Detection of Web-Site Usability Problems: Empirical Comparison of Two Testing Methods

Mikael B. Skov and Jan Stage

Department of Computer Science
Aalborg University
Fredrik Bajers Vej 7
9220 Aalborg East, Denmark
{dubois, jans}@cs.auc.dk

SUMMARY
This article reports from an empirical study of web site usability testing. The study is used to compare two usability-testing methods that have a prominent position in the literature on web-site usability. We have denoted these as the think-aloud method and the questionnaire method. The empirical study involved 36 teams of four to eight university students who were taught the two methods and applied one of them for conducting a usability test of a commercial web-site. The results of the study shows that the application of the two usability methods gives rise to different results with respect to the detection of usability problem. It is concluded that the questionnaire method depends more on individual skills of the people conducting the test since more teams were only able to detect none or very few problems. The think-aloud teams found on average more problems than the questionnaire teams.

KEYWORDS: Usability testing methods, interface evaluation methodologies, empirical study, World Wide Web

INTRODUCTION
Usability engineering is characterized by numerous methods for conducting usability tests. Key textbooks describes a broad variety of methods such as usability testing, usability audits, expert evaluations, thinking aloud, participatory design, focus groups, surveys based on questionnaires, open interviews, cognitive walkthroughs, heuristic inspections, field studies, logging, user feedback, observation, cf. [8, 9]. This variety in terms of methods complicates both teaching and practice. Which method should we apply in a certain situation? What methods should we teach and how can we relate them to each other?

The area of usability engineering is also characterized by a rich variety of attempts to answer these questions. Some approaches provide overviews where a selection of methods are presented and compared in terms of some determinants. Examples of key determinants are: the point in time compared to a complete project lifecycle, or the need to involve users [8, 9]. These overviews are helpful but it is often not clear to what extent they are normative statements or based on empirical evidence.

Other contributions employ detailed empirical studies to provide evaluations of specific usability testing methods. Examples of this are studies of the extent to which the outcome of using one method is affected by test monitor/evaluator [3, 5]. Others have compared different methods in terms of certain factors. Here the idea is that in a given situation, these factors might be used as determinants for selecting a method. Typical examples of this are comparisons of tests with and without users [1,4]. Other contributions focus on different user-based test methods [2].

When we focus on web-site usability test, the amount of guidelines for selecting usability testing methods is much more limited. However, there are examples of normative overview [7] where the primary method is thinking aloud. There are also empirical investigations [11] where the method is based on questionnaires in order to reduce the extent to which the test monitor can influence the outcome. Yet none of these attempt to make an empirically based comparison.

This article describes how we designed and conducted an empirical study of methods for web-site usability. In the study, we have compared two of the methods that have been suggested for testing web-site usability: think-aloud and questionnaires. The following section presents the two testing methods in detail and how the empirical study was planned and conducted. The section that follows then presents the results by comparing the effectiveness of the two methods in terms of the number of problems detected. Finally, we discuss the results and the motivation behind the study design and provide a conclusion to the questions raised above.

METHOD

The Usability Evaluation Methods
The two methods we have compared were denoted as the think-aloud method and the questionnaire method. Below, we summarize their specific characteristics.

The think-aloud method implies that test subjects are encouraged to think aloud while solving a set of tasks by means of the system that is tested [8]. The challenge with this method is to make good tasks and to act consistently as a test monitor. On these issues, we have employed specific guidelines from the literature [6, 7, 9]. The test procedure is that each subject enters the test room, receives a short introduction, is presented with a sheet de-scribing the tasks to be solved, thinks aloud while solving the tasks, and is interviewed after the completion of all tasks. During a usability test session, one or more loggers take notes about the things that the test subject expresses and the problem he or she faces.

The data analysis after using the think-aloud method focuses on the problems that were expressed explicitly by the test subjects. In addition, there will be problems that are implicitly expressed by the test subject. Both categories are listed and described as the main result.

