Behaviour & Information Technology
2009, 1–11, i*First* article

Taylor & Francis
Taylor & Francis Group

Training software developers and designers to conduct usability evaluations

Mikael Brasholt Skov and Jan Stage*

Department of Computer Science, Aalborg University, Selma Lagerlöfs Vej 300, DK-9220 Aalborg East, Denmark

(*Received 30 September 2008; final version received 8 October 2009*)

Many efforts to improve the interplay between usability evaluation and software development rely either on better methods for conducting usability evaluations or on better formats for presenting evaluation results in ways that are useful for software designers and developers. Both of these approaches depend on a complete division of work between developers and evaluators. This article takes a different approach by exploring whether software developers and designers can be trained to conduct their own usability evaluations. The article is based on an empirical study where 36 teams with a total of 234 first-year university students on software development and design educations were trained through an introductory course in user-based website usability testing that was taught in 40 h. They used the techniques from this course for planning, conducting, and interpreting the results of a usability evaluation of an interactive website. They gained good competence in conducting the evaluation, defining user tasks and producing a usability report, while they were less successful in acquiring skills for identifying and describing usability problems.

Keywords: usability; user-based evaluation; training of software developers; dissemination of usability skills; empirical study

1. Introduction

Usability evaluation and user interaction design are two key activities in the development of an interactive system. The two activities are mutually dependent, but in practice there is often too little or no fruitful interplay between them (Hornbæk and Stage 2006). Considerable efforts have been devoted to improve the interplay between usability evaluation and software development. A substantial part of these efforts reflect two typical approaches.

The first approach focuses on better methods. The aim is to improve the products of usability evaluations through use of methods that provide better support to the evaluators that carry out usability evaluations. During the last 20 years, a whole range of methods have been developed within this approach. A prominent and influential example is Rubin (1994) that covers all activities in a usability evaluation. There are many others that cover all or some selected evaluation activities.

The second approach focuses on better feedback. The aim is to improve the impact of usability evaluations on user interaction design. This is achieved in a variety of ways, typically by improving the format that is used to feed the results of usability evaluations back into user interaction design. The classical format for feedback is an extensive written report, but there have been numerous experiments with alternatives to the report; see Høegh *et al.* (2006) for an overview.

For both of these approaches, website development is particularly challenging. Websites exhibit a huge and unprecedented amount of information, services and purchasing possibilities, and the users of websites are a tremendously heterogeneous group that use websites for a multitude of purposes any time, any place. Because of this, website developers must accommodate a massive variety of user preferences and capabilities.

A conventional usability evaluation that involves the prospective users of an interactive system facilitates a rich understanding of the actual problems that real users will experience (Rubin 1994). The main drawback of user-based usability evaluations is that they are exceedingly demanding in terms of time and other resources; some researchers have reported that durations of 1 month or more and efforts amounting to around 150 person-hours are not unusual (Molich and Nielsen 1990, Nielsen 1992, Molich *et al.* 2004). These figures are simply not feasible for many website projects. Often, they do not have such an amount of resources, and they cannot wait for the usability evaluators to conduct the evaluation and provide the needed feedback.

Time pressure is a key reason why established knowledge and methodologies are ignored in many

*Corresponding author. Email: jans@cs.aau.dk

ISSN 0144-929X print/ISSN 1362-3001 online

DOI: 10.1080/01449290903398208
http://www.informaworld.com

website development projects (Baskerville and Pries-Heje 2001). Website developers experience a strong push for speed and users of websites rapidly change preferences and patterns of use, and new ideas for design and functionality emerge constantly. This makes customers and management demand development cycles that are considerably shorter than in traditional software development (Anderson 2000, Broadbent and Cara 2000).

The two approaches that were emphasised above share the key characteristic that they involve a complete division of work between developers and evaluators. Software is made by developers, and its usability is assessed by evaluators. This division of work may create difficulties for fast-paced projects, as it necessitates handovers between the two groups. This increases project complexity and tends to lengthen development time. On the other hand, the division of work between developers and evaluators ensures that the usability evaluation is unbiased.

