Usability Evaluation

Jean Scholtz
National Institute of Standards and Technology

Introduction

The International Organization for Standardization (ISO) defines Usability of a product as “the extent to which the product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.” Usability has five attributes: learnability, efficiency, memorability, errors, and user satisfaction. Depending on the type of application one attribute might be more critical than another. For example, if the software will be used infrequently then it is essential that users can easily remember the actions necessary for desired tasks. If the application is time critical then efficiency will be critical along with the prevention of errors.

Usability engineering is the discipline that provides structured methods for achieving usability in user interface design during product development. Usability evaluation is part of this process. While theoretically any software product could be evaluated for usability, the evaluation is unlikely to produce good results unless a usability engineering process has been followed. Usability engineering has three basic phases: requirements analysis, design/testing/development, and installation. Usability goals are established during requirements analysis. Iterative testing is done during the design/testing/development phases and the results are compared to the usability goals. User feedback should also be obtained after installation as a check on the usability and functionality of the product.

In this chapter we will focus on the three basic methods for evaluating usability. The methods differ depending on the source used for the evaluation. This source can be users, usability experts, or models. Figure 1 shows a timeline for usability evaluations in the last 30 years. Users were first used as the source of usability feedback but models have been also used for over 20 years. Expert feedback was developed in heuristic reviews and cognitive walkthroughs and has been used since the early 90s. All three methods rely on usability engineers or usability professionals to design, conduct, analyze, and report on the evaluations.

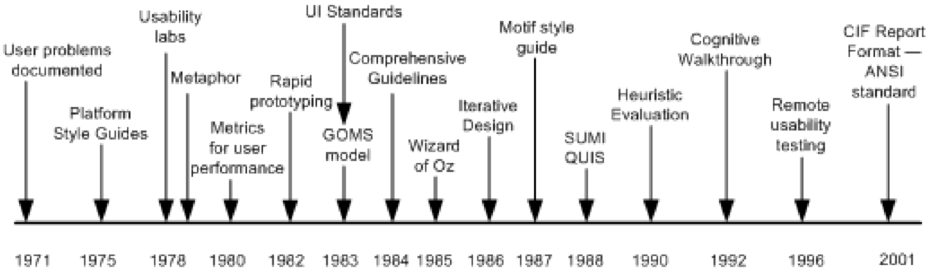

Fig. 1. 30 years of highlights in the development of desktop computing user evaluations from 1971 - 2001

User-Centered Evaluations

User-centered evaluations are accomplished by identifying representative users, representative tasks, and developing a procedure for capturing the problems that users have in trying to apply a particular software product in accomplishing these tasks. During the design/testing/development cycle of software development, two types of user evaluations are carried out. Formative evaluations are used to obtain information used in design. Summative evaluations are usability evaluations that document the effectiveness, efficiency, and user satisfaction of a product at the end of the development cycle. These two types of evaluation differ in the purpose of the evaluation, the methods used, the formality of the evaluation, the robustness of the software being evaluated, the measures collected, and the number of participants used. In both types of evaluation representative users are recruited to participate, some method of collecting information is used, and some way of disseminating the results of the evaluation to members of the software development team is needed.

The Role of Usability Laboratories

Usability laboratories are used in some companies to conduct the evaluations. Usability laboratories are useful in making a company's commitment to usability visible. Laboratories are usually outfitted with audio and video recoding equipment to record what the user is doing on the computer. The computer screen, the user's hand motions, and the facial expression of the user are usually captured on video. In addition, logging software is used to capture keystrokes to determine what the user is typing and what menu items are selected. Many laboratories are designed with rooms for the user as well as for observers. These rooms can be separated by one-way glass or the video from the user's computer can be piped into a separate room where managers and developers can observe the testing. Remote usability labs are also sold. These consist of audio and video recording equipment bundled conveniently to allow usability engineers to travel to users, rather than having users come to them. Digital video software is now available for recording and is greatly facilitating user-centered evaluations.

