Supporting Web Developers in Evaluating Usability and Identifying Usability Problems

Mikael B. Skov and Jan Stage
Aalborg University, Department of Computer Science,
Fredrik Bajers Vej 7, DK-9220 Aalborg East, Denmark

ABSTRACT

Support to website developers without formal training in human-computer interaction that enable them to conduct their own usability evaluations would radically advance integration of usability engineering in web development. This chapter presents experiences from usability evaluations conducted by developers and results from an empirical study of means to support non-experts in identifying usability problems. A group of software developers who were novices in usability engineering analyzed a usability test session with the task of identifying usability problems experienced by the user. In their analysis they employed a simple one-page tool that has been developed to support identification of usability problems. The non-experts were able to conduct a well-organized usability evaluation and identify a reasonable amount of usability problems with a performance that was comparable to usability experts.

KEYWORDS

Usability evaluation, think-aloud, usability problem identification

INTRODUCTION

Over the last decade, software usability as a discipline has made considerable progress. An important indicator of this is that more and more software organizations are beginning to take usability seriously as an important aspect of development. Yet there are still significant obstacles to a full integration of usability engineering into software development (Bak et al., 2008). The average developer has not adopted the concern for usability, and evaluators are not being involved until late in development, when most substantial changes are too costly to implement (Anderson et al., 2001).

There are several areas of software development where the limited integration of usability efforts is apparent. Development of sites for the World Wide Web is one such area. It is usually argued that the web is qualitatively different from conventional software systems. For the typical web application, the user group is more varied and fluent, and it has a considerably shorter lifetime compared to other kinds of software. For web development, the main difference is that it is done by a broad variety of companies, ranging from one or two person companies to large corporations, and many of the development companies, in particular the smaller ones; do not have any usability experts available. Budget constraints prohibit hiring specialists, and the development schedule does not leave time for usability testing and feedback to iterative design (Scholtz et al., 1998). Research indicates that work practices in web-site development seem to largely ignore the body of knowledge and experience that has been established in the disciplines of software engineering, human-computer interaction, and usability engineering (Sullivan and Matson, 2000). Conventional usability evaluation is expensive, time consuming and requires usability experts. This is incompatible with web development, where many web sites are designed and implemented in fast-paced projects by multidisciplinary teams that involve such diverse professions as information architects, Web developers, graphic designers, brand and content strategists, etc. Such teams are usually not familiar with established knowledge on human-computer interaction (Braiterman et al., 2000). The consequence of this is clear. A large number of websites have severe usability problems that prohibit effective and successful use (Spool et al., 1999). An investigation of usability through content accessibility found that 29 of 50 popular web sites were either inaccessible or only partly accessible (Spool et al., 1999; Sullivan and Matson, 2000).

At least two ways exist for organizing usability expertise in website development projects. First, developers can adapt and use tailored usability heuristics in the evaluation and let these heuristics guide the usability work in the development team (Agarwal and Venkatesh, 2002; Sutcliffe, 2001). The practical implications of usability heuristics in software design have been discussed for several years, but traditional heuristics is not of focus in this paper. Secondly, a possible solution to the limited integration of usability in software development is to involve non-experts in the usability engineering activities. This could be accomplished by offering ordinary software developers means for creating usable web sites and for evaluating them in a systematic manner (Skov and Stage, 2001). This would bring usability into the earliest possible phases of software development where it could have most impact by improving initial design and eliminating rework. It would also solve a potential problem with availability of usability experts. The professional evaluator resource is very scarce. Evaluating the usability of just a fraction of all new web sites would be well beyond their capacity.

This chapter presents an empirical study of a specific means to support non-experts in web usability in conducting a web site usability evaluation. We have explored to what extent a simple one-page usability problem identification tool can support and stimulate the analytical skills of novice usability evaluators. By doing this, we wish to explore whether people with a basic foundation in software engineering and programming through methodological support can build a capability to identify, describe and classify usability problems. The following section gives an overview of existing literature on identification of problems. The next section describes the design of an empirical study we have conducted in order to examine the usefulness of the usability problem identification tool we have developed for problem identification. Then the results of the empirical study are presented and discussed. Finally, we conclude on our study.