This article presents results from an empirical study of a course where first-year students in software development and design educations were trained to conduct their own user-based usability evaluations. The aim of the approach behind this course is to facilitate direct integration of usability evaluation into software development by removing the division between evaluators and developers. In the study, we explored whether future designers and software developers who had received a 40 h training course could conduct a usability evaluation of a reasonable quality. In Section 2, we present previous work related to our study. In Section 3, we describe the study in detail. The results of the study are presented in Section 4, and Section 5 discusses additional aspects of the results. Finally, Section 6 provides the conclusion.

2. Related work

There is a significant body of work on problems with website usability. It has been shown that many contemporary websites suffer from problems with low usability, e.g. an investigation of content accessibility found that 29 of 50 popular websites were either inaccessible or only partly accessible (Sullivan and Matson 2000). This is in line with the suggestions that usability evaluations of websites should focus on the extent to which users can navigate the website and exploit the information and possibilities for interaction that are available (Spool *et al*. 1999). More recent studies of usability and accessibility of websites for specialised areas confirm that there are still considerable problems; for example of websites of top universities in the USA (Zaphiris and Ellis 2001), of aging and health-related websites (Zaphiris *et al*. 2001) and selected tourist websites (Maswera *et al*. 2005). A comparison of usability between the website of Fortune 30 companies and the 30 fastest growing companies in the USA also revealed considerable usability problems, especially for the fast growing companies (Brown *et al*. 2006).

It has been emphasised that there is a gap between software development and usability evaluation, because the results of usability evaluations often have little or no effect on the software (Hornbæk and Stage 2006). It has been suggested that education to broaden the usability engineering skills of software developers could contribute to close this gap by reducing the usability problems that characterise many software products. This suggestion focuses on a general awareness of usability issues and on the early activities in a development project (Karat and Dayton 1995).

It has also been discussed on a more general level how development teams could be better trained to use fundamental techniques from the usability engineering discipline. This requires systematic empirical studies of the true costs of learning and applying usability engineering techniques (John 1996).

We conducted a search on the Web on training of software developers in usability engineering. We found a group of companies that offer training courses for software developers in various methods from the usability engineering discipline. The two most common methods were the so-called discount usability evaluation techniques (expert inspection and walkthrough) and user-based empirical testing based on a think-aloud protocol. There were much fewer and mostly shorter courses on general usability topics. Such courses for practitioners respond to the request for training of practitioners in usability topics (Karat and Dayton 1995). Unfortunately, they are not complemented by the research studies of cost and effects that were also requested (John 1996). In fact, we have only been able to find very few systematic studies of efforts to train software developers in key usability engineering topics.

A notable exception to this limited amount of research is an empirical study of training of software engineering students in a language for describing and analysing user interface designs (Blandford *et al*. 1998). This study measured the effect of a training course and also provided improved insight into the way experts work with description of user interface designs. Another study introduced user-centred techniques in a small software development company, with focus on design and inspection, and got positive results (Häkli 2005). Nielsen (1992) showed that usability experts found considerably more usability problems in a specific system compared to novices. Yet, if the novice evaluators are experts in the work domain for the system that is evaluated, their performance is

considerably better, and the impact of the problems they identify is very high (Følstad 2007). In a study of novice evaluators, it was shown that knowledge about business goals increased the utility of the usability problems that were identified (Hornbæk and Nielsen 2008).

Other studies have dealt with certain aspects in relation to novices. Law and Hvannberg (2008) have inquired into the process of merging lists of usability with focus on novices. Another study of novices focused on the idea of including tools for usability evaluation, and it was concluded that novice usability practitioners can benefit from such tools. A particular benefit was that the novices produced higher quality usability reports (Howarth 2007).

3. Method

We have conducted an empirical study of a training course that is intended to teach software developers and designers to conduct user-based usability evaluations. The aim of the study was to provide the participants with skills in formative usability evaluation. To provide a benchmark for the performance of the participants in the training course, we have compared it to a study that is part of the Comparative Usability Evaluation efforts or CUE (Molich, undated); the CUE study we have used is CUE-2 (Molich *et al.* 2004), where a number of usability expert teams evaluated the same website as our students. The CUE-2 study is described in Section 3.3.