Formative Evaluations

Formative evaluations obtain user feedback for early concepts or designs of software products. Formative evaluations are typically more informal in that the goal is to collect information to be used for design as opposed to collecting measures of usability. The primary source of data in formative evaluations is verbal data from the user. Early evaluations may use paper prototypes or initial screen designs. Later evaluations can be done on partial prototypes or prototypes of only one portion of a user interface. When possible, logging software is also used to capture user interaction with the software. Additionally, usability engineers often take notes of critical incidents that occur during the evaluation. The debriefing or post-evaluation interview is an excellent source of information in formative evaluations. Usability engineers can probe in depth to understand sources of confusion in the interface.

Formative evaluations need to be conducted in a fairly rapid pace in order to provide design input when it is needed. As a consequence, the evaluations usually focus on a small portion of the user interface, involve relatively few user-participants, and have less formal reporting mechanisms than summative evaluations. Ideally, software designers and developers can observe the evaluations and discuss results and potential solutions with the usability engineers after the evaluations.

Summative Evaluations

Summative evaluations are more formal evaluations conducted to document the usability characteristics of a software product. These evaluations involve a number of users. The recommendation is 5-7 users per cell, where a cell represents a class of end-users. For example, if a product is being design for both home and small business users, then representatives of users of each type must be included in the evaluation. If both adults and teenagers will be using the home product, then representatives from both groups need to be included as evaluation participants. Good experimental design is essential to summative evaluation. The metrics of efficiency, effectiveness, and user satisfaction are typically used and the design of the evaluation must include the measures and collection methodology. Tasks used in the evaluation usually represent core functionality of the software but may also include new or improved functionality. Directions and materials given to the users need to be designed and tested in a pilot evaluation to make sure that they are understandable.

Usability evaluation has always tried to make the "context-of-use" as realistic as possible. However, a usability laboratory cannot duplicate actual conditions of use within an office or home. Interruptions and other demands for attention do not, and cannot, occur during usability evaluation conditions. As such these evaluations represent the best case condition. If the software is not usable in the laboratory, it will certainly not be usable in real-world use. However, usability in the laboratory does not guarantee usability in more realistic conditions.

The desired level of usability should be defined early in the usability engineering lifecycle and the actual results from the summative evaluation are compared to this. If good usability engineering practices have been followed, including a number of formative evaluations, it is likely that the desired levels will be achieved. If this is not the case, then a decision must be made as to whether or not to release the software or to redesign and re-evaluate the usability.

Advantages and Disadvantages

The chief advantage of user-centered evaluation is the involvement of users. Results are based on actually seeing what aspects of the user interface cause problems for representative users. The downside is that user evaluations are expensive and time consuming. Finding and scheduling an appropriate number of representative users for each user type is difficult. Laboratory and usability engineering resources are needed to conduct the evaluations and analyze the results. There are also issues involved as to the realism of the evaluation. Have the correct tasks been selected? How will the product work in real work environments? Beta testing and user feedback after installation are used to gather data about usability aspects of the product in the actual context of use.

Expert-based Evaluations

Expert evaluations of usability are similar to design reviews of software projects and code walkthroughs. Inspection methods include heuristic evaluation, guideline reviews, pluralistic walkthroughs, consistency inspections, standards inspections, cognitive walkthroughs, formal usability inspections, and feature inspections.

Guideline and Standards Reviews

Guideline and standard reviews need a basis for making these judgments. Standards for user interfaces have at this point been built into the graphic user interface development software. For example, there are standard widgets used by software applications developed in the Microsoft Platform. Some companies have style guides for software applications and compile a checklist to be used to assure that applications conform to this style.

Numerous guidelines have been developed to facilitate design and evaluation of user interfaces. The guidelines have been obtained from many studies and are sometimes contradictory. The application of the guideline depends on many factors including the type of application, the expertise of the end-users, and the environment in which the application will be used.

