AN EMPIRICAL STUDY OF USABILITY TESTING: HEURISTIC EVALUATION VS. USER TESTING

Enlie Wang Barrett Caldwell
Industrial Engineering, Purdue University
West Lafayette, Indiana

In this study, two different usability-testing methods (Heuristic Evaluation and User Testing) were selected to test the usability of a pre-release version of software searching for Science, Mathematics and Engineering education materials. Our major goal is to compare Heuristic Evaluation and User Testing in terms of efficiency, effectiveness and cost/benefit analysis. We found that Heuristic Evaluation was more efficient than User Testing in finding usability problems (41 vs. 10), while User Testing was more effective than Heuristic Evaluation in finding major problems (70% vs.12%). In general, Heuristic Evaluation appears to be more economic in finding a wide range of usability problems by incurring a low cost in comparison to User Testing. However, User Testing can provide more insightful data from real users such as user's performance and satisfaction.

INTRODUCTION

The Science Math and Engineering Learning Technologies (SMELT) software prototype is an online educational resource repository designed for teachers, students, and administrators in K-12 schools. Many of these users do not have a reliable Internet connection or enough time and Internet experience to find good educational resources on the web. Thus, SMELT was developed as a solution for these problems. It can help users quickly assess the resources they wanted with minimal search effort. In order to improve the usability and quality of this special educational software, we decided to use two different usability-testing methods to conduct parallel usability testing for the pre-release version of SMELT. Although improving usability of SMELT software is our ultimate goal, comparing the effectiveness and efficiency of different usability testing methods is our major research interest in this study.

A recent literature review in usability testing (Wang, Caldwell and Salvendy, 2002) summarized major usability testing methods (see table 1). Usability testing methods can be classified into four categories (Nielsen 1994a): automatic, usability measures computed by running usability inspection software during the evaluation task; informal, based on rules of thumb and general skill and experience of evaluators; empirical, tested by real users; and formal, using exact models and formulas to calculate usability measurements. After examining the strengths and weaknesses of each, two usability-testing methods (Heuristic Evaluation and User Testing) were chosen as optimal testing methods for testing in this study.

Table 1. Usability methods taxonomy (expended from Nielsen 1994c)

Method Classification	Representative Methods
Automatic	Webtesting software, such as WebArch by *Addwise*, and Web Trends by *NetIQ*
Informal	Heuristic Evaluation (Nielsen 1990) Cognitive Walkthroughs (Lewis, Wharton and Rieman, 1990) Pluralistic Walkthroughs (Bias, 1994) Guidelines, Questionnaires (i.e. QUIS) and Checklists (i.e. Lin, Choong and Salvendy, 1997) Feature Inspection (Bell, 1992)
Empirical	User Testing (response time, error rate and subjective judgment), User Feedback
Formal	GOMS (Card, Moran & Newell 1983, Gong and Kieras 1994) Thinking aloud analysis (Simon 1989, Lewis 1982, Wright 1992)

Heuristics Evaluation

Heuristic Evaluation is done by having usability experts look at an interface and trying to come up with a judgment about what is good and bad about the interface. The most popular heuristics are: *simple and natural dialogue, speak the user's language, minimize user memory load, be consistent, provide feedback, provide clearly marked exits, provide shortcuts, good error messages, prevent errors, help and documentation* (Nielsen and Molich, 1990). This testing technique is intuitive, inexpensive, and quick way of getting a fairly comprehensive usability problem report. Because of these characteristics, it is easy to motivate both management and evaluators to apply this usability testing technique. This method was chosen because the SMELT development team needed quick feedback for redesigning and improving the software before the official release date.

User Testing

Heuristic Evaluation is a quick and inexpensive method, but it is not perfect or without defect. Many studies (Nielsen, 1992, 1993a, 1994a) reported that it tended to find too many false positives (false alarms) and minor problems. A lot of the developer's time and effort may be spent working on these less critical problems. User Testing is the best way to identify what are real problems that will impact user's performance and preference. Because the problems found in User Testing are identified by actual users, there are fewer false positives identified, and all problems found are worth investigating. It also can measure user's performance and satisfaction such as delay tolerance during human-computer interaction (Wang and Caldwell, 2002). However, it takes time to prepare the testing materials and to recruit test subjects. Bailey (1996) did a study to compare the User Testing and Heuristic Evaluation directly by evaluating a series of systems with both methods. He reported that User Testing performed better than Heuristic Evaluation, and each method tended to find different types of usability problems. We also want to examine this claim through our study and compare Heuristic Evaluation and User Testing by evaluating a real software product. Thus, User Testing became the second testing method in this study.