3.1. Training course

We studied the training course in a first-year university curriculum. The course included 10 class meetings, cf. Table 1, each lasting 4 h that was divided evenly between 2 h of lecture, and 2 h of exercises in smaller teams. The course required no specific skills in information technology which is the reason why class meeting numbers one and five included introductions to basic technological issues. The purpose of the exercises was to practise selected methods and techniques from the lectures. In the first four class meetings, the exercises made the students conduct small usability pilot tests to train and practice their practical skills with selected methods. The exercises in the last six class meetings were devoted to conducting a realistic usability evaluation of a specified website.

The course introduced a number of methods for usability testing. The first was the conventional method for user-based testing with the think-aloud protocol (Nielsen 1993, Rubin 1994). The second method was based on questionnaires that test subjects fill in after completing each task and after completion of the entire test (Spool *et al.* 1999). The students were also introduced to additional methods such as interviewing, heuristic inspection, cognitive walkthroughs, etc. The content of the course was ambitious, but this was defined in the curriculum and, therefore, not within our control.

The students were required to document their work by handing in a usability report. The instructors suggested to the students that the usability report should consist of (1) executive summary (1 page), (2) description of the usability evaluation method applied (2 pages), (3) results of the evaluation, primarily a list and detailed description of the identified usability problems for the website that was evaluated (5–6 pages), and (4) discussion of the method that was applied (1 page). The report would typically amount to around 10 pages of text. It was further emphasised that

Table 1. The 10 class meetings of the training course.

No.	Lecture	Exercises
1	Introduction to the course and basic website technology	*Pilot test*: Each team conducts simple pilot usability tests of websites to train their practical skills in usability evaluation
2	Basic introduction to usability issues and guidelines for interaction design	
3	The think-aloud protocol and how to set up a test scenario. User groups and their different needs	The teams choose the website themselves. Experience with conducting tests and the results achieved are discussed afterwards
4	Application of questionnaires for collecting data and how to use different kinds of questions	
5	Computer architecture and website technology	*Usability evaluation*: The teams conduct a usability evaluation of the Hotmail website according to a specification provided by the course instructors
6	Describing the usability testing method and how to collect and analyse empirical data	
7	Other usability evaluation methods and how to conduct a full-scale usability test session	The usability evaluations are conducted at the university in assigned rooms for each team
8	Website structures, information search and web surfing	After the usability test sessions, the teams analyse the empirical data and make a usability report that describes the identified usability problems
9	Guidelines for website design and principles for orientation and navigation	
10	Principles for visual design and different interaction styles	

the problems identified should be categorised, at least as major and minor usability problems. In addition, the report should include appendices with all data material produced such as log-files, tasks assignments for test subjects, questionnaires, etc. A prototypical example of a usability report was given to the students.

3.2. Website

We chose www.hotmail.com as the website for our study. This website provides advanced interactive features and functionalities appropriate for an extensive usability test. Furthermore, it facilitates evaluations with both novice and expert test subjects because of its vast popularity. Finally, it had been used in other usability evaluations that have been published, which enabled us to compare the results of the student teams in our study with other result (this is further explained in Section 3.7).

Our study was conducted in the fall semester of 2000. The CUE-2 study that we use as benchmark (see Section 3.3) was conducted late in 1998 and early in 1999 (Molich *et al*. 2004). The CUE-2 results were presented for the Hotmail usability team, and they reported that only 4% of the findings were new to them. We have used the hotmail.com website in our teaching both before and after the study, and we have seen the same problems being identified over and over. Thus even if the website was fine-tuned, the basic problems were not solved. The lack of changes of the website is also reflected in a high degree of correspondence between the problems found by the students and the experts.

3.3. Participants

The participants were first-year university students enrolled in four different studies at a faculty for natural sciences and engineering. The first of the four studies was informatics, which is a user-oriented IT education with focus on software development but also with elements of design in general. The other three studies were architecture and design, planning and environment, and chartered surveyor, which all shared a focus on design in general but also had elements of software development. All students in the four groups of students participated together in the course described in this article. All students attending the course participated in our study. Thus the selection of participants was not within our control. None of the participants had any experience with usability evaluation prior to the study.