A current set of guidelines for web sites is maintained by the National Cancer Institute (NCI). These guidelines provide a measure of evidence supporting the guidelines, ranging from research experiments to observational evidence to expert opinion.

Cognitive Walkthroughs

Guidelines are more oriented towards a static view of the user interface. Aspects of the design are evaluated separately and not in the context of use in a real-world task. The Cognitive Walkthrough technique was developed to address this. To conduct a cognitive walkthrough, one needs some form of a detailed description of the user interface, a scenario for the task, the demographics of the end user population, the context in which the use would occur, and a successful sequence of actions that should be performed. The person doing the walkthrough then looks at the actions for each task and evaluates whether the user will be able to select the right action, will be able to determine that the selected action is correct, and will see progress towards the solution. Critical information to capture during the walkthrough is what the user should know prior to performing the task and what the user should learn while performing the task.

The developers of the cognitive walkthrough point out that the focus is on the learnability aspect of usability. Cognitive walkthroughs can be performed by individuals or by groups. Thus other evaluation methods must also be applied to assess other aspects of usability.

Heuristic Evaluation

The heuristic evaluation technique is the most widely used inspection method. Heuristic evaluation uses a small set of evaluators who judge a user interface for compliance with usability design principles. Table 1 provides some example heuristics used in evaluations. Each inspector judges the user interface separately and after the assessments have been, they meet to compile their findings. Five evaluators is the recommended number for critical systems and no fewer than three evaluators for any heuristic review.

Visibility of system status
Match between system and the real world
User control and freedom
Consistency and standards
Error prevention
Recognition rather than recall
Flexibility and efficiency of use
Aesthetic and minimalist design
Help users recognize, diagnose, and recover from errors
Help and documentation

Table 1: Usability heuristics (Nielsen, 1994)

Advantages and Disadvantage

Heuristic reviews are less expensive and less time-consuming to conduct than user-centered evaluations. The cognitive walkthrough can be accomplished using only a text description of the user interface and therefore can be used very early in the software development process.

Inspection techniques do not provide possible solutions to the usability problem. Moreover, it is difficult to summarize the findings from multiple evaluators as they report problems differently and at different levels. There is also the issue of severity. Not all usability problems are equal. Development teams need to be able to prioritize what problems get fixed according to the seriousness of the problem. There is currently no agreement on how to judge the severity of usability problems.

One question is how accurately these inspection methods predict problems that real user encounter? An early study found that heuristic reviews were better predictors than cognitive walkthroughs and guideline based evaluations. This was compared to results from laboratory usability tests. However, none of these methods found more than 50% of the problems discovered in laboratory testing. The last section of this chapter discusses issues in comparing evaluation methodologies.

Model-based Evaluations

GOMS Models

A model of the human information processor has been developed based on data derived from psychology research on the human systems of perception, cognition, and memory. The model incorporated capabilities of short term and long term memory, along with capabilities of the human visual and audio processing. Times for cognitive processing and motor processing are also included. This allows human-computer interaction researchers to evaluate user interface designs based on predictions of performance from the model.

The GOMS model consists of Goals, Operators, Methods and Selection rules. A method is a set of steps or operators that will accomplish a goal. In any interface there may be more than one method that will provide the same result. The user than chooses the method based on selection rules. A GOMS model can be used only for evaluating the efficiency of the procedural aspect of usability but cannot evaluate potential errors due to screen design or terminology. Natural GOMS Language (NGOMSL) has extended the GOMS model to predict learning time. For both of these techniques a task analysis must be done to construct the goals, operators, methods and selection rules.

Other Modeling Techniques

The EPIC (Executive-Process/Interactive Control) system simulates the human perceptual and motor performance system. Epic can interact as a human would with a simulation of a user interface system. EPIC is being used to study users engaged in multiple tasks, such as using a car navigation system while driving. Using EPIC involves writing production rules for using the interface and writing a task environment to simulate the behavior of the user interface.