METHODS

Heuristic Evaluation

Participants. The experiment was conducted with 5 graduate students who were, at the time, taking a Usability Testing and Evaluation Techniques seminar at University of Wisconsin-Madison. Five evaluators are sufficient to conduct Heuristic Evaluation according to Nielsen's research (Nielsen 1994 c). Evaluators had no previous experience with the SMELT software before testing. A small gift was given to each evaluator upon the completion of evaluation.

Apparatus. The experiment was run on a Pentium 3 class computer with the SMELT software installed locally on a hard drive. Instruction form and the ten heuristics list were given to the evaluators as guidelines during evaluation. The ten heuristics we used in this study are: (1) visibility of system status, (2) match between system and the real world, (3) user control and free, (4) consistency and standard, (5) error prevention, (6) recognition rather than recall, (7) flexibility and efficiency of use, (8) aesthetic and minimalist design, (9) help users recognize, diagnose, and recover from errors , and (10) help and documentation (Nielsen 1994 b).

Test Procedures. It took the evaluator about 1.5 hours to complete one evaluation session. The evaluation included following steps:

1. Introduction to the SMELT software. The experimenter gave a brief description about the software and explained the functions in the software. Next, the evaluators were asked to familiarize themselves with the software. The whole process took about 15 minutes.
2. The evaluation. Two main parts of the software were evaluated: the "User's Guide" section and the "Search" section. The evaluators were given a set of heuristic principles. They were asked to inspect dialogue elements and compare them with the heuristic principles. Then, they were asked to write the problems found and the heuristic principles that were violated by each problem. Each evaluation lasted about 30 minutes.
3. Aggregate results. After the evaluation was done each evaluator was asked to give general comments about the software, and then the results were aggregated.

Usability Testing

Participants. Ten participants participated in the User Testing. They came from a wide scholastic background including instructors and students at both high school and college levels. Academic disciplines represented Engineering, Food Science and other disciplines. This variety of backgrounds was thought to be more representative of the SMELT user population (see Table 2. Participant background). User's previous

Table 2. Participant Background

Participant Classification	Number	Age (Mean)	Experience (Mean)		
			Computer	Internet	Search Engine
Student	5	21.5 (6.5)	4.2 (2.2)	4.6 (1.1)	4.4 (0.9)
Teacher	5	26.8 (2.8)	5.0 (0.7)	4.8 (1.1)	5.0 (1.4)
Male	5	27.6 (3.1)	4.2 (1.9)	4.4 (0.9)	4.4 (0.9)
Female	5	30.6 (5.3)	5.0 (1.2)	5.0 (1.2)	5.0 (1.2)
Total	10	24.1 (5.5)	4.6 (1.6)	4.7 (1.1)	4.7 (1.2)

computer and software knowledge and experience may effect user's satisfaction and performance toward new software. Three kinds of related experience, computer, internet and search engine experience were measured by seven-point scale ranging form 1(no experience) to 7 (very experienced). The results show that our users have adequate experience in these aspects.

Apparatus. The test was run on an IBM compatible laptop computer (Pentium II, 400 MHz) with the SMELT software installed locally. A video camera recorder was used to record users' activities during testing. A series of forms were developed for this testing process: *Consent Form, General Instructions, Background Survey, User Testing Instructions, Test Task, Formal Tasks Result Recording Sheet and User Satisfaction Survey.*

The user satisfaction survey is composed of two parts. The first part is adapted form of the QUIS 5.0 survey from (Chin, Diehl and Norman, 1988), and the second part is developed by authors to measure users' views of task time, user help, user expectations, memory effort, and search effort.

Test Procedures. A standard procedure was developed that included 9 separate steps: (1) Sign consent form, (2) Complete background survey, (3) Read general instructions, (4) Learn and practice software operation, (5) Read test instructions, (6) Execute test task, (7) Execute four formal simulation tasks (search for online educational resources in a specific area and recommend the best Web site for the users in the simulation scenario), (8)Complete satisfaction survey, (9) Report problems.
In order to limit initial learning effects, users were given enough time to explore the software and the User's Guide before testing. The test was not started until the users indicated that they were familiar with the software. After the formal tasks, the user was to complete user satisfaction survey and report the problems they found during testing. Users were also given access to the SMELT software at this time to aid in recalling usability problems. Each testing session lasted about 1 hour.