The questionnaire method implies test that subjects fill in a questionnaire after completing each task and after finishing the entire test [11]. The tasks are also a key challenge with this method. The other important point is the design of the questionnaires. Here the test monitor is likely to have less influence on the outcome. The test procedure is that each subject enters the test room, receives a short introduction, is presented with a sheet describing the tasks to be solved, solves each task quietly, fills in a questionnaire about each specific task immediately after completing it, and fills in another questionnaire after having completed all tasks. The test monitor is only there to provide assistance if a test subject asks for it. During a test one or more loggers take notes about the actions of the test subject and the problem he or she seems to face.

The data analysis after using the questionnaire method focuses on the way test subjects experienced the usability of a web site for solving a task. The post-task questionnaires deal specifically with the mental state of the test subjects. The analysis can also involve the observations made by the loggers. The result is a list of usability problems.

The two methods were combined with general techniques for test planning, interviewing, questionnaire design, etc. that originate from [6, 7, 9]. It should be noted that the two methods could be combined in a usability test session but for this study, they primarily define the main procedure for collecting data.

Empirical Study

We designed and conducted an empirical study to compare the relative strengths and weaknesses of the two usability-testing methods, e.g. think-aloud and questionnaires, as described above.

The study was conducted in connection with a course that is part of a curriculum for the first year at Aalborg University, Denmark. The title of the course is use of information technology, and the overall purpose is to teach and train students in fundamentals of computerized systems with a particular emphasis on usability issues. 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.

The ten class meetings compromised the following topics: #1 introduction and computer networks; #2 usability issues: guidelines and measurement; #3 usability testing: think-aloud method; #4 usability testing: questionnaire method; #5 computer architecture; #6 usability testing: data analysis; #7 usability issues: techniques and documentation; #8 web-site: usability; #9 web-site: orientation and navigation; and #10 web-site: visual design. Thus the two methods were presented at class meeting #3 and #4. All class meetings, except number one and five, addressed aspects of usability and web-sites. 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 smaller usability pilot tests in order to train and practice their practical skills. The exercises of the last six class meetings were devoted to conducting a complete usability test of a web-site.

The empirical study related to the course involved 36 teams of first year university students who conducted a usability evaluation of the email services at the Hotmail web site (`http://www.hotmail.com/`). The 36 teams consisted of students from such diverse educations as architecture and design, informatics, planning and environment and chartered surveyor. These studies are all part of a natural science or engineering program at Aalborg University. Figure 1 describes the teams that participated in the study.

Total number of students	Total number of teams	Team size *Average*	Team size *Min / Max*
234	36	6.5	4 / 8
Number of test subjects *Average*	**Number of test subjects *Min / Max***	**Age of test subjects *Average***	**Age of test subjects *Min / Max***
3.6	2 / 5	21,2	19 / 30

***Figure 1*:** Team and test subject data for the teams that participated in the empirical study.

Each student team was required to apply one of the two methods described above, and they were allowed to supplement this with other techniques according to their own choice. The distribution of teams on the two methods was made randomly when the course started by the authors of this article. Each team should among themselves choose a test monitor and a number of loggers (they were recommended to use two loggers), who should examine the system, design task assignments for

Method	Number of detected problems																				Average	Standard deviation
T	7	5	5	13	10	8	5	11	11	12	11	16	5	6	8	8	5	5	3		8,6	3,372
Q	3	11	9	3	4	10	0	13	10	7	14	11	6	11	5	1	12				7,6	4,358

***Figure 2*:** Usability problems detected by the teams employing either think-aloud (T) or questionnaires (Q).

the test subjects, and prepare the test. The rest of each team acted as test subjects, and the web site used for testing was kept secret to them until their test started.

All teams were also given a very detailed two-page scenario stating that they should conduct a usability test of the Hotmail web-site (`http://www.hotmail.com/`). The scenario included a list of features that emphasized the parts of Hotmail they were supposed to test. Each usability test session was planned to last approximately one hour. Due to the pedagogical approach of the university, each team has their own office. Most teams conducted the tests in this office, which was equipped with a personal computer and Internet access. After the test, the entire team worked together on the analysis and identification of usability problems and produced the usability report.