Thirty-six teams with a total of 234 students (87 females, 37%) participated in the course and our study. Each team was required to distribute the roles of test subjects, loggers, and test monitor among themselves. This was done before the second class meeting, well before they started the evaluation of the Hotmail website. One hundred and twenty-nine (55%) of the students acted as test subjects, 69 (30%) as loggers, and 36 (15%) as test monitors, cf. the description of roles in Rubin (1994). The average team size was 6.5 students (SD = 0.91). The average number of test subject in the teams was 3.6 (SD = 0.65), and their average age was 21.2 years old (SD = 1.58). Forty-two (33%) of the 129 test subjects had never used www.hotmail.com before the evaluation, whereas the remaining 86 subjects had varied experience with the website. These data are summarised in Table 2.

The students were required to attend the course. The students did not take an exam in the course, but it was part of an examination of a project they did in the same semester. Thus making the report was voluntary, but the students were promised specific feedback, and they all worked seriously during the class meetings. They were not told about our study before they handed in the reports.

We compare the results produced by the students to evaluation results produced by teams from professional laboratories. These reports were selected from a pool of usability reports produced in the CUE-2 study where nine usability laboratories conducted similar usability tests of www.hotmail.com, cf. Molich (undated) and Molich *et al*. (2004). Of the nine professional teams, we discarded one because it only used heuristic inspection, which was different from our focus on user-based evaluation. The nine teams were six commercial or industrial usability evaluation organisations, one university lab that conducted paid usability evaluations and two teams of university students. The two student teams could not be identified in the CUE-2 material,

Table 2. Team and test subject data.

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

but it is stated that their performance was similar to the industry and university teams. The expert teams participated voluntarily in the study. They applied the same procedure as our students, and the evaluation was based on the same scenario. Each team produced a usability report that is available on the CUE-2 website. Below, we describe where and how the expert teams conducted their evaluations.

3.4. Setting

Because of the pedagogical approach of the university, each team had their own office equipped with a personal computer and Internet access. Most teams conducted the tests in their office, while the rest did it in one of their homes. After the tests, the entire team worked together on the analysis and identification of usability problems and produced the usability report.

The expert teams conducted their evaluations in their professional environment, e.g. in their usability laboratories.

3.5. Procedure

The student teams were required to apply the techniques presented in the course. After the second class meeting, the test monitor and loggers of each team received a two-page scenario specifying the web-based mail service www.hotmail.com that they should focus on in the usability evaluation. The scenario also specified a comprehensive list of features that emphasised the specific parts of www.hotmail.com they were supposed to evaluate. The test monitor and the loggers examined the system, designed tasks, and prepared the evaluation, cf. Rubin (1994). The use of www.hotmail.com as the website to be evaluated in the study was kept secret to the test subjects until the actual test was conducted.

The expert teams had received the same scenario as the students. It is still available on the CUE-2 website (http://www.dialogdesign.dk/tekster/cue2/scenario.pdf). Thus, the students and experts worked from exactly the same task description.

3.6. Data collection

The main data collected in the study were the usability reports that were handed in by the teams. The 36 reports had an average length of 11.4 pages (SD 2.76) excluding the appendices that had an average length of 9.14 pages (SD = 5.02). Thirty (83%) of the 36 teams provided information on task completion times for 107 (83%) of the 129 subjects, and they had an average session time (with one user) of 38.10 min (SD = 15.32 min).

We did not collect any data on the way the students performed during the evaluation, and we did not monitor or record how they carried out the evaluations.

The expert reports were obtained from the CUE-2 website. As mentioned above, we discarded one report, because it was exclusively based on heuristic inspection. Thus, we had eight expert reports available.

3.7. Data analysis

All the student reports were analysed, evaluated, and marked by the two authors of this article according to the following three steps.