RESULTS

Usability Problems and Types

The five evaluators using Heuristic Evaluation found 103 problems in total (before aggregating). On average, each evaluator found 20 problems. After aggregation, 58 unique problems were identified. However, the subjects in User Testing only reported 10 usability problems. In order to study the problem types and severities, we used Nielsen's five-point severity-rating scale (Nielsen, 1994c) to evaluate the problems found by both methods. The scores (0-4) stand for *false alarm, cosmetic problem, minor usability problem, major usability problem,* and *usability catastrophe* respectively. Results showed that among the 58 problems identified by Heuristic Evaluation, there were 5(8.7%) major problems, 18(31%) minor problems, 18(31%) cosmetic problems, and 17(29.3%) false alarms. The ten usability problems found by User Testing were classified into two types: 7(70%) major problems and 3(30%) minor problems; no problem was rated as a false alarm. Table 3 listed some examples found by Heuristic Evaluation and User Testing.

More User Testing Results

The quantitative data collected in this study includes task completion times and user satisfaction (from the surveys). Table 4 summarized task completion times according to tasks and user groups. T-Test showed that no significant difference was found in learning-time or task-time between teachers and students, or between males and females. The average time to complete a search task was 2.24 min (SD= 0.87). These results indicated that SMELT software is easy to learn and use for both teachers and students.

Table 3. Some examples of the problems identified

Problem Type	Heuristic Evaluation	User Testing
Major	"It is difficult to switch to SMELT from other application when multiple applications are running simultaneously because there is no software icon on the task bar "	" Confused by the difference between ***Target*** and ***User***"
Minor	"User's Guide doesn't explain the function of ***Accurate Search***"	"Database is too small"
Cosmetic	"The term ***More Details*** is not explained in User's Guide"	NA
False Alarm	"I prefer green background."	NA

Table 4. Analysis of task completion times

	Learning	Test	Task1	Task2	Task3	Task 4
Students	4.68 (1.92)	2.78 (1.60)	3.52 (1.78)	2.89 (1.64)	1.07 (0.36)	2.20 (0.93)
Teachers	6.21(2.26)	3.95 (0.14)	2.84 (0.57)	2.11 (0.47)	1.18 (0.22)	2.11 (0.48)
Male	5.99(2.35)	4.10 (0.36)	2.74 (0.62)	2.02 (0.50)	1.07 (0.17)	1.98 (0.49)
Female	5.08(1.96)	2.64 (1.40)	3.62 (1.71)	2.98 (1.58)	1.18 (0.41)	2.34 (0.88)
Total	5.54(2.10)	3.37 (1.23)	3.18 (1.30)	2.50 (1.21)	1.13 (0.30)	2.16 (0.70)

Table5. Break down analysis of user satisfaction

Dimensions	Teacher	Student
Part one (9-point scale)		
Overall	6.36 (0.52)	6.88 (1.79)
Learning	7.12 (0.97)	7.84 (1.01)
Terminology	7.32 (0.90)	7.08 (0.33)
Screen	7.50 (1.38)	7.75 (1.06)
System Capabilities	7.20 (1.37)	7.30 (0.84)
Part two (7-point scale)		
Learning time	4.60 (1.14)	5.40 (1.52)
Expectation	6.40 (0.55)	6.60 (0.89)
User's guide	3.60 (1.14)	4.20 (1.10)
Help	4.00 (0.71)	4.20 (0.45)
Quality of results	5.20 (0.84)	4.80 (1.64)
Description	5.20 (1.30)	5.00 (1.00)
Memory effort	4.80 (1.79)	4.80 (1.92)
Search effort	5.80 (0.84)	5.20 (1.30)

The reliability of our usability survey was fairly good (a=0.91, 32 items) and comparable to the original QUIS 5.0 survey (a= 0.94, 27 items). For this study, we divided our survey into two parts: the first part focuses on general criteria such as *Overall impression, Learning, Terminology, Screen and System Capabilities;* while the second part deals with more specific features including *Learning time, User's expectation, User's guide, Help, Quality of search results, Description, Memory effort and Search effort.* In general, users are fairly satisfied with SMELT software. They rated SMELT software 7.18 in the first part (9-point scale) and 4.99 in the second part (7-point scale). However, we also noticed that the overall rating in the first part was relatively lower than other criteria. It was the *User's guide* (3.9 out of 7) and Help (4.1 out of 7) that affected user's overall rating. A break down analysis is provided in the Table 5.