A model of information foraging useful in evaluating information seeking in web sites is based on the ACT-R model. The ACT-IF model was developed to use in testing simulated users interacting with designs for web sites and predicts optimal behavior in large collections of web documents. The information foraging model is being used to understand the decisions that users of the web make in following various links to satisfy information goals.

Advantages and Disadvantages of Model- based Evaluations

The use of models to predict user behavior is less expensive than empirical, user-centered evaluations. Thus many more iterations of the design can be tested. However, a necessary first step is conducting the task-level cognitive task analysis to use in producing model description. This is time consuming but can be used for testing many user interface designs.

Models must be tested for validity. This is accomplished by watching humans perform the tasks and coding their behavior for comparison with the model. This is time consuming but necessary to determine if the model predicts are accurate.

Current Issues Concerning Evaluation Methodologies

While the HCI community has come a long way in developing and using methods to evaluate usability, the problem is by no means solved. This chapter has described three basic methods for evaluation but there is not yet agreement in the community about which evaluation is more useful than another. Although a number of studies have been done to compare these methods, the comparison is difficult and flaws have been pointed out in a number of these studies. First, there is the issue of using experimental (user-centered) methods to obtain answers to large questions of usability as opposed to the more narrow questions that are the more traditional use for experimental methods. A second issue is what should be used for the comparison? Should user-centered methods be considered as the ground truth? All usability tests are not created equal. There are certainly flaws in the way tests are design, conducted, and analyzed. . While individual methods have limitations and can be flawed in their implementation, it is certain that performing some evaluation methodology is better than doing nothing. The current best practice is to use a number of different evaluation methodologies to provide rich data on usability.

Evaluation methodologies were, for the most part, developed to evaluate the usability of desk-top systems. The current focus in technology development of mobile and ubiquitous computing presents challenges for current usability evaluation methods. Laboratory evaluations will be hard pressed to simulate use conditions for these applications. Going out into the field to evaluate use places constraints on how early evaluations can be done. Mobile and multi-user systems must be evaluated for privacy and any usability issues entailed in setting up, configuring, and using such policies. The use of such devices in the context of doing other work also has implications for determining the context of use for usability testing. We need to test car navigation systems in the car – not the usability lab.

Technology is being used by more users. The range of users using mobile phones, for example, means that representative users need to be selected from teenagers to grandparents. The accessibility laws in the United States require that federal information is accessible by persons with disabilities. Again, this requires inclusion of more users from the disable population in user-centered evaluations.

Web sites are also of interest in usability evaluation. Again, there is a matter of a broad user population. Design and development cycles in web site development are extremely fast and doing extensive usability evaluation is usually not feasible. Usability practitioners are looking at remote testing methods to more closely replicate context of usage for web site evaluation.

International standards exist for user centered design processes, documentation, and user interfaces. Usability is becoming a requirement for companies in purchasing software as they recognize that unusable software will increase the total cost of ownership.

New usability evaluation methodologies will be developed to meet the demands of our technology-focused society. Researchers and practitioners in usability will need to join forces to meet this challenge.

Further Reading

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human Computer Interaction*, 12(4), 439-462.

Behaviour & Information Technology, vol. **13**, nos. 1-2, January-April 1994. [Special issue devoted to usability laboratories]

Card, S.K., Moran, T.P., and Newell, A. 1983. *The psychology of human-computer interaction*. Hillsdale, NJ: Erlbaum Associates.

Card, S.K., Moran, T.P., and Newell, A. 1980. The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7), 396-410.

The Industry usability Reporting (IUSR) Project. Retrieved September 12, 2003 from http://www.nist.gov/iusr.

Dumas, J. and Redish, J. 1993. *A Practical Guide to Usability Testing*. Norwood, NJ: Ablex.

