Integrating Usability Design and Evaluation: Training Novice Evaluators in Usability Testing

Mikael B. Skov and Jan Stage
Department of Computer Science
Aalborg University
Aalborg Øst, Denmark
+45 9635 8080
{dubois, jans}@cs.aau.dk

ABSTRACT
This paper reports from an empirical study of training of usability testing skills. 36 teams of novice evaluators with an interest but with no education in information technology were trained in a simple approach to web-site usability testing that can be taught in less than one week. The evaluators were all first-year university students. The paper describes how they applied this approach for planning, conducting, and interpreting a usability evaluation of the same web site.

We discover that basic usability testing skills can be developed. The student teams gained competence in defining good task assignments and ability to express the problems they found. On the other hand, they were less successful when it came to interpretation and analytical skills. They found quite few problems, and they seemed to lack an understanding of the characteristics that makes a problem list applicable.

Keywords
Usability test, training novices, dissemination of usability skills

INTRODUCTION
Despite several years of research on usability testing and engineering, many computer-based information systems still suffer from low usability [4]. One problem arises from the fact that planning and conducting full-scale usability tests yields key challenges of e.g. user integration [7]. Considerable costs arise when a large group of users is involved in a series of tests. Furthermore for some applications it is difficult to recruit prospective test subjects [2].

The theoretical usability evaluation approach denoted as heuristic inspection evolved as a creative attempt to reduce such costs of usability evaluations [5, 6, 8]. The idea in heuristic inspection is that an interface design is evaluated by relating it to a set of guidelines, called heuristics [8]. The aim of the heuristics is to equip people who are not usability specialists to conduct heuristic inspections. Some of the empirical studies of the approach have been based on university students or readers of a computer magazine who act as evaluators [8].The idea behind heuristic inspection is to accomplish a simplified way of conducting usability tests. However, the empirical results indicate that we move the problem from finding users to finding user interface specialists. For a small organization developing web-based systems both of these problems may be equally hard to overcome. On a more general level the relevance of heuristic inspection can also be questioned. It has been argued that real users are an indispensable prerequisite for usability testing. If they are removed, it is at the expense of realism [10].

In this paper, we pursue a different idea of enhancing the knowledge of usability for software designers. One key problem in improving the usability of systems is the challenges involved in the interplay between the design and the evaluation of the system. Sometimes these activities are separated and detached making the interplay difficult and challenging e.g. one potential problem arises from the fact that designers and evaluators do not share a common language or set of tools in order to communicate. Our study explores how we can enhance usability testing competences for novice evaluators. Our aim is to train novice evaluators and compare their usability testing performances against the performances of professional usability testing labs. For our study, we use first-year university students as novice evaluators. First, we outline the taught usability testing approach and present the experiment behind the paper. Secondly, we compare the performances of the novice evaluators to the performances of professional labs on 17 different variables. Finally, we discuss and conclude our study.

METHOD
We have made an empirical study of the usability approach that was taught to the novice evaluators.

Usability Testing Approach
The approach to usability testing was developed through a course that was part of a curriculum for the first year at Aalborg University, Denmark. The overall purpose of the course was to teach and train students in fundamentals of

usability issues and testing. The course included ten class meetings each lasting four hours that was divided between two hours of class lectures and two hours of exercises in smaller teams. All class meetings except for two addressed aspects of usability and testing. The course required no specific skills within information technology that explains the introduction of course number one and five. The purpose of the exercises was to practice selected techniques from the lectures. In the first four class meetings, the exercises made the students conduct small usability pilot tests in order to train and practice their practical skills. The last six exercises were devoted to conducting a more realistic usability test of a specified web site.

The course introduced a number of techniques for usability testing. The first one was the technique known as the think-aloud protocol, which is a technique where test subjects are encouraged to think aloud while solving a set of tasks by means of the system that is tested, cf. [7]. The second technique is based on questionnaires that test subjects fill in after completing each task and after completion of the entire test, cf. [11]. Additional techniques such as interviewing, heuristic inspection, cognitive walkthroughs, etc. were additionally briefly presented to the students.

