18

Usability Testing

There are two major considerations when conducting usability testing. The first is to ensure that the best possible method for testing is used. Generally, the best method is to conduct a test where representative participants interact with representative scenarios. The tester collects data on the participant's success, speed of performance, and satisfaction. The findings, including both quantitative data and qualitative observations information, are provided to designers in a test report. Using 'inspection evaluations,' in place of well-controlled usability tests, must be done with caution. Inspection methods, such as heuristic evaluations or expert reviews, tend to generate large numbers of potential usability 'problems' that never turn out to be actual usability problems.

The second major consideration is to ensure that an iterative approach is used. After the first test results are provided to designers, they should make changes and then have the Web site tested again. Generally, the more iterations, the better the Web site.

18:1 Use an Iterative Design Approach

Guideline: Develop and test prototypes through an iterative design approach to create the most useful and usable Web site.

Relative Importance:
1 2 3 4 ○

Strength of Evidence:
1 2 3 4 5

Comments: Iterative design consists of creating paper or computer prototypes, testing the prototypes, and then making changes based on the test results. The 'test and make changes' process is repeated until the Web site meets performance benchmarks (usability goals). When these goals are met, the iterative process ends.

The iterative design process helps to substantially improve the usability of Web sites. One recent study found that the improvements made between the original Web site and the redesigned Web site resulted in thirty percent more task completions, twenty-five percent less time to complete the tasks, and sixty-seven percent greater user satisfaction. A second study reported that eight of ten tasks were performed faster on the Web site that had been iteratively designed. Finally, a third study found that forty-six percent of the original set of issues were resolved by making design changes to the interface.

Sources: Badre, 2002; Bailey, 1993; Bailey and Wolfson, 2005; Bradley and Johnk, 1995; Egan, et al., 1989; Hong, et al., 2001; Jeffries, et al., 1991; Karat, Campbell, and Fiegel, 1992; LeDoux, Connor and Tullis, 2005; Norman and Murphy, 2004; Redish and Dumas, 1993; Tan, et al., 2001.

See page xxii for detailed descriptions of the rating scales
1 2 3 4 ○

18:2 Solicit Test Participants' Comments

Guideline: Solicit usability testing participants' comments either during or after the performance of tasks.

Relative Importance: 1 2 3 ○ ○

Strength of Evidence: 1 2 3 4 ○

Comments: Participants may be asked to give their comments either while performing each task ('think aloud') or after finishing all tasks (retrospectively). When using the 'think aloud' method, participants report on incidents as soon as they happen. When using the retrospective approach, participants perform all tasks uninterrupted, and then watch their session video and report any observations (critical incidents).

Studies have reported no significant difference between the 'think aloud' versus retrospective approaches in terms of the number of useful incident reports given by participants. However, the reports (with both approaches) tended to be positively biased and 'think aloud' participants may complete fewer tasks. Participants tend not to voice negative reports. In one study, when using the 'think aloud' approach, users tended to read text on the screen and verbalize more of what they were doing rather than what they were thinking.

Sources: Bailey, 2003; Bowers and Snyder, 1990; Capra, 2002; Hoc and Leplat, 1983; Ohnemus and Biers, 1993; Page and Rahimi, 1995; Van Den Haak, De Jong, and Schellens, 2003; Wright and Converse, 1992.

18:3 Evaluate Web Sites Before and After Making Changes

Guideline: Conduct 'before and after' studies when revising a Web site to determine changes in usability.

Relative Importance: 1 2 3 ○ ○

Strength of Evidence: 1 2 3 ○ ○

Comments: Conducting usability studies prior to and after a redesign will help designers determine if changes actually made a difference in the usability of the site. One study reported that only twenty-two percent of users were able to buy items on an original Web site. After a major redesign effort, eighty-eight percent of users successfully purchased products on that site.

Sources: John and Marks, 1997; Karat, 1994a; Ramey, 2000; Rehman, 2000; Williams, 2000; Wixon and Jones, 1996.

18:4 Prioritize Tasks

Relative Importance: 1 2 3 ○ ○

Strength of Evidence: 1 2 ○ ○ ○

Guideline: Give high priority to usability issues preventing 'easy' tasks from being easy.