BACKGROUND

With the prevalent role of contemporary websites in today's societies, website usability has received increased attention over the last years and several textbooks on website usability has been published (Badre, 2002; Krug, 2000; Nielsen, 2000; Nielsen and Tahir, 2002). While such literature primarily focuses on specific elements of usability in websites, e.g. Nielsen and Tahir (2002) analyze 50 different websites on their usability; some references in the research literature provide methodological support of the usability evaluation process for web sites. Primarily, some research attempts have proposed heuristics for website evaluation (Agarwal and Venkatesh, 2002; Sutcliffe, 2001). On the other hand, the more general literature on usability evaluation practices and means to support it is varied and rich. On the overall level, there are methods to support the whole process of a usability evaluation, e.g. (Rubin, 1994). The literature that compares usability evaluation methods also includes detailed descriptions of procedure for conducting evaluations, e.g. how to identify, group and merge lists of usability problems (Hornbæk and Frøkjær, 2004; Jeffries et al., 1991; Karat et al., 1992). All of this deals with user-based usability evaluation.

Heuristic methods for usability evaluation have been suggested as means to reduce the resources required to conduct a usability evaluation. In many cases, strong limitations in terms of development time effectively prohibits conventional usability testing as it is described in classical methods (Dumas and Redish, 1993; Fath et al., 1994; Nielsen, 1993; Nielsen et al., 1992; Rubin, 1994). Such evaluations are very time-consuming, and considerable costs arise when a large group of users is involved in a series of tests. Heuristic inspection evolved as an attempt to reduce these costs (Lavery et al., 1997; Nielsen, 1993; Nielsen et al., 1992). The basic idea is that a group of usability experts evaluate an interface design by comparing it to a set of guidelines, called heuristics (Nielsen, 1992). The first heuristics consisted of nine principles (Lavery et al., 1997), which have been developed further over the last ten years. The literature on heuristic inspection also includes empirical studies of its capability for finding usability problems. The first studies indicated that the method was very effective (Agarwal and Venkatesh, 2002; Lavery et al., 1997; Nielsen, 1992; Nielsen et al., 1992). Other studies have produced less promising results as they conclude that a conventional user-based usability test yields similar or better results compared to inspection (Karat et al. 1992), and heuristic inspection tends to find many low-priority problems (Jeffries et al., 1991). But usability heuristics designated for website design and evaluation have been proposed and successfully adapted in some research studies (Agarwal and Venkatesh, 2002, Sutcliffe, 2001). Finally, the basic idea in heuristic evaluation is also the key characteristic of the usability evaluation method called MOT, where five metaphors of human thinking are used as a basis for evaluation (Hornbæk and Frøkjær, 2004).

There is also research that describes how usability experts actually conduct evaluations. It has been established that expert evaluators find different usability problems. This has been denoted as the evaluator effect (Hertzum and Jacobsen, 2001; Jacobseb et al., 1998). There is a remarkable difference both in the number of problems and the specific problems they find. The strength is that if we introduce more evaluators, we find more problems. The weakness is that it seems random and difficult to trust.

Changes in software development with new development approaches such as open source development, global software development and outsourcing are challenging conventional usability evaluation practices. With outsourcing and global software development, developers, evaluators and users are distributed across multiple organizations and time zones. This also characterizes several website development projects and makes conventional user-based usability testing considerably more complex and challenging (Murphy et al., 2004). This makes remote usability testing increasingly important as an alternative to conventional usability testing (Andreasen et al., 2007). Remote usability testing denotes a situation where "the evaluators are separated in space and/or time from users" (Castillo et al., 1998). The first methods for remote usability testing emerged about ten years ago. At that time, some empirical studies were conducted that showed results comparable to conventional methods (Hartson et al., 1998). A very interesting method was based on the idea that users should report the critical incidents they experienced while using the system (Hartson et al., 1996; Hartson et al., 1998). A recent study of remote usability evaluation methods concluded that users report significantly fewer problems compared to a classical usability evaluation but the method imposes considerably less effort on the evaluators (Andreasen et al., 2007; Bak et al., 2009).

A related line of research has inquired into the ability of novice usability evaluators to identify usability problems. Based on a comparison with experts it is concluded that novice evaluators can quickly learn to plan and conduct user-based usability evaluations and to write up the related reports. However, when it comes to identification, description and categorization of usability problems, they perform at a significantly lower level than expert evaluators (Skov and Stage, 2001; Skov and Stage, 2004).

The amount of research on user-based usability evaluation conducted by novices is very limited. We have only been able to find one reference where novices conducted the evaluation, and this was heuristic evaluation and not user-based (Slavkovic and Cross, 1999). An effort with training focused on transfer of developers' skills in design of user interfaces from one technology to another (Nielsen et al., 1992).