The tangible product of the usability evaluation was described as a usability report, that identifies usability problems on the web-site in question. It was suggested that a usability report should consist of an executive summary (1 page), description of the approach applied (2 pages), results of the evaluation (5-6 pages), and a discussion of methodology (1 page). It was also emphasized that the problems identified should be categorized, at least in terms of major and minor usability problems. In addition, a report should include all data material collected such as log-files, tasks for test subjects, questionnaires etc. A prototypical example of a usability report was given to the students for inspiration.

The usability reports were the primary source of data for our empirical study. All reports were analysed, evaluated, and marked by both authors of this paper. Firstly, we worked individually and marked a collection of reports in terms of 16 different factors. Secondly, these markings were compared and negotiated, a new factor was added, and the criteria for marking each of the 17 factors were specified. Thirdly, the authors individually marked all reports according to the 17 factors. Fourthly, all reports and evaluations were compared and a final evaluation on each factor was negotiated.

One factor was the number of usability problems that each group found. We went through their reports and noted all problems that were emphasized by the team considered. This produced an absolute number of problems found. In this article, we focus on that factor.

The specific conditions of this study limit its validity in a number of 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. Finally, the demographics of the test subjects are not varied with respect to age and education. Most test subjects are a female or 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 curriculum they are following.

RESULTS

This section presents and compares the results from the empirical study. The comparison is based on the number of problems that the teams were able to detect. The basic underlying data are listed in figure 2. The figure is divided vertically between think-aloud teams (T) and questionnaire teams (Q). The horizontal dimension provides the numbers of problems detected by each of the 36 teams applying either the think-aloud protocol or questionnaires. The order in which the teams are listed in the figure is arbitrary. The average number of problems detected and the standard deviation is illustrated for both methods.

The first impression is that the results exhibit a wide distribution ranging from one team detecting no problems at all to one team detecting 16 problems. Almost half of all teams (41.7%) were able to detect at least ten problems and six teams (16.7%) were only able to detect four problems or less

The 19 think-aloud teams found on average 8.6 problems. The specific number of problems ranges from 3 to 16 problems. The wideness of the distribution is also indicated by the standard variation being as high as 3.4. The 17 questionnaire teams found on average 7.6 problems ranging from 0 to 14 problems. Here, the standard deviation is even higher with a value of 4.4.

In comparing the two methods, we can start noticing that the think-aloud teams on average detected one more problem than the questionnaire teams, a difference of 13%. We can also notice that only 1 of the 19 think-

aloud teams (5.3%) detected less than five usability problems whereas 5 of the 17 questionnaire teams (35.3%) detected less than five problems. Thus at the lower end of the scale, much more questionnaire teams found very few problems. At the other end of the scale, the results are more similar.

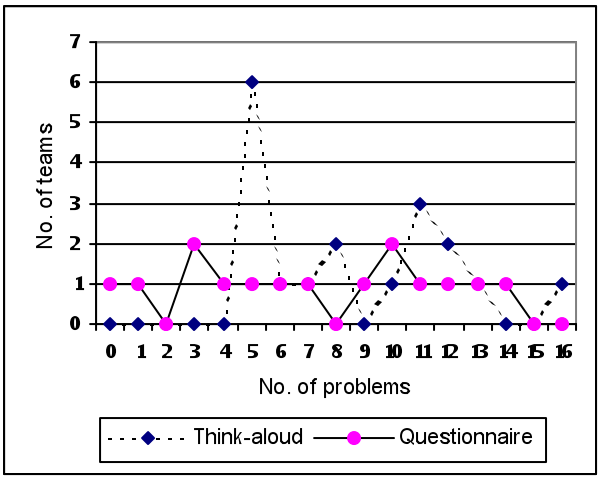

***Figure 3*:** Number of teams detecting precise number of usability problems.

The differences between the teams are illustrated in figure 3, where the number of teams detecting each specific number of problems is shown for both the think-aloud and questionnaire teams. The questionnaire teams are almost evenly distributed across the scale. In addition, no more than two questionnaire teams have detected the same number of problems, and 11 out of the 17 questionnaire teams (64.7%) detected a number of problems that is different from the other questionnaire teams. The think-aloud teams are distributed quite differently. They have 6 teams that detected exactly six usability problems, 3 teams detected twelve problems, and the remaining 10 teams are distributed on only seven numbers of problems.