Comments: When deciding which usability issues to fix first, address the tasks that users believe to be easy but are actually difficult. The Usability Magnitude Estimation (UME) is a measure that can be used to assess user expectations of the difficulty of each task. Participants judge how difficult or easy a task will be before trying to do it, and then make a second judgment after trying to complete the task. Each task is eventually put into one of four categories based on these expected versus actual ratings:

- Tasks that were expected to be easy, but were actually difficult;
- Tasks that were expected to be difficult, but were actually easy;
- Tasks that were expected to be easy and were actually easy; and
- Tasks that were expected to be difficult and were difficult to complete.

Sources: Rich and McGee, 2004.

18:5 Distinguish Between Frequency and Severity

Relative Importance: 1 2 3 ○ ○

Strength of Evidence: 1 2 3 ○ ○

Guideline: Distinguish between frequency and severity when reporting on usability issues and problems.

Comments: The number of users affected determines the frequency of a problem. To be most useful, the severity of a problem should be defined by analyzing difficulties encountered by individual users. Both frequency and severity data can be used to prioritize usability issues that need to be changed. For example, designers should focus first on fixing those usability issues that were shown to be most severe. Those usability issues that were encountered by many participants, but had a severity rating of 'nuisance,' should be given much less priority.

Sources: Woolrych and Cockton, 2001.

See page xxii for detailed descriptions of the rating scales 1 2 3 4 ○

18:6 Select the Right Number of Participants

Guideline: Select the right number of participants when using different usability techniques. Using too few may reduce the usability of a Web site; using too many wastes valuable resources.

Relative Importance: 1 2 3 ○ ○

Strength of Evidence: 1 2 3 4 ○

Comments: Selecting the number of participants to use when conducting usability evaluations depends on the method being used:

- Inspection evaluation by usability specialists:
  - The typical goal of an inspection evaluation is to have usability experts separately inspect a user interface by applying a set of broad usability guidelines. This is usually done with two to five people.
  - The research shows that as more experts are involved in evaluating the usability of the product, the greater the number of usability issues will be identified. However, for every true usability problem identified, there will be at least one usability issue that is not a real problem. Having more evaluators does decrease the number of misses, but is also increases the number of false positives. Generally, the more expert the usability specialists, the more useful the results.

- Performance usability testing with users:
  - Early in the design process, usability testing with a small number of users (approximately six) is sufficient to identify problems with the information architecture (navigation) and overall design issues. If the Web site has very different types of users (e.g., novices and experts), it is important to test with six or more of each type of user. Another critical factor in this preliminary testing is having trained usability specialists as the usability test facilitator and primary observers.
  - Once the navigation, basic content, and display features are in place, quantitative performance testing (measuring times, wrong pathways, failure to find content, etc.) can be conducted to ensure that usability objectives are being met. To measure each usability objective to a particular confidence level, such as ninety-five percent, requires a larger number of users in the usability tests.
  - When the performance of two sites is compared (i.e., an original site and a revised site), quantitative usability testing should be employed. Depending on how confident the usability specialist wants to be in the results, the tests could require a larger number of participants.

– It is best to perform iterative cycles of usability testing over the course of the Web site's development. This enables usability specialists and designers to observe and listen to many users.

Sources: Bailey, 1996; Bailey, 2000c; Bailey, 2000d; Brinck and Hofer, 2002; Chin, 2001; Dumas, 2001; Gray and Salzman, 1998; Lewis, 1993; Lewis, 1994; Nielsen and Landauer, 1993; Perfetti and Landesman, 2001; Virzi, 1990; Virzi, 1992.

18:7 Use the Appropriate Prototyping Technology

Guideline: Create prototypes using the most appropriate technology for the phase of the design, the required fidelity of the prototype, and skill of the person creating the prototype.

Relative Importance: 1 2 ○ ○ ○

Strength of Evidence: 1 2 3 ○ ○

Comments: Designers can use either paper-based or computer-based prototypes. Paper-based prototyping appears to be as effective as computer-based prototyping when trying to identify most usability issues. Several studies have shown that there was no reliable difference in the number of usability issues detected between computer and paper prototypes. However, usability test participants usually prefer interacting with computer-based prototypes. Paper prototypes can be used when it is necessary to view and evaluate many different (usually early) design ideas, or when computer-based prototyping does not support the ideas the designer wants to implement, or when all members of the design team need to be included–even those that do not know how to create computer-based prototypes.