Step 1

We designed a scheme for the evaluation of the 36 reports by analysing, evaluating and marking 5 randomly selected reports out of the total of 36 reports. Through discussions and negotiations we came up with an evaluation scheme with 17 variables as illustrated in Table 3. The 17 variables were divided into the following 3 overall categories: evaluation (the way the evaluation was conducted), report (the presentation of the evaluation and the results), and results (the outcome of the usability evaluation). Finally, we described, defined, and illustrated all 17 variables in a 2-page marking guide.

Step 2

We worked individually and marked each of the 36 reports in terms of the 17 variables by using the

Table 3. The 17 experimentally identified variables used in the assessment of the 36 usability reports.

Category	Variable
Evaluation	1. Conducting the evaluation
	2. Task quality and relevance
	3. Questionnaires/interviews quality and relevance
Report	4. Test procedure description
	5. Data quality
	6. Clarity of usability problem list
	7. Executive summary
	8. Clarity of report
	9. Report layout
Results	10. Number of identified usability problems
	11. Usability problem categorisation
	12. Practical relevance of usability problems
	13. Qualitative results overview
	14. Quantitative results overview
	15. Use of literature
	16. Conclusion
	17. Test procedure evaluation

marking guide. The markings were made on the following scale of 1 to 5: 1 = wrong answer or no answer at all, 2 = poor or imprecise answer, 3 = average answer, 4 = good answer, and 5 = excellent answer.

We also counted the number of identified usability problems in each of the 36 usability reports. We defined a usability problem as something in the user interaction that prevents or delays users in realising their objectives. Each time a report described such an obstacle or delay, we would count that as a usability problem. Thus, it was our decision whether an element in the report was considered to be a relevant usability problem or an irrelevant observation. Finally, we specified intervals for grading of the identification of usability problems based on their distribution on the following scale: 1 = 0–3 problems, 2 = 4–7 problems, 3 = 8–12 problems, 4 = 12–17 problems, and 5 $\geq$ 17 problems.

Step 3

All reports and grades were compared and a final assessment on each variable was negotiated. In case of disagreements on a grade, we employed the following procedure: (1) if the difference was one grade, we would renegotiate the grade based upon our separate notes; (2) if the difference was two grades, we would reread and reassess the report together focusing only on the variable in question. In our study, no disagreement exceeded two grades. For each report, we also went through the set of usability problems that each of us thought they had identified. We negotiated each team's list of usability problems until we had consensus on that as well.

The expert reports were analysed after we had completed the analysis of the student reports. The eight expert reports were analysed, assessed, and marked through the same procedure as the student reports.

To facilitate comparison with the general performance of the students, we included data about their grades in other courses. This was done by calculating a combined score for each team based on the grades that the individual team members had obtained in other courses they attended in the same semester. The aim was to explore the correlation between the overall skills of the students and their ability to conduct a usability evaluation.

4. Results

The overall results show that the student teams did quite well in conducting the usability evaluation. It is not surprising that the professionals did better on most variables. It was, however, surprising to us that on some variables, the students had a comparable performance and on a few variables they even performed better than the professional teams.

4.1. Evaluation

These three variables relate to the way the usability evaluation was conducted, see Table 4. On variable 1, conducting the evaluation, the professional teams have an average of 4.38 (SD = 0.74). This is almost one grade higher than the student teams and a Mann–Whitney U Test shows strong significant difference between the student teams and the professional teams ($z = -2.68$, $p = 0.0074$). On variable 2, task quality and relevance, the students performed slightly better than the professionals, but this difference is not significant ($z = 0.02$, $p = 0.984$). No significant difference was found on variable 3, questionnaire/interviews quality and relevance ($z = -1.63$, $p = 0.1031$).

4.2. Report

These six variables relate to the quality of the usability report that was the tangible result of the usability evaluations, see Table 5.

The student teams did not perform as well as the professionals on the description of the test, and this difference is significant ($z = -2.15$, $p = 0.0316$). On the other hand, the student teams actually performed significantly better than the professional teams on the quality of the data material in the appendices ($z = 2.07$, $p = 0.0385$).

On the clarity of the usability problem list, we found a strong significant difference in favour of the professional teams ($z = -2.98$, $p = 0.0029$). There is also a significant difference on the teams' executive summary, where the professionals are better ($z = -2.27$, $p = 0.0232$), and a strong significant difference on the clarity of the entire report ($z = -3.15$, $p = 0.0016$). Finally, no significant difference was found for the layout of the report ($z = -1.02$, $p = 0.3077$) although the number for the professional teams is slightly higher.