The tangible product of the usability evaluation should be a usability report that identifies usability problems of the product, system, or web site in question. We proposed to the students that the usability report should consist of 1) an executive summary (1 page), 2) description of the applied methodology (2 pages), 3) results of the evaluation (5-6 pages), and 4) a discussion of the applied methodology (1 page). Thus, the report would typically integrate around 10 pages of text. It was further emphasized that the problems identified should be categorized, at least in terms of major and minor usability problems. In addition, the report should include all data material collected such as log-files, tasks for test subjects, questionnaires etc.

Web-Site

Hotmail.com was chosen as object for our study mainly for two reasons. First, hotmail.com is one of the web-sites that provides advanced features and functionalities appropriate for an extensive usability test. Furthermore, hotmail.com facilitates evaluations with both novice and expert test subjects due to its vast popularity. Secondly, hotmail.com has been of focus in other usability evaluations and we compare the results of the student teams in our study with other results on usability evaluations of hotmail.com (further explained under Data Analysis).

Subjects

The subjects were all first-year university students enrolled in four different studies at the faculty for natural sciences and engineering at Aalborg University; the four studies were architecture and design, informatics, planning and environment, and chartered surveyor. None of the subjects indicated any experiences with usability tests prior to the study.

36 teams involving a total of 234 subjects (87 females, 37%) participated in our study of which 129 (55%) acted as test subjects, 69 (30%) acted as loggers, and 36 (15%) acted as test monitors, cf. [10]. The average subject age was 21.2 years old (SD=1.58) and the average team size was 6.5 subjects (SD=0.91). The average size of number of test subject in the teams was 3.6 subjects (SD=0.65). 42 (33%) of the 129 test subjects had never used hotmail.com before the conduction of test, whereas the remaining 86 subjects had rather varied experience.

Procedure

The student teams were required to apply the techniques presented in the course. Additionally, each team was required to select among themselves the roles of test subjects, loggers, and test monitor.

The test monitor and the loggers received after the second lecture a two-page scenario specifying the web-based mail service www.hotmail.com as the object of focus in the test. The scenario also specified a comprehensive list of features that emphasized the specific parts of www.hotmail.com they were supposed to test. The test monitor and the loggers would then start to examine the system, design tasks, and prepare the test in general, cf. [10]. The www.hotmail.com web site in the study was kept secret to test subjects until the actual test conduction.

30 (83%) of the 36 teams provided information on task completion times for 107 (83%) of the 129 subjects resulting in an average session time of 38.10 minutes (SD=15.32 minutes). Due to the pedagogical approach of the university, each team was allocated their own offices equipped with a personal computer and Internet access. Most teams conducted the tests in these offices. After the tests, the entire team worked together on the analysis and identification of usability problems and produced the usability report.

Data Analysis

The 36 usability reports were the primary source of data for our empirical study. The 36 reports had an average size of 11.36 pages (SD=2.76) excluding the appendences, which had an average size of 9.14 pages (SD=5.02). All reports were analyzed, evaluated, and marked by both authors of this paper according to the following three steps.

1) We designed a scheme for the evaluation of the 36 reports by analyzing and evaluating five randomly selected reports from the 36 reports. Through discussions and negotiations we came up with an evaluation scheme with 17 variables as illustrated in table 3. The 17 variables was divided into three overall categories of evaluation (relates the conduction of the test), report (relates the presentation of the test and the results), and results (relates the results and outcome of the usability test). Finally, we described, defined, and illustrated all 17 variables in a two-page marking guide.

2) We worked individually and marked each report in terms of the 17 variables using the two-page marking guide. The

Team	Conduction			Documentation					
	Test procedure conduction	Task quality and relevance	Question-naire / Interviews	Test description	Data quality	Clarity of problem list	Executive summary	Clarity of report	Layout of report
Student (N=36)	3.42 (0.73)	3.22 (1.05)	2.72 (1.00)	3.03 (0.94)	3.19 (1.33)	2.53 (1.00)	2.39 (0.80)	2.97 (0.84)	2.94 (0.89)
Professional (N=8)	**4.38 (0.74)**	3.13 (1.64)	3.50 (1.69)	**4.00 (1.31)**	2.13 (0.83)	**3.50 (0.93)**	**3.38 (1.06)**	**4.25 (0.71)**	3.25 (0.71)