These streams of research emphasize a need for methodological support to novice or non-expert usability evaluators in identifying usability problems. They also illustrate that the literature is limited in this area.

CONCEPTUAL TOOL FOR USABILITY PROBLEM IDENTIFICATION

During a series of courses on usability testing for under-graduate students, we discovered a clear fundamental need for support on usability problem identification for novice usability evaluators. Especially, we found that even though test participants experienced usability problems, novice evaluators were incapable of identifying and classifying such problems (Skov and Stage, 2001). As a solution, we came up with the idea of the usability problem identification tool (see table 1).

The basic idea in the usability problem identification tool is that it provides a conceptual or overall interpretation of what constitutes a problem. Inspired by previous research (Molich, 2000; Nielsen, 1993; Rubin, 1994) and our own practical experiences with usability test teaching, we identified four overall categories of usability problems as experienced by users:

1) slowed down
2) understanding
3) frustration
4) test monitor intervention.

These four episodes often reveal some sort of usability problem. 1) The first category includes problems where the test participant is being slowed down relatively to normal speed. Several usability problems denotes and describes some sort of users being slowed down while interacting with a website. Thus, they are not able to complete assigned tasks in an efficient manner. 2) The second category of problems deals with users' understanding of the website. Often users find it difficult to understand how website are constructed, what functionality the website offers, and how information is organized in the website. 3) The third category describes problems related to the user's level of frustration. This is a classical metric in usability evaluation studies where researchers focus on the user frustration as an indicator of website usability. Users may (or may not) show their frustration during a usability test session, however if they do so, it is often due to interaction problems with the interface. 4) The fourth category shows problems where the test monitor has intervened or helped the test participant in order to complete the assigned tasks. A good acting test monitor will intervene (and only intervene) if the participant experience severe problems in task completion.

On the other dimension, we distinguish between three severities of problem namely critical problem, serious problems, and cosmetic problems – inspired by previous research (Molich, 2000).

Table 1. Usability problem identification tool

	Slowed down *relative to normal work speed*	**Understanding**	**Frustration**	**Test monitor intervention**
Critical	Hindered in solving the task	Does not understand how information in the system can be used for solving a task. Repeats the same information in different parts of the system.		Receives substantial assistance (could not have solved the task without it).
Serious	Delayed for several seconds	Does not understand how a specific functionality operates or is activated. Cannot explain the functioning of the system.	Is clearly annoyed by something that cannot be done or remembered or something illogical that you must do. Believes he has damaged something.	Receives a hint.
Cosmetic	Delayed for a few seconds	Does actions without being able to explain why (you just have to do it).		Is asked a question that makes him come up with the solution

EMPIRICAL STUDY

We have conducted an empirical study with a usability problem identification tool that is intended to support novice or non-expert evaluators in identifying usability problems in a user-based evaluation. The purpose of the empirical study was to examine whether this tool was useful for such inexperienced evaluators.

Setting: The empirical study was conducted in relation to a course that one of the authors of this paper was teaching. The course was an introduction to design and implementation of user interfaces. It consisted of the following three modules:

A. Introduction to human computer interaction and a method for user interaction design
B. Implementation of user interfaces in Java
C. Usability evaluation of interactive systems

Each module consisted of five class meetings with a two-hour lecture and an equal amount of time for exercises. The experiment was part of the last module (module C). The content of that module was a presentation of the activities of a usability evaluation and techniques that are relevant in each activity. The main literature was (Preece et al., 2002) supplemented with selected articles. The five lectures of this module had the following contents:

1. The purpose of a usability evaluation, the concept of usability and overview of the activities involved in a usability evaluation
2. Basic decisions, field versus lab, the test monitor role and the test report
3. Creation of test context, tasks assignments, conducting the test and the think-aloud technique
4. Interpretation of data, the ISO definition, task load, identification of usability problems, exercises in identification and categorization of usability problems
5. Presentation of experiences from our evaluation, heuristic evaluation, comparison with think-aloud and training of novices in usability evaluation

Subjects: The participants in the experiment were 24 undergraduate second-year students in computer science. They had a basic qualification in programming and software engineering. They were offered to participate in this experiment as a voluntary exercise, and they were promised feedback on their products.