Table 4. Results for conducting the evaluations.

	Evaluation		
Teams	Conducting the evaluation	Task quality and relevance	Questionnaire/ interviews
Student ($N = 36$)	3.42 (0.73)	3.22 (1.05)	2.72 (1.00)
Professional ($N = 8$)	**4.38 (0.74)**	3.13 (1.64)	3.50 (1.69)

Boldface numbers indicate significant differences between the student and professional teams.

4.3. Results

The pivotal result of the usability reports was the usability problems that were identified and the descriptions of them. There are eight variables on this category, see Table 6.

On the number of problems identified, the student and professional teams performed rather differently. The student teams were on average able to identify 7.9 usability problems (in the marking scale: Mean 2.56, SD 0.84) whereas the professional teams on average identified 21.0 usability problems (in the marking scale: Mean 4.13, SD 1.13). A Mann–Whitney *U* Test confirms strong significant difference between the student and professional teams on this variable ($z = -3.09$, $p = 0.002$). It is, however, interesting that the professional teams actually performed very dissimilar on this variable, as they identified from 7 to 44 usability problems. Thus, the professional team that identified the lowest number of usability problems actually performed worse than the average student team.

The professional teams performed better than the student teams on categorisation of the usability problems that were identified, but the difference is not significant ($z = -1.84$, $p = 0.0658$). On the practical relevance of the identified usability problems, the professional teams performed better, and this difference is significant ($z = -2.56$, $p = 0.0105$).

On the overview of the qualitative results, the professional teams did significantly better than the students ($z = -1.99$, $p = 0.0466$). On the other hand, the student teams provided better overview of the quantitative results, but this difference is not significant ($z = 0.90$, $p = 0.3681$).

There is no significant difference on the use of literature ($z = -0.05$, $p = 0.9601$). The conclusions are better in the usability reports from the professional teams, and this difference is strong significant ($z = -3.13$, $p = 0.0017$). No significance was found for the teams' own evaluations of the test procedure they employed ($z = -1.00$, $p = 0.3173$).

4.4. Usability problem correlations

The strong differences between the student teams and the professionals in the production of results, e.g. the usability problem identified, made us conduct a more detailed analysis of potential causes.

A Spearman Rank Correlation shows a weak positive correlation between the way the evaluation was conducted and the number of identified usability problems, but this correlation is not significant (marking ($r^2 = 0.061$, $p > 0.718$), actual ($r^2 = 0.089$, $p > 0.599$)). The same can be concluded for the correlation between the quality and relevance of the tasks and the number of identified usability problems (marking ($r^2 = 0.239$, $p > 0.157$), actual ($r^2 = 0.235$, $p > 0.165$)). Thus, our study indicates that the student's competence in planning and conducting a usability test does not necessarily influence the outcome of the evaluation in terms of the number of usability problems identified.

When looking at the corresponding variables for the professional teams, we find that there is a high correlation between the quality and relevance of the

Table 5. Results for the usability reports.

	Report					
Teams	Test description	Data quality	Clarity of problem list	Executive summary	Clarity of report	Layout of report
Student ($N = 36$)	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 ($N = 8$)	**4.00 (1.31)**	2.13 (0.83)	**3.50 (0.93)**	**3.38 (1.06)**	**4.25 (0.71)**	3.25 (0.71)

Boldface numbers indicate significant differences between the student and professional teams.

Table 6. Results for the outcome of the usability evaluations.

	Results							
Team	Number of problems	Problem categorisation	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)

Boldface numbers indicate significant differences between the student and professional teams.

tasks and the number of identified usability problems for the professional teams and this correlation is significant ($r^2 = 0.741$, $p < 0.05$). Furthermore, a weak correlation exists between the way the evaluation was conducted and the number of identified usability problems, but this correlation is not significant ($r^2 = 0.336$, $p > 0.374$).