These observations raise the question: what is the nature of these variations in the number of detected problems? In order to explore this further, we have grouped the number of problems detected, as illustrated in figure 4. Teams detecting between zero and two problems are summarized in one bullet, three to five problems in one bullet, and so forth.

The think-aloud teams have one peak at 3 to 5 detected problems and the number of teams detecting the subsequent numbers of problems decreases almost linearly. It is also illustrated that no think-aloud team performs really poor in the usability test since none of them detected zero to two problems. The questionnaire teams are distributed differently. Here, the figure exhibits two peaks. The first is around 4 problems, and the second peak is around 10 problems. The point illustrated in figure 4 is that none of the think-aloud teams completely missed the task of finding usability problems. That is different with the questionnaire teams. Very few of the questionnaire teams are distributed around the average. Finally, some of the questionnaire teams are doing so much better than the worst teams of this category.

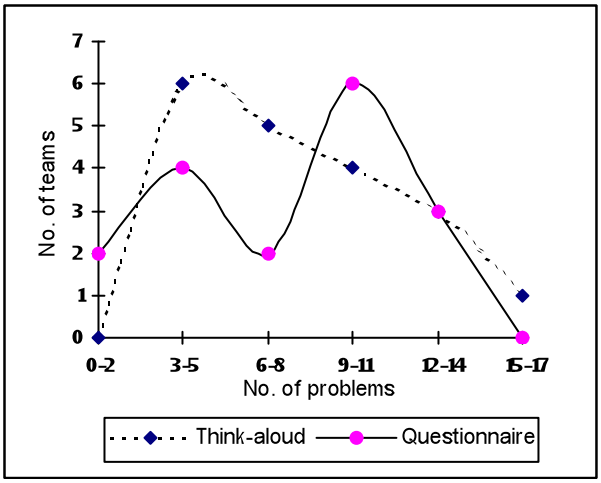

***Figure 4*:** Number of teams summarised on blocks of detected usability problems.

DISCUSSION

The results outlined above raise the question as to why some of the questionnaire teams are doing so exceptionally bad, and why many of the other questionnaire teams are doing so much better. The empirical basis provides no direct answer to these questions, but some answers could be given by nature of the study design and by some of the results.

The empirical study of this paper focuses merely on the number of detected problems by each team. In this paper, we have not discussed the kinds of problems detected and the different granularity in the description of these problems. The study seems to support the fact that the think-aloud protocol prohibits usability tests that completely fails to generate results in terms of problem detection in contrast to the teams applying questionnaires. No teams in the study detected zero to two problems when applying the think-aloud protocol. It seems as if the questionnaire teams are either good at designing appropriate questions in the questionnaire (detecting app. 10 problems) or else they are unable to design appropriate questions (detecting app. 4 problems). The think-aloud teams should not design questions for the test but could merely observe the users when they used and navigated the web site. This relates the ability to learn and train practical usability testing skills.

The aspect of training practical usability testing skills is discussed in another paper, cf. [10]. In this paper, we compare the results of the usability testing sessions by the students with usability tests of Hotmail conducted by professional usability laboratories. In this paper, we con-

clude that the students are actually able to construct and design a usability test session almost as well as professional usability laboratories. However, the students detect fewer problems on average and the practical relevance of the problems is significant lower than the problems described by the professional laboratories.