Software tools that are available to assist in the rapid development of prototypes include PowerPoint, Visio, including other HTML base tools. PowerPoint can be used to create medium fidelity prototypes. These prototypes can be both interactive and dynamic, and are useful when the design requires more than a 'pencil-and-paper' prototype.

Sources: Sefelin, Tscheligi and Giller, 2003; Silvers, Voorheis and Anders, 2004; Walker, Takayama and Landay, 2002.

See page xxii for detailed descriptions of the rating scales 1 2 3 4 ○

18:8 Use Inspection Evaluation Results Cautiously

Guideline: Use inspection evaluation results with caution.

Relative Importance: 1 2 ○ ○ ○

Strength of Evidence: 1 2 3 4 ○

Comments: Inspection evaluations include heuristic evaluations, expert reviews, and cognitive walkthroughs. It is a common practice to conduct an inspection evaluation to try to detect and resolve obvious problems before conducting usability tests. Inspection evaluations should be used cautiously because several studies have shown that they appear to detect far more potential problems than actually exist, and they also tend to miss some real problems. On average, for every hit there will be about 1.3 false positives and .5 misses.

Another recent study concluded that the low effectiveness of heuristic evaluations as a whole was worrisome because of the low problem detection rate (p=.09), and the large number of evaluators required (16) to uncover seventy-five percent of the potential usability issues.

Another difficulty when conducting heuristic evaluations is that evaluators frequently apply the wrong heuristic, which can mislead designers that are trying to fix the problem. One study reported that only thirty-nine percent of the heuristics were appropriately applied.

Evaluators seem to have the most success identifying usability issues that can be seen by merely looking at the display, and the least success finding issues that require users to take several steps (clicks) to a target.

Heuristic evaluations and expert reviews may best be used to identify potential usability issues to evaluate during usability testing. To improve somewhat on the performance of heuristic evaluations, evaluators can use the 'usability problem inspector' (UPI) method or the 'Discovery and Analysis Resource' (DARe) method.

Sources: Andre, Hartson and Williges, 2003; Bailey, Allen and Raiello, 1992; Catani and Biers, 1998; Cockton and Woolrych 2001; Cockton and Woolrych, 2002; Cockton, et al., 2003; Fu, Salvendy and Turley, 1998; Fu, Salvendy and Turley, 2002; Law and Hvannberg, 2002; Law and Hvannberg, 2004; Nielsen and Landauer, 1993; Nielsen and Mack, 1994; Rooden, Green and Kanis, 1999; Stanton and Stevenage, 1998; Virzi, Sorce and Herbert, 1993; Wang and Caldwell, 2002.

See page xxii for detailed descriptions of the rating scales 1 2 3 4 ○

18:9 Recognize the 'Evaluator Effect'

Relative Importance: 1 2 ○ ○ ○
Strength of Evidence: 1 2 3 4 ○

Guideline: Beware of the 'evaluator effect' when conducting inspection evaluations.

Comments: The 'evaluator effect' occurs when multiple evaluators evaluating the same interface detect markedly different sets of problems. The evaluators may be doing an expert review, heuristic evaluation, or cognitive walkthrough.
The evaluator effect exists for evaluators who are novice or experienced, while detecting cosmetic and severe problems, and when evaluating simple or complex Web sites. In fact, when using multiple evaluators, any one evaluator is unlikely to detect the majority of the 'severe' problems that will be detected collectively by all evaluators. Evaluators also tend to perceive the problems they detected as more severe than the problems detected by others.

The main cause of the 'evaluator effect' seems to be that usability evaluation is a complex cognitive activity that requires evaluators to exercise difficult judgments.

Sources: Hertzum and Jacobsen, 2001; Jacobsen, Hertzum and John, 1998; Molich, et al., 1998; Molich, et al., 1999; Nielsen and Molich, 1990; Nielsen, 1992; Nielsen, 1993; Redish and Dumas, 1993; Selvidge, 2000.