Usability problem identification tool: The empirical study involved a one-page usability problem identification tool that the authors had developed during earlier usability evaluations, see Table 1. The authors also used this tool in their own data analysis

Experimental procedure: The empirical study was conducted between lecture 3 and 4. The students were only told in advance that there would be an exercise about usability problems, but no details were given. The empirical study lasted for three and a half hour. All students came into the class at 8:30. They were handed a CD-ROM with the same recording of a usability test session and a few practical guidelines for carrying out the exercise. The test session was app. 30 minutes. The students also received the usability problem identification tool they were asked to use, cf. Table 1. The recording was of a user that solved a series of tasks on a web-site for a large furniture store. The think-aloud technique was used. The empirical study ended at noon when they delivered their problem lists and diaries by email.

The students were asked to work individually on the task. They would see the recording and note down usability problems as they occurred. In doing so, they were encouraged to use the usability problem identification tool. Thus the tool gave a practical definition of usability problems, and it was supposed to be used in the detailed analysis. For each usability problem they identified, they were also asked to record in the diary if they used the tool and which field in the table the problem was related to.

Data collection: The main result was the problem list from each student. In addition, they were asked to maintain a diary with reasons why they decided that something was a usability problem and why they categorized it at a certain level. In this paper, we only deal with the problem lists.

Table 2. Example of a usability problem

No.	Window	Description	Severity
13	Product page	Does not know how to buy the article that is described in the page; is uncertain about the procedure to buy an article on-line	Serious

Data analysis: The two authors of this paper analysed the recording independently of each other and produced an individual problem list where each problem was described as illustrated in Table 2. The first column contains the unique number of the problem. The second column specifies the window or screen where the problem occurred. The third column contains the description of the way the user experiences the problem. In the individual problem lists, each evaluator also made a severity assessment for each usability problem. This was expressed on a three-point scale, e.g. cosmetic, serious, or critical (Molich, 2000). The individual problem lists from the two authors were merged through negotiation into one overall list of usability problems. The resulting problem list was the basis for evaluating the problem lists produced by the participants in the experiment. Thus the problem list from each student was compared to the authors' joint problem list.

Validity: The specific conditions of this study limit its validity in a number of ways. First, the students participated in the empirical study on a voluntary basis receiving no immediate credit for their participation. Thus, motivation and stress factors could prove important. 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. Secondly, the demographics of the test subjects are not varied with respect to age and education. Most subjects were students of approximately 22 years of age with approximately the same school background and recently started on a computer science education.

RESULTS

This section presents the key results from our empirical study. First, we present the problem identification by the 24 participants and compare their reporting with the usability experts. Second, we analyze the identified problems according to their categorization as done by the participants.

Identifying and Reporting Usability Problems

The participants identified very different numbers of usability problems. This is illustrated by two participants reporting no usability problems while one participant identified and reported 18 different usability problems. On average, the participants identified 8.00 usability problems (SD=4.63). This is illustrated in table 3. The high variety in numbers of reported problems suggests strong presence of the evaluator effect. Therefore, the usability problem identification tool did not in it self remove this effect and indicates that some participants only marginally used the tool.

From our data it seemed from the reporting of usability problems that our participants could be divided into three different groups regarding numbers of reported problems. The first group reported no or very few problems (0-3), the second group reported up to ten problems (4-10), and the third group reported more than ten problems (>10). Six participants belonged to the first group, and 11 participants belonged to the second group, while seven participants belonged to the third group. Interestingly, we saw a gap between participants from the first group compared to the second group as the "best" participants in group one identified three problems whereas none in the second group reported less than seven problems

Table 3: Mean numbers of identified problems and non-problems.

	Problems	**Non-Problems**	**Sum**

Tool Participants (N=24)	8.00 (4.63)	2.95 (2.87)	10.39 (5.83)

As stated earlier and further illustrated above, novice evaluators often find it difficult just to see and identify usability problems. Additionally, they are typically faced with challenges when trying to describe (or illustrate) identified problems. Several participants reported issues from the usability test as problems but it was impossible for us to figure out or extract the actual problem from the descriptions. We denote such issues as non-problems (see table 3). In several cases, these issues were even described in a non-problematic way (e.g. as a positive or neutral feature of the tested system). The participants reported on average 2.95 non-problems (SD=2.87). Again, the numbers of reported non-problems were very diverse between participants having some participants reporting zero non-problems while one participant reported 10 non-problems.

Having discussed numbers of problems identified per participant, we will now outline the reporting of problems for all participants as one group. Two usability experts also conducted a video analysis of the test session and reported usability problems. We will in the following compare the participants in the experiment with these usability experts.