Team	Results							
	Number of problems*	Problem categorization	Practical relevance	Qualitative results overview	Quantitative results overview	Use of literature	Conclusion	Evaluation of test
Student (N=36)	2.56 (0.84)	2.06 (1.22)	3.03 (1.00)	3.03 (1.00)	2.28 (1.14)	3.08 (0.81)	2.64 (0.90)	2.44 (1.08)
Professional (N=8)	**4.13 (1.13)**	3.25 (1.75)	**4.25 (1.49)**	**3.75 (1.16)**	2.00 (1.51)	3.13 (0.35)	**3.88 (0.64)**	2.88 (1.13)

Table 3. Mean values and standard deviation (in parentheses) of all 17 variables for the student and professional teams. The grade for the number of identified problems is calculated from the actual number of identified usability problems in each usability report according to the following procedure: 1 = 0-3 identified problems; 2 = 4-7 identified problems; 3 = 8-12 identified problems; 4 = 13-17 identified problems; 5 = >17 identified problems. Boldfaced numbers indicate significant differences between the student and professional teams.

markings were made on a scale of 1 to 5 (1=no or wrong answer, 2=poor or imprecise answer, 3=average answer, 4=good answer, and 5=outstanding answer). We furthermore counted the number of identified usability problems in all 36 usability reports. In our study, we define a usability problem as the prevention or impediment of realization of user objectives through the interface. Furthermore, we specified limits for grading afterwards based on their distribution on the scale (1=0-3 problems, 2=4-7 problems, 3=8-12 problems, 4=12-17 problems, and 5>17 problems).

3) All reports and evaluations were compared and a final evaluation on each variable was negotiated. In case of disagreements on marking, we pursued the following two-folded procedure - 1) if the difference was equal to one grade we would renegotiate the grade based upon our textual notes 2) if the difference was equal to two grades, we would reread and reevaluate the report in a collaborative effort focusing only on the corresponding variable. For our study, no disagreement exceeded more than two grades.

To examine the overall performance of the students, we included two additional sets of data in the study. First, we compared the student reports to usability reports produced by teams from professional laboratories. These reports were selected from a pool of usability reports produced in another research study where nine different usability laboratories received the same scenario as outlined above and conducted similar usability tests of www.hotmail.com, cf. [2]. Of these nine usability reports, we dropped one due to its application of only theoretical usability evaluation techniques, e.g. heuristic inspection, thereby not explicitly dealing with the focus of our study namely user-based testing techniques. The remaining eight usability reports were analyzed, evaluated, and marked through the same procedure as the student reports. We analyze the data using Mann-Whitney U Test for testing the significance between means for all 17 variables

RESULTS

The general impression of the results as outlined in table 3 suggests that the professional laboratories performed better than the student teams on most variables. However, on three, e.g. 2), 5), 14), of the 17 variables the student teams actually performed best, whereas on the remaining 14 variables the professional teams on average did better and for six variables, e.g. 1), 8), 10), 11), 12), 16, the professional teams were marked one grade (or more) higher than the students.

Conducting and Documenting the Usability Test

Test conduction relates the actual conduction of the usability test. The professional teams have average of 4.38 (SD=0.74) almost one grade higher than the student teams and a Mann-Whitney U Test shows strong significant difference between test conduction of the student teams and test conduction of the professional teams ($z=-2.68$, $p=0.0074$). On the other hand, even though the students performed slightly better on the quality and relevance of tasks, this difference is not significant ($z=0.02$, $p=.984$). Finally, no significant variation was found for the questionnaires and interview guidelines quality and relevance ($z=-1.63$, $p=0.1031$).

Concerning presentation of the usability testing results, the professional teams did better than the student teams on clarity of the usability problem list and we found strong

significant variance on this variable (z=-2.98, p=0.0029) and we also found strong significant difference on the clarity of the entire report (z=-3.15, p=0.0016). Further, there is significant difference on the teams' description of the test (z=-2.15, p=0.0316) and on the executive summary (z=-2.27, p=0.0232). The student teams actually performed significantly better than the professional teams on the quality of the data material in the appendix (z=2.07, p=0.0385). Finally, no significance was identified for the layout of the report (z=-1.02, p=0.3077).

