test was highly worthwhile despite this lack of a clear preference. We asked the participants not only which alternative design they preferred, but why. Their answers to "why?" provided the designers with considerable insight for decision-making. Also, the problems participants encountered with various behavioral features and their strong preferences (and dislikes) for certain features gave additional design guidance.

Our recommendation from this usability test was not to choose either alternative design exactly as shown in its early prototype, but to design a new version of the home page that included the usable and appealing features from both alternatives, while avoiding (or solving) their problems. If a comparative usability test is performed early enough in the development of a web site or other product—as this one was—the test results will often help create a product that's better than either alternative tested.

CONCLUSION

In considering which of the two methods presented in this paper to try first for evaluating web site usability, let's suppose an organization or company has just a small window of time in which to prove the value of usability research in the development cycle. In that case, the authors recommend starting with collection of primary user data through laboratory testing. The reason is that actual user data will convince more people, especially in engineering-driven companies, than will "just another opinion." Of course, the methodology of an expert heuristic evaluation yields more than a personal opinion, as described earlier in this paper.

If an organization already has a usability program in place, an iterative sequence of heuristic evaluation followed by laboratory testing achieves the greatest value from each method. The heuristic evaluation makes a first pass at catching the most visible usability problems ("the low-hanging fruit"), enabling laboratory testing to focus on deeper issues. In addition, iterative testing is critical to uncovering issues arising from resolution of earlier problems.

Would the authors ever recommend heuristic evaluation alone? Yes, for a site that has already undergone iterative testing and is now receiving minor revisions, or for a site that has an extremely small usability budget—because some usability evaluation is better than none at all.

BIBLIOGRAPHY

1. Dumas, J.S. & Redish, J.C. (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex.

2. Kantner, L. (1994). Techniques for Managing a Usability Test. *IEEE Transactions of Professional Communication*, 37(3) pp. 143-148.

3. Nielsen, J. (1993). *Usability Engineering*. New York, NY: Academic Press, Inc.

4. Rosenbaum, S. and Kantner, L. (1995). "Alternative Methods for Usability Testing." *ErgoCon '95 Proceedings*, San Jose, CA, pp. 47-51.

5. Virzi, R.A. (1992). "Refining the Test Phase of Usability Evaluation: How Many Subjects is Enough?" *Human Factors*, 34(4).