Gray, W. D., John, B. E., & Atwood, M. E. 1993. Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world performance. *Human-Computer Interaction, 8*(3), 237-309.

Gray, W. D. and Salzman, M. D. 1998. Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-computer Interaction*, Vol. 13 (3), 203-262.

Hartson, H.R., Castillo, J.C., Kelso, J., Kamler, J., and Neale, W.C. (1996). Remote Evaluation: The Network as an Extension of the Usability Laboratory. Proceedings of ACM CHI'96, (location, date). 228-235.

International Organization for Standardization. accessed Sept. 12, 2003 from http://www.iso.ch/iso/en/ISOOnline.openerpage

Jeffries, R., Miller, J.R., Wharton, C., and Uyeda, K.M. 1991. User interface evaluation in the real world: A comparison of four techniques. Proceedings ACM CHI'91 Conference (New Orleans, LA, April 28-May 2): 119-124.

Karat, J., Jeffries, R., and Miller, J., Lund, A., McClelland, I., John, B., Monk, A., Oviatt, S., Carroll, J.,, Mackay, W., and Newman, W.,Olsen, G., and Moran, T.P. 1998. commentary on "Damaged Merchandise?". *Human-computer Interaction.* Vol. 13 (3), 263-324.

Kieras, D.E. 1997. A Guide to GOMS Model Usability Evaluation Using NGOMSL. In Helander M., Landauer T. (eds.): *The handbook of human-computer interaction*. Amsterdam: North-Holland: 733-766.

Kieras, D. & Meyer, D.E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*., 12, 391-438.

Lewis, C., Polson, P., Wharton, C., and Rieman, J. 1990. Testing a Walkthrough Methodology for Theory-based Design of Walk-up-and-user Interfaces. Proceedings of ACM CHI'90 Conference (Seattle, Wa. April 1-5): 235-242.

Mack, R. and Nielsen, J. 1994. *Usability Inspection Methods*. New York, NY: John Wiley and Sons.

Mayhew, D. 1999. *The Usability Engineering Lifecycle*. San Francisco,CA: Morgan Kauffman.

Microsoft Corporation, 1995. *The Windows Interface Guidelines for Software Design.* Redmond, WA: Microsoft Press.

Muller, M., Haslwanter, J., and Dayton, T. 1997. In Helander M., Landauer T. (eds.): *The handbook of human-computer interaction*. Amsterdam: North-Holland: 255-298.

National Cancer Institute. 2003. Research-Based Web Design and Usability Guidelines. Retrieved from September 12, 2003 from http://www.usability.gov.

Nielsen, J. 1993. *Usability Engineering*. San Diego, CA: Academic Press.

Nielsen, J., and Molich, R. 1990. Heuristic evaluation of user interfaces. Proceedings ACM CHI'90 conference (Seattle, WA, April 1-5): 249-256.

Neilsen, J. Enhancing the explanatory power of usability heuristics. Proceedings ACM CHI'94 Conference , (Boston, MA. April 24-28): 152-158.

Pirolli, P. and Card, S. 1995. Information Foraging in Information Access Environments. Proceedings ACM CHI '95 Conference (locate, date), pages

Pirolli, P. (1997). Computational models of information scent-following in a very large browsable text collection. Proceedings ACM CHI '97 (Atlanta, GA, March 22-27) 3-10.

Section 508 of the Rehabilitation Act (29 U.S.C. 794d), as amended by the Workforce Investment Act of 1998 (P.L. 105-220), August 7, 1998. Retrieved September 12, 2003 from http://www.section508.gov/.

Smith, S. L, and Mosier, J. N. 1986. Guidelines for Designing User Interface Software, technical report NTIS No. A177 198. Hansom Air Force Base, MA: USAF Electronic Systems Division.

Yourdon, E. 1989. *Structured Walkthroughs* (4th ed.) Englewood cliffs, NJ: Yourdon Press.