Finally, the web site Hotmail is quite well known with more than 100 million users. When conducting an empirical study like the one described in this paper, many test subjects may be familiar with the web site. The empirical data from our study showed most test subjects were familiar with or used Hotmail as mail provider. However, a significant number of test subjects did not use Hotmail and had never used it. The number of detected usability problems by the student teams and the professional usability laboratories indicate that there is still plenty of room for improvement. More test subjects faced problems when using Hotmail even though they used it on a regular basis. This can mainly be explained by the fact that the usability test sessions tested advanced features and functions of Hotmail something the test subjects were not familiar with.

CONCLUSION

Usability engineering includes a variety of different methods and techniques for evaluating the usability of computer-based information systems. Many of these methods are well established and are being applied in various settings for testing usability. Emerging technologies, e.g. the world-wide-web, challenge existing methods with respect to their general applicability. Today methods are been applied to evaluate their usefulness for testing the usability of web sites.

We have conducted an empirical study of an international well-known and heavily used web site where 36 teams applied two different usability methods namely think-aloud and questionnaires for detection of usability problems. Our study shows that the application of the two methods provides different results in terms of the number of detected usability problems. Firstly, none of the think-aloud teams performed really badly by detecting none or very few problems. On average, the think-aloud teams detected 8.6 problems ranging from three problems to 16 problems. On the other hand, some of the questionnaire teams performed pretty poor where one team detected no usability problems at all and another team detected only one problem. On average, the questionnaire teams detected 7.6 problems ranging from zero to 14 problems. In addition, the number of problems detected by the teams is differently distributed for the two methods. It seems that the questionnaire method forces usability testers to consult additional sources, e.g. log-files, in order to detect relevant usability problems.

Further research in this field is needed. First, we need to understand to exact nature of the detected problems of the various methods in order to be able to improve the methods. What are the characteristics of the detected problems, how are they described and what is their severity? This could be investigated in an in-depth empirical study of usability testing. Secondly, the comparison could be done by letting experienced usability professionals conduct the same test in order to identify similarities and differences between them and the students.

ACKNOWLEDGEMENTS

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.

BIBLIOGRAPHY

1. Bailey, R. W., Allan, R. W. and Raiello, P. Usability Testing vs. Heuristic Evaluation: A Head-to-Head Comparison. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, HFES, Santa Monica, 1992, pp. 409-413.
2. Henderson, R., Podd, J., Smith, M. and Varela-Alvarez, H. An Examination of Four User-Based Software Evaluation Methods. *Interacting with Computers*, Vol. 7, No. 4, 1995, pp. 412-432.
3. Jacobsen, N. E., Hertzum, M. and John, B. E. The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgments. In *Proceedings of the Human Factors and Ergonomics Society 42nd annual meeting*, HFES, Santa Monica, 1998, pp. 1336-1340.
4. Karat, C. M., Campbell, R. and Fiegel, T. Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In *Proceedings of CHI'92*, ACM, New York, 1992, pp. 397-404.
5. Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D. and Kirakowski, J. Comparative Evaluation of Usability Tests. In *Proceedings of the Usability Professionals Association Conference*, 1998, pp. 189-200.
6. Molich, R. *User-Friendly Computer Systems (in Danish)*. Teknisk Forlag, Copenhagen, 1994
7. Molich, R. *User-Friendly Web Design (in Danish)*. Teknisk Forlag, Copenhagen, 2000
8. Nielsen, J. *Usability Engineering*. Morgan Kaufmann Publishers, Inc., San Francisco, California, 1993
9. Rubin, J. *Handbook of Usability Testing – How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons, Inc., New York , 1994
10. Skov, M. B. and Stage, J. (2001) A Simple Approach to Web-Site Usability Testing. In *Proceedings of the first International Conference on Universal Access in Human-Computer Interaction* (to appear)

11. Spool, J. M., Scanlon, T., Schroeder, W., Snyder, C., & DeAngelo T. *Web Site Usability – A Designer's Guide*. Morgan Kaufmann Publishers, Inc., San Francisco, California, 1999