Cost comparisons

Although all the subjects participated in both usability tests were volunteers, we still found that Heuristic Evaluation was cheaper than User Testing in terms of direct cost and time cost. The direct office supply costs including testing materials and gifts for testing users were: $10.54 for Heuristic Evaluation; $47.30 for User testing. It only took experimenter 15.5 hours to conduct the HE including data analysis, but User Testing need 45 hours to complete the whole process.

DISCUSSION

Looking at the results of the cost analysis, Heuristic Evaluation seems to be a more appealing usability technique than User Testing. The Heuristic Evaluation ended up with less expense and less time invested than the User Testing technique. Accordingly, Heuristic Evaluation appears to be more economic in finding a wide range of usability problems by incurring a low cost in comparison to User Testing. Nielsen (1993b) once pointed out user testing was 4.9 times as expensive as the cheapest heuristic method but provided better performance estimates. The cost comparison of this study supports Nielsen's conclusion.

According to Cost Comparison, Heuristic Evaluation was found to be more *efficient* than User Testing, identifying a larger number of problems for a smaller cost in both time and money. However, as graduate students, the evaluators did not demand the salary that an experienced, professional usability expert would be paid. Our results, therefore, may not be particularly accurate in an off-campus setting. We also found some problems while using Heuristic Evaluation to detect usability problems. First, it yields a significant false alarm rate. Second, once a person becomes an expert, he/she will not think and behave like a novice user. Thus, some problems associated with learning are hard to be identified. For example, the test users in User Testing identified two major problems that had not been identified by Heuristic Evaluation. Based on observations made during this study, Heuristic Evaluation is a useful testing method in the earlier stages of software development. It was found that Heuristic Evaluation is a quick method and identifies a wide range of problems. As is widely known, it is easier and cheaper to correct errors in earlier design stages than in the later design stages. We recommended that software designers and developers use Heuristic Evaluation to eliminate as many usability problems as possible in the earlier stages. When a functional prototype of the software is available, User Testing should be considered as major testing method because it can capture real usability problems from the real users. This can help software development team to better meet the needs and expectations of the user population before release of the software.

ACKNOWLEDGMENTS

Portions of the work presented in this paper were supported by grants from the National Institute for Science Education (funded by NSF) and the Wisconsin Space Grant Consortium (funded by NASA), while the authors were at the University of Wisconsin-Madison. The opinions and findings represent the authors' own perspectives, and do not reflect official positions by these or any other agencies.

REFERENCES

Bailey, W. R. (1996). *Human performance engineering: Designing high quality, professional user interfaces for computer products, applications, and systems* (3rd ed.). Upper Saddle River, NJ: Prentice Hall PTR.

Chin, J.P., Diehl, V.A., & Norman, K.L. (1988). Development of an instrument measuring user satisfaction of the Human-Computer Interface. *ACM CHI'88 Proceedings*, 213-218.

Lin, X. H., Choong, Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology,* 16 (4/5), 267-278.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proc. ACM CHI'90 Conf.,* 249-256.

Nielsen, J. & Phillips (1993a). Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared. *Proc. ACM CHI'93 Conf.,* 214-221.

Nielsen, J. (1993b). *Usability Engineering*, Boston, MA: Academic Press.

Nielsen, J. (1994a). Usability inspection methods. *Proc. ACM CHI'94,* Boston, Massachusetts USA, April 24-28.

Nielsen, J. (1994b), *Jakob Nielsen's Online Writings on HE.* Retrieved January 16, 2000, from http://www.useit.com/papers/heuristic/.

Nielsen, J. (1994c). Enhancing the explanatory power of usability heuristics. *CHI'94 Con.,* 152-155.

Rubin, J. (1994). *Handbook of Usability Testing: How to plan, design, and conduct effective tests.* New York: John Wiley & Sons, Inc.

Wang, E., & Caldwell, B.S. & Zhang K. (2002). Time delay tolerance in Computer Supported Cooperative Work. *Proceeding of 6th International Scientific Conference on Work With Display Unit,* 199-201.

Wang, E., Caldwell, B.S. & Salvendy, G (2002). Usability Comparison: Similarity and differences between E-commerce and World Wide Web. *Journal of Chinese Industrial Engineering.* Accepted for Publication.