Introducing more test subjects in usability evaluations will usually (at least in theory) generate a higher number of identified usability problems. In our study, the average number of test subjects was 3.6 (SD 0.65), ranging from one team using only two test subjects to one team using five test subjects. However, we found only a negligible positive correlation between the number of test subjects and the number of identified usability problems, as this correlation was not significant (marking ($r^2 = 0.247$, $p > 0.143$), actual ($r^2 = 0.238$, $p > 0.159$)). The test subjects had a rather varied experience with www.hotmail.com, but there is no significant correlation between the number of novice subjects and the number of identified problems (marking ($r^2 = 0.119$, $p > 0.482$), actual ($r^2 = 0.119$, $p > 0.481$)).

Correlations between the length of the tests and the number of identified usability problems for the 36 teams (grading and actual numbers) are illustrated in Figure 1. Considering the total time spent on all tests in each team, we identify a great variation ranging from 56 min to 225 min (mean = 113.26 min, SD 65.59 min). A minor correlation exists between the total time spent on the test and the number of identified problems, but the correlation is not significant ($r^2 = 0.280$, $p > 0.098$). This is also the case when looking at the actual number of problems against time spent ($r^2 = 0.329$, $p > 0.051$). This correlation is, however, close to being significant.

As a complementary perspective, we analysed the basic skills of the students and their performances in other university activities in the same semester. We examined the correlation between the combined grade obtained by each of the 36 teams (based on the individual grades of team members) in other major coursework and the number of identified usability problems.

The grade is reported on a scale from zero (not satisfactory) to nine (outstanding). A Spearman Rank Correlation Test shows only a slight positive correlation between the grade of the students and the number of identified usability problems (marking ($r^2 = 0.103$, $p > 0.542$), actual ($r^2 = 0.130$, $p > 0.441$)). This correlation between grades and identified number of usability problems is illustrated in Figure 2.

5. Discussion

As emphasised in the introduction, several early studies found that many websites suffer from low usability (Sullivan and Matson 2000), and more recent studies confirm that there are still considerable usability problems (Zaphiris and Ellis 2001, Zaphiris *et al.* 2001, Maswera *et al.* 2005, Brown *et al.* 2006). The purpose of our study was to explore to what extent

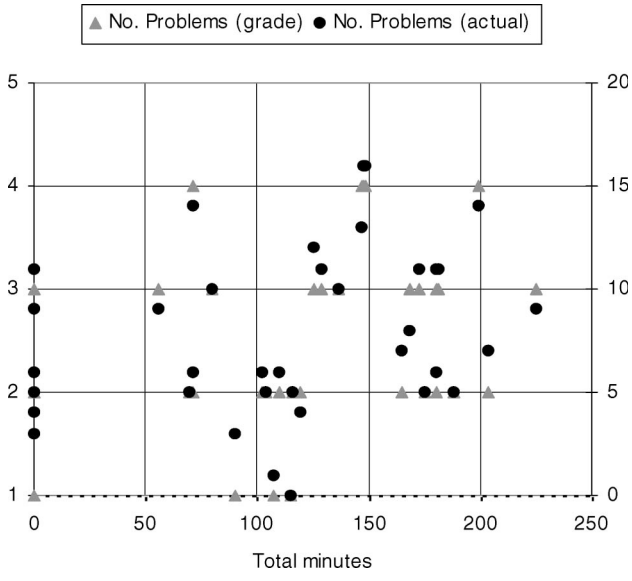

Figure 1. Correlation between the length of all tests in the 36 teams and the number of identified usability problems (reported as grading 1–5). Six teams did not report the time spent on the tests.