Identification and Categorization of Test Results

The pivotal results of all student and professional usability reports were the identification (and categorization) of various usability problems. However, the student and professional teams performed rather differently on this issue. The student teams were on average able to identify 7.9 usability problems (in the marking scale: Mean 2.50, SD 0.88) whereas the professional teams on average identified 21.0 usability problems (in the marking scale: Mean 4.13, SD 1.13) and a Mann-Whitney U Test confirms strong significance (z=-3.09, p=0.002). However, the professional teams actually performed rather dissimilar identifying from seven to 44 usability problems.

The student teams provided better overview of the quantitative results, but this difference was not significant (z=0.90, p=0.3681). On the hand, the practical relevance of the identified usability problems was significantly higher for the professional teams (z=-2.56, p=0.0105). Furthermore, the conclusion are better in the professional team reports and this difference was strong significant (z=-3.13, p=0.0017). The overview of the qualitative results also showed significant variance (z=-1.99, p=0.0466). No significance was found for the problem categorization (z=-1.84, p=0.0658), the use of literature (z=-0.05, p=0.9601), or the evaluations of the test procedure (z=-1.00, p=0.3173).

DISCUSSION

Our aim with this study was to explore dissemination of usability testing skills to people with no formal training in information technology design or use. Previous studies have suggested heuristic inspection as a creative attempt to reduce costs of usability evaluations. Research has shown that planning and conducting full-scale usability tests yields key challenges of e.g. user integration [7]. Considerable costs arise when a large group of users is involved in a series of tests. Further, for some applications it is difficult to acquire prospective test subjects [2]. However, user-based evaluations may provide more valid results.

Our study documents experiences from a course with 234 students that conducted a usability test of hotmail.com in teams of four to eight students. The results of these tests were documented in 36 individual usability reports. Our study reveals a number of interesting issues to consider when novices are to conduct full-scale user-based usability evaluations.

One key finding of our study is characteristics of usability problem identification (and categorization). The student teams are only able to identify significantly fewer problems than the professional teams. A key aim in usability testing is to uncover and identify usability problems, and the student teams on average found 7.9 usability problems whereas the professional teams on average found 21 usability problems. The student teams perform rather differently on this variable as one team identify no problems (it seems this team misunderstood the assignment) to two teams identifying 16 problems. Most of the teams identify no more than 10 problems. The professional teams also perform rather differently and this is perhaps more surprising where one team identify 44 problems and one team identify only seven problems. The latter is actually rather disappointing for a professional laboratory. We are in process of analyzing the severity of the problems and we do not have any results on this issue so far.

Related the conduction of the usability test sessions, the majority of student teams score 4, which indicates well-conducted tests with a couple of problematic characteristics. The average on 3.43 also reflects the general quality of the test processes. The professional laboratories score an average of 4.6 on this factor, and 6 out of 8 score the top mark. This is as it should be expected because experience will tend to raise this variable. However, the student teams perform rather well with respect to planning and conducting the usability testing sessions. On the other hand, there seems to be no direct correlation between the quality of the test conduction or the quality of the assigned tasks and the number of identified problems. Thus, the students may plan their evaluations carefully, but

Another variable that exhibits a difference is the practical relevance of the problem list, cf. figure 5. The student teams are almost evenly distributed on the five marks of the scale, and their average is 3.2. Yet when we compare these to the professional laboratories, there is a clear difference. The professionals score an average of 4.6 where 6 out of 8 laboratories score the top mark. This difference can partly be explained from the experience of the professionals in expressing problems in a way that make them relevant to their customers. Another source may be that the course has focused too little on discussing the nature of a problem; it has not been treated specifically with examples of relevant and irrelevant problems.