18:10 Apply Automatic Evaluation Methods

Relative Importance: 1 ○ ○ ○ ○
Strength of Evidence: 1 2 3 ○ ○

Guideline: Use appropriate automatic evaluation methods to conduct initial evaluations on Web sites.

Comments: An automatic evaluation method is one where software is used to evaluate a Web site. An automatic evaluation tool can help find certain types of design difficulties, such as pages that will load slowly, missing links, use of jargon, potential accessibility problems, etc. While automatic evaluation methods are useful, they should not be used as a substitute for evaluations or usability testing with typical users. There are many commercially available automatic evaluation methods available for checking on a variety of Web site parameters.

Sources: Brajnik, 2000; Campbell and Stanley, 1963; Gray and Salzman, 1998; Holleran, 1991; Ivory and Hearst, 2002; Ramey, 2000; Scholtz, 1998; World Wide Web Consortium, 2001.

18:11 Use Cognitive Walkthroughs Cautiously

Guideline: Use cognitive walkthroughs with caution.

Relative Importance: 1

Strength of Evidence: 1 2 3 4

Comments: Cognitive walkthroughs are often conducted to resolve obvious problems before conducting performance tests. The cognitive walkthrough appears to detect far more potential problems than actually exist, when compared with performance usability testing results. Several studies have shown that only about twenty-five percent of the potential problems predicted by the cognitive walkthrough were found to be actual problems in a performance test. About thirteen percent of actual problems in the performance test were missed altogether in the cognitive walkthrough. Cognitive walkthroughs may best be used to identify potential usability issues to evaluate during usability testing.

Sources: Blackmon, et al., 2002; Desurvire, Kondziela and Atwood, 1992; Hassenzahl, 2000; Jacobsen and John, 2000; Jeffries and Desurvire, 1992; John and Mashyna, 1997; Karat, 1994b; Karat, Campbell and Fiegel, 1992; Spencer, 2000.

18:12 Choosing Laboratory vs. Remote Testing

Guideline: Testers can use either laboratory or remote usability testing because they both elicit similar results.

Relative Importance: 1

Strength of Evidence: 1 2 3 4

Comments: In laboratory-based testing, the participant and the tester are in the same physical location. In remote testing, the tester and the participant are in different physical locations. Remote testing provides the opportunity for participants to take a test in their home or office. It is convenient for participants because it requires no travel to a test facility.

Studies have evaluated whether remote testing is as effective as traditional, lab-based testing. To date, they have found no reliable differences between lab-based and remote testing in terms of the number of types of usability issues identified. Also, they report no reliable differences in task completion rate, time to complete the tasks, or satisfaction scores.

Sources: Brush, Ames and Davis, 2004; Hartson, et al., 1996; Thompson, Rozanski and Rochester, 2004; Tullis, et al., 2002.

18:13 Use Severity Ratings Cautiously

Guideline: Use severity ratings with caution.

Relative Importance:
1○○○○

Strength of Evidence:
1234○

Comments: Most designers would like usability specialists to prioritize design problems that they found either by inspection evaluations or expert reviews. So that they can decide which issues to fix first, designers would like the list of potential usability problems ranked by each one's 'severity level'. The research literature is fairly clear that even highly experienced usability specialists cannot agree on which usability issues will have the greatest impact on usability.

One study had 17 expert review and usability test teams evaluate and test the same Web page. The teams had one week to do an expert review, or two weeks to do a usability test. Each team classified each usability issue as a minor problem, serious problem, or critical problem. There was considerable disagreement in which problems the teams judged as minor, serious or critical, and there was little agreement on which were the 'top five problems'. Another study reported that heuristic evaluators overestimated severity twenty-two percent of the time, and underestimated severity seventy-eight percent of the time when compared with usability testing results.

Sources: Bailey, 2005; Catani and Biers, 1998; Cockton and Woolrych, 2001; Dumas, Molich and Jeffries, 2004; Hertzum and Jacobsen, 2001; Jacobsen, Hertzum and John, 1998; Law and Hvannberg, 2004; Molich, 2005.

See page xxii for detailed descriptions of the rating scales
1234○