The 24 participants together identified and reported a total of 28 different usability problems. Thus on one hand, they were not able to identify all known problems as reported by the usability experts, but they were able to report on a substantial amount of these problems (72%). When looking at problem severity, we found that the participants were able to identify many of the more severe problems. As a group, they identified 86% of the most severe problems (critical and serious problems) where they identified both critical problems and 16 out of the 19 serious problems. On the other hand, they reported on the identification of 12 cosmetic problems out of a total of 18 problems. One participant identified a usability problem not identified or reported by any of the two usability experts. In summary, the participants as a group were able to identify most severe problems as reported by the usability experts while they missed some cosmetic problems in their reporting.

Table 4: Total numbers of identified problems for the two approaches.

	Usability Experts (N=2)	**Participants (N=24)**	**Sum (N=26)**
Critical	2	2	2
Serious	19	16	19
Cosmetic	17	12	18
Total	38	28	39

Considering numbers of participants reporting the 28 problems, further analysis show that problem severity had an impact on identification and reporting. Thus, the more severe a problem was the higher the chance of identification and reporting. On average, the critical problems were reported by 67% of the participants. The two critical problems were reported by 18 participants respectively 14 participants. The same figures are considerably lower for the serious and cosmetic problems. In fact, our analysis show that a critical usability problem was significantly more likely to be reported by a participant than a serious or a cosmetic usability problem according to two-tailed Chi-square tests ($\chi^2[1]=47.691$, p=0.0001; $\chi^2[1]=66.012$, p=0.0001). However, we only discovered a tendency towards a serious problem being more likely to be reported than a cosmetic problem, but this finding was not significant ($\chi^2[1]=3.725$, $p=0.0536$). Summarized, it appeared that severity had considerable impact on identification as severe problems were more likely to be reported.

Categorization of Usability Problems

As an integrated part of the usability problem identification tool, problems should be categorized according to severity. The usability problem identification tool integrates three levels of severity namely critical, serious, and cosmetic problems (see Table 1). The categorization was characterized by some diversity but also by agreement between the participants. All problems had been categorized according to severity by the two usability experts.

The two problems categorized as critical by the usability experts were identified by 18 respectively 14 participants out of the total number of 24 participants (as discussed previously). However, the two problems were categorized very differently by the participants. The first critical problem (reported by 18) was unanimously categorized as critical by all participants who identified it. This particular problem is that the test subject was unable to complete a purchase on the website. Interestingly, the second critical usability problem was categorized rather differently, where only one participant categorized it as critical, and nine participants categorized it as serious, while four categorized it as cosmetic. This problem is subtle as it reflects how the test subject understands interface elements which make her navigate wrongly. The problem was categorized as critical by both of the usability experts as it delayed her task completion for several minutes. Either the participants did not see this long delay or they disagreed that she was delayed this long. This is not clear from the descriptions, but their reporting typically lacked information on task delay in this situation. This was quite the opposite for the other problem where she failed to complete the task and the delay was obvious.

The remaining 26 serious and cosmetic problems were categorized quite differently by the participants compared the categorization made by the usability experts. Three problems received all three categorizations ranging from critical to cosmetic, but most problems were categorized as either serious or cosmetic. Also, five problems received unanimously categorizations by the participants. Summarized, our analysis of usability problem categorization confirms that this is a highly difficult and challenging task. Furthermore, it seems that individual differences between evaluators are very prominent.

DISCUSSION

Our aim with the usability problem identification tool is to provide software designers and programmers support in constructing more usable interfaces. Thus, we strive to contribute to the body of knowledge within discount usability evaluation by integrating the activities of usability testing into the knowledge of the software developer. In addition, we are interested in providing software projects that are distributed physically with tools or techniques that can support remote usability testing. Inspired by previous research on usability evaluation and particularly on the challenges related to usability problem identification, we developed a usability problem identification tool for use in user-based usability evaluations. The tool is supposed to support evaluators during video analysis of user interaction with a computerized system by emphasizing different levels of problems and modes of experiencing problems. We evaluated the usability problem identification tool in an experiment with 24 participants. All participants had only introductory knowledge of human-computer interaction issues and no specific training in analysis of usability test sessions.

Our empirical study shows that the participants were able to identify and report many of the more severe problems from the test session. Two critical problems were identified by more than half of the participants; in fact, one critical problem was discovered by 75% of the participants. Several participants used and applied the tool in the identification of the problems and tried to express the problems in terms of the different suggested modes. Especially user delay was commonly used in the reporting. Not surprisingly, less severe problems were not identified to the same extent as the critical problems. More of these problems were only reported by one or two participants, while only four problems were reported by at least ten participants. No problem was reported by all participants partly as a consequence of the fact that two participants reported no usability problems at all.