Our study is limited in a number of different ways. First, the environment in which the tests were conducted was in many cases not optimal for a usability test session. In some cases, the students were faced with slow Internet access that influenced the results. Second, motivation and stress factors could prove important in this study. None of the teams volunteered for the course (and the study) and none of them received any payment or other kind of compensation; all teams participated in the course because

it was a mandatory part of their curriculum. This implies that students did not have the same kinds of incentives for conducting the usability test sessions as people in a professional usability laboratory. Thirdly, the demographics of the test subjects are not varied with respect to age and education. Most test subjects were a female or a male of approximately 21 years of age with approximately the same school background and recently started on a design-oriented education. The main difference is the different curricula they follow. Fourthly, the hotmail.com website is a general website in the sense it provides no or little domain knowledge. Different distributions on the variable may emerge for more specialized user interfaces, see [2] for examples.

CONCLUSION

The existing low level of skills in usability engineering among web-site development teams is likely to prohibit moves towards the ideal of universal access and the idea of anyone, anywhere, anytime. This article has described a simple approach to usability testing that aims at quickly teaching fundamental usability skills to people without any formal education in software development and usability engineering. Whether this approach is practical has been explored through a large empirical study where 36 student teams have learned and applied the approach.

The student teams gained competence in two important areas. They were able to define good tasks for the test subjects, and they were able to express the problems they found in a clear and straightforward manner. Overall, this reflects competence in planning and writing. The students were less successful when it came to the identification of problems, which is the main purpose of a usability test. Most of the teams found too few problems. It was also difficult for them to express the problems found in a manner that would be relevant to a practicing software developer.

The idea of this approach is to reduce the efforts needed to conduct usability testing. This is consistent with the ideas behind heuristic inspection and other walkthrough techniques. On a more general level, it would be interesting to identify other potential areas for reducing effort.

This approach to usability testing did provide the students with fundamental skills in usability engineering. Thus it is possible to have usability work conducted by people with primary occupations and competencies that are far away from software development and usability engineering. We see the approach as a valuable contribution to the necessary development emphasized here: "Organizations and individuals stuck in the hierarchies and rigidity of the past will not foster what it takes to be successful in the age of creativity, the age of the user, and the age of the Internet economy" [1].

ACKNOWLEDGMENTS

We would like to thank the participating students in the study. In addition, we would like to thank the anonymous reviewers for comments for earlier drafts.

REFERENCES

1. Anderson, R. I. Making an E-Business Conceptualization and Design Process More "User"-Centered. *interactions 7*, 4 (July-August), 27-30.
2. Kjeldskov, J. and Skov, M. B. (2003) Evaluating the Usability of a Mobile Collaborative System: Exploring Two Different Laboratory Approaches. In Proceedings of the 4th International Symposium on Collaborative Technologies and Systems, pp. 134 - 141
3. Kjeldskov, J. and Skov, M. B. (2003) Creating Realistic Laboratory Settings: Comparative Studies of Three Think-Aloud Usability Evaluations of a Mobile System. In Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction (Interact2003), IOS Press, pp. 663 – 670.
4. Molich, R. *Comparative Usability Evaluation Reports*. Available at http://www.dialogdesign.dk/cue.html.
5. Molich, R., and Nielsen, J. Improving a Human-Computer Dialogue. *Comm. ACM 33*, 3, 338-348.
6. Nielsen, J. Finding Usability Problems Through Heuristic Evaluation. In *Proceedings of CHI '92*, ACM Press, 373-380.
7. Nielsen, J. *Usability Engineering*. Morgan Kaufmann Publishers, 1993.
8. Nielsen, J., and Molich, R. Heuristic Evaluation of User Interfaces. In *Proceedings of CHI '90*, ACM Press, 249-256.
9. Rohn, J. A. The Usability Engineering Laboratories at Sun Microsystems. *Behaviour & Information Technology 13*, 1-2, 25-35.
10. Rubin, J. *Handbook of Usability Testing. How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons, 1994.
11. Spool, J. M., Scanlon, T., Schroeder, W., Snyder, C., and DeAngelo, T. *Web Site Usability. A Designer's Guide*. Morgan Kaufmann Publishers, 1999.
12. Sullivan, T., and Matson, R. Barriers to Use: Usability and Content Accessibility on the Web's Most Popular Sites. In *Proceedings of Conference on Universal Usability* (Washington, November 2000), ACM Press, 139-144