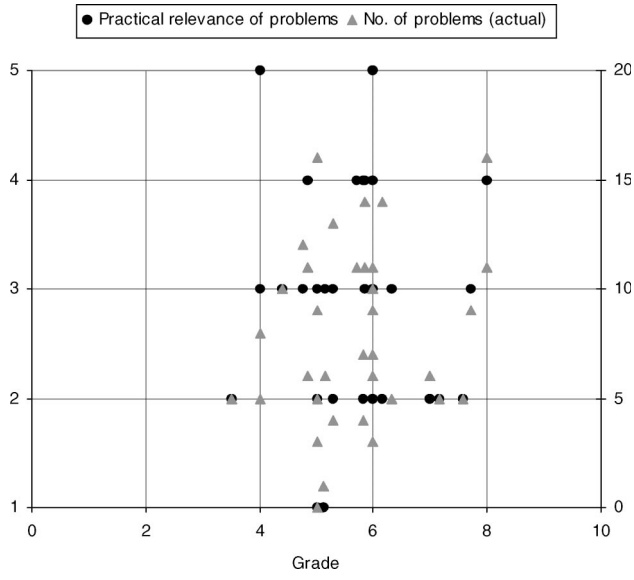

Figure 2. Correlation between the team grading (reported as zero to nine) and the number of identified usability problems (reported as grading 1–5 and the actual number identified).

people working with software development and design but with no formal training in usability engineering could be trained to conduct website usability evaluations of a reasonable quality.

One of our key findings concerns identification and categorisation of usability problems. The student teams identified significantly fewer problems than the professional teams. On average, the student teams found 7.9 usability problems, whereas the professional teams on average found 21 usability problems. This difference is important because uncovering of usability problems is a key purpose of a formative usability evaluation. The student teams did, however, perform rather differently on this variable. One student team identified no problems at all. This team might have misunderstood the assignment, but we cannot tell from their usability report, which was the basis for our analysis. The best performing students were two teams that identified 16 problems. Most of the student teams identified no more than 10 problems.

The professional teams also performed rather differently. It has been shown before that usability evaluators find different problems; this has been denoted as the evaluator effect (Hertzum and Jacobsen 2003). Yet, we also found a substantial difference in terms of the number of problems identified, and this is perhaps more surprising. One professional team identified 44 usability problems whereas another team identified only 7 problems. The latter is actually rather disappointing for a professional team. We have analysed the problems they found in more detail. The professional teams identified several critical problems on the website, but some of the critical problems were identified by relatively more student teams than professionals. For example, it was discovered by relatively more student teams that test subjects were unable to locate the functionality to change password. Thus, even though the student teams identified significantly fewer problems, they still identified some of the most severe problems on the website.

A variable that also exhibits a remarkable difference is the practical relevance of the problem list. This variable measures the extent to which the descriptions of the usability problems identified are useful for a software developer that will solve the problem. The student teams are almost evenly distributed on the five marks of the scale, and their average is 3.2. When we compare this to the professional teams, there is a clear difference. The professionals score an average of 4.6, and six out of eight teams score the top mark. This difference can, at least partly, be explained from the experience that the professionals have acquired in describing usability problems in a way that make them relevant to their customers.

Another reason for the differences between student teams and professionals in identifying and describing usability problems may be the specific design of the training course. We might have focused too little on discussing the nature of a usability problem and provided too few examples. We could also have treated this in more detail by presenting specific examples of relevant and irrelevant problems. Our analysis of the reports from the student teams clearly suggests that this topic received too little attention.

One of the key factors in the assessment of the performance of the student teams is the number of usability problems they identified. We determined this number by merging all the individual problem lists into one overall problem list, which allowed us to compare the numbers. This merging process includes combination of usability problems that have been identified separately by others as well as breaking overall problems into smaller problems. Thus, the same problems count the same way between teams. Despite these efforts to facilitate comparison, the use of problem count is discussed in the research literature. It has been argued that compared to the mere problem count it is more interesting to consider the effect or impact that the problem descriptions have on the developers of the system that is being evaluated (Hornbæk and Frøkjær 2005, Følstad 2007).

There are differences in the literature about the definition of a usability problem. The definition we have used is inspired by Molich (2007) who provides a definition of a usability problem as an aspect of the software that prevents a user from fulfilling his task. This definition is very operational as you just observe when the user is prevented from continuing. Examples of usability problems that the students found are: 'Does not understand that the Inbox is not automatically updated with new incoming emails', 'Cannot find the function to change password' and 'Try to log in as an existing user when wanting to register as a new user'. It should be noted that these statements are problems that are identified from the users' work with the website. To correct such a problems, a developer first needs to