Promoting remote or distance usability testing conducted by the users themselves require some sort of framework to guide the testing or the analysis. As a group, somewhat surprisingly the participants performed well by identifying a substantial amount of the usability problems. In fact, the most severe problems namely the critical and serious problems were identified almost completely by the group taken as a whole.

A major challenge in usability problem identification and categorization is the so-called evaluator effect. Previous studies have found that the evaluator effect is challenging in user-based usability evaluations such as think-aloud tests as evaluators identify substantial amounts of unique problems (Hertzum and Jacobsen, 2001; Jacobsen et al., 1998). Furthermore, evaluators also suffer from the fact that they identify very few common problems. Our results seem to confirm the evaluator effect as our participants identified several unique problems. This is not surprising. At this stage, we cannot conclude whether the tool addressed or solved some of the inherent problems of the evaluator effect, but we can see that participants in several cases used the tool actively in their descriptions. But further studies are needed to confirm or reject the effects of different evaluators.

Categorization of usability problems is very difficult and challenging. This was confirmed in our experiment. It seemed that the tool only marginally supported the categorization. The tool was designed to integrate key aspects of severity by illustrating different modes of usability problems for different severity ratings. In certain situations, it seemed to help the participants in understanding the situation and therefore more easily being able to categorize the observed problem. As an example, most participants actively used the tool in the categorization of one of the critical problems. However, several problems were categorized rather differently by the participants sometimes reflecting differences in the assessed scope of the problem.

We have only involved novice evaluators as participants in our study, just like the studies in (Hornbæk and Frøkjær, 2004). Studies involving expert evaluators tend to identify more and different kinds of problems (Nielsen, 1992). However, to compensate against this potential problem, we measured the participants' performance against experienced usability evaluators. The participants taken together identified a significant proportion of the problems identified by the experts.

FURTHER DEVELOPMENT

The usability problem identification tool used in this experiment has been developed entirely through introspective observation of our own problem identification and categorization process in other usability evaluations. In addition, certain parts are still vaguely defined. In other domains, there is a considerable confidence in the use of checklists. Schamel (2008) present the history behind introduction of checklists in aviation.

Hales and Provonost (2006) describes a checklist as a list of action items or criteria that are arranged in a systematic manner, which allow the user to record the presence/absence of the individual items listed to ensure that all of them are considered or completed. They emphasize that the objectives of a checklist may be to support memory recall, standardization and regulation of processes or methodologies. They present an overview of the use of checklists in the areas of aviation, product manufacturing, healthcare and critical care.

Hales et al. (2008) have conducted a systematic study of literature on checklist. They provide a categorization of different checklists with examples from medicine. Some of these categories are clearly relevant for usability evaluation. The tool we have presented in this chapter resembles what they call a Diagnostic checklist or a Criteria of merit checklist. They also provide guidelines for development of checklists. This could be a useful basis for enhancing our problem identification tool.

Verdaasdonk et al. (2008) argue that the use of checklists is a promising strategy for improving patient safety in all types of surgical processes. They present requirements and guidelines for implementation of checklists for surgical processes. Helander (2006) describes checklists in relation to HCI. He emphasizes the checklist as a memory aid that can be used to support systematic assessment of ergonomics in a workplace.

All of these examples deals with checklists for professionals. Our tool can be considered as a simple checklist. It has been shown that a usability problem identification tool like the one presented in this chapter combined with education provides solid support to problem identification and categorization (Skov and Stage, 2005). The results presented in this chapter show that even without training, the tool provides some assistance. However, the results also indicate that the tool could be improved. Other researchers in the HCI area have worked with definition of usability problems from a usability problem identification platform (Cockton et al., 2004; Lavery et al., 1997). It may be possible to combine this with the guidelines for developing checklist in order to create an enhanced usability problem identification tool.

CONCLUSION

This paper has presented results from an empirical study of methodological support to identification of usability problems as part of a usability evaluation. The key element of the support was a usability problem identification tool for identification of usability problems.

The non-expert participants in the experiment found on average 8 usability problems, but with substantial differences between them. Two usability experts found 38 problems. Compared to this, the performance of the participants is limited. On the other hand, the 24 participants together identified 72% of the problems found by the experts. And they found nearly all critical and serious problems. This is very interesting given that the time spent on data analysis of the problem lists produced by the participants is very limited. This indicates that even with a very limited expert effort you are able to get a large proportion of the severe problems provided that you involve a group of participants that is larger than what we normally are used to. This gives a reason to be optimistic about the ideas of having developers report usability problems.

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.

These conclusions are based on a single experiment with 24 participants. Unfortunately, there are very few results in the literature to compare with. Therefore, it would be interesting to repeat the empirical study. The usability problem identification tool could also be developed further, especially in order to support categorization of usability problems.

ACKNOWLEDGMENTS

The work behind this paper received financial support from the Danish Research Agency (grant no. 2106-04-0022). We would especially like to thank all the participating test subjects. Finally, we want to thank the anonymous reviewers for comments on the draft of this paper

REFERENCES

Agarwal, R. and Venkatesh, V (2002) Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability. Information Systems Research, Vol. 13, No. 2, pp. 168-186

Anderson, J., Fleek, F., Garrity, K. and Drake, F. (2001) Integrating Usability Techniques into Software Development. IEEE Software, 18(1):46-53.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O. and Stage, J. Usability in open source software development: Opinions and practice. Information Technology and Control 35A, 3 (2006), 303-312.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O. and Stage, J. What Happened to Remote Usability Testing? An Empirical Study of Three Methods. Proceedings of CHI 2007. ACM Press (2007).

Badre, A. N. (2002) Shaping Web Usability – Interaction Design in Context. Addison-Wesley, Boston, USA.

Bak, J. O., Nguyen, K., Risgaard, P. and Stage, J. (2008) Obstacles to Usability Evaluation in Practice: A Survey of Software Organizations. Proceedings of NordiCHI 2008, ACM Press.

Benson, C., Muller-Prove, M. and Mzourek, J. Professional usability in open source projects: Gnome, openoffice.org, netbeans. Proceedings of CHI 2004, ACM Press (2004), 1083-1084.

Braiterman, J., Verhage, S., and Choo, R. (2000). Designing with Users in Internet Time. interactions, 7, 5 (September–October), pp. 23-27.

Bruun, A., Gull, P., Hofmeister, L. and Stage, J. (2009) Let your users do the testing: a comparison of

three remote asynchronous usability testing methods. Proceedings of CHI 2009, ACM Press.

Castillo, J. C., Hartson, H. R. and Hix, D. Remote usability evaluation: Can users report their own critical incidents? Proceedings of CHI 1998, ACM Press (1998), 253-254.

Cockton, G., Woolrych, A. & Hindmarch, M. (2004) Reconditioned Merchandise: Extended Structured Report Formats in Usability Inspection. CHI 2004 Extended Ab-stracts, pp. 1433-36. ACM Press: New York

Dempsey, B. J., Weiss, D., Jones, P. and Greenberg, J. Who is an open source software developer? Communications of the ACM 45, 2 (2002), 67-72.

Dumas, J. S. & Redish, J. C. (1993). A practical guide to usability testing, Norwood, NJ: Ablex Publishing.

Fath, J. L., Mann, T. L., and Holzman, T. G. (1994) A Practical Guide to Using Software Usability Labs: Lessons Learned at IBM. Behaviour & Information Technology 13, 1-2, 25-35.

Frishberg, N., Dirks, A. M., Benson, C., Nickell, S. and Smith, S. Getting to know you: Open source development meets usability. Proceedings of CHI 2002, ACM Press (2002), 932-933.

Hales, B. M. and Provonost, P. J. (2006) The checklist. A tool for error management and performance improvement. Journal of Critical Care, 21:231-235.

Hales, B. M., Terblanche, M. Fowler, R. and Sibbald, W. (2008) Development of medical checklists for improved quality of patient care. International Journal for Quality in Health Care, 20(1): 22-30.

Hartson, H. R. and Castillo, J. C. Remote evaluation for post-deployment usability improvement. Proceedings of AVI 1998, ACM Press (1998), 22-29.

Hartson, H. R., Castillo, J. C., Kelso, J. and Neale, W. C. Remote evaluation: The network as an extension of the usability laboratory. Proceedings of CHI 1996, ACM Press (1996), 228-235.

Helander, M. (2006) A Guide to Human Factors and Ergonomics, 2nd ed. CRC Press.

Hertzum, M. and Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about us-ability evaluation methods. International Journal of Human-Computer Interaction, 13, 4, 421-443.

Hornbæk, K. & Frøkjær, E. (2004) Usability Inspection by Metaphors of Human Thinking Compared to Heuristic Evaluation. International Journal of Human-Computer Interaction, 17(3) 357-374.

ISO 9241-11. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) Part 11: Guidance on usability. ISO 1997

Jacobsen, N.E., Hertzum M. and John, B.E. (1998) The Evaluator Effect in Usability Tests. Proc. CHI'98, ACM Press

Jeffries, R., Miller, J. R., Wharton, C., and Uyeda, K. M. (1991) User Interface Evaluation in the Real World: A Comparison of Four Techniques. In Proceedings of CHI '91, ACM Press, 119-124.

Karat, C.-M., Campbell, R., and Fiegel, T. (1992) Comparison of Empirical Testing and Walk-through Methods in User Interface Evaluation. In Proceedings of CHI '92, ACM Press, 397-404.

Krug, S. (2000) Don't Make Me Think – A Common Sense Approach to Web Usability. Circle.com Library, USA

Lavery, D. Cockton, G. and Atkinson, M.P. (1997) Comparison of Evaluation Methods Using Structured Usability Problem Reports. Behaviour and Information Technology, 16(4), 246-266.

Molich, R. (2000) User-Friendly Web Design (in Danish). Copenhagen: Ingeniøren Books.

Murphy J., Howard S., Kjeldskov K. and Goschnick, S. Location, location, location: Challenges of outsourced usability evaluation. Proceedings of the Workshop on Improving the Interplay between Usability Evaluation and User Interface Design, NordiCHI 2004, Aalborg University, Department of Computer Science, HCI-Lab Report no. 2004/2 (2004), 12-15.

Nielsen, J. (1992) Finding Usability Problems Through Heuristic Evaluation. In Proceedings of CHI '92, ACM Press, 373-380.

Nielsen, J. (1993) Usability Engineering. Morgan Kaufmann Publishers

Nielsen, J., Bush, R. M., Dayton, T., Mond, N. E., Muller, M. J., and Root, R. W. Teaching experienced developers to design graphical user interfaces. Proceedings of CHI 1992. ACM Press (1992) 557-564.

Nielsen, J. (2000) Designing Web Usability. New Riders Publishing, USA

Nielsen, J. and Tahir, M. (2002) Homepage Usability – 50 Websites Deconstructed. New Riders Publishing, USA

Preece, J., Rogers, Y. and Sharp, H. (2002) Interaction Design: Beyond Human-Computer Interaction. New York: John Wiley and Sons.

Rohn, J. A. (1994) The Usability Engineering Laboratories at Sun Microsystems. Behaviour & Information Technology 13, 1-2, 25-35.

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

Schamel, J. (2008) How the pilot's checklist came about. http://www.atchistory.org/History/checklst.htm

Scholtz, J., Laskowski S. and Downey L. (1998) Developing Usability Tools and Techniques for Designing and Testing Web Sites. Proceedings of the 4th Conference on Human Factors & the Web. AT&T.

Skov, M. B. and Stage, J. A Simple Approach to Web-Site Usability Testing. In Proceedings of 1st International Conference on Universal Access in Human-Computer Interaction. Lawrence-Erlbaum (2001) 737-741

Skov, M. B. and Stage, J. (2004) Integrating Usability Design and Evaluation: Training Novice Evaluators in Usability Testing. K. Hornbæk and J. Stage (Eds.), Proceedings of the Workshop on Improving the Interplay between Usability Evaluation and User Interface Design, NordiCHI 2004, pp. 31-35. Aalborg University, Department of Computer Science, HCI-Lab Report no. 2004/2.

Skov, M. B. and Stage, J. (2005) Supporting Problem Identification in Usability Evaluations. Proceedings of the Australian Computer-Human Interaction Conference 2005 (OzCHI'05), ACM Press.

Slavkovic, A. and Cross, K. Novice heuristic evaluations of a complex interface. In proceedings of CHI 1999. ACM Press (1999) 304-305.

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

Sullivan, T. and Matson, R. (2000). Barriers to Use: Usability and Content Accessibility on the Web's Most Popular Sites. In Proceedings of Conference on Universal Usability, November 16-17, Washington, ACM, pp. 139-144.

Sutcliffe, A. (2001) Heuristic Evaluation of Website Attractiveness and Usability. Interactive Systems: Design, Specification, and Verification. Lecture Notes in Computer Science Vol. 2220, 183-198

Verdaasdonk, E. G. G., Stassen, L. P. S., Widhiasmara P. P. and Dankelman, J. (2008) Requirements for the design and implementation of checklists for surgical processes. Surgical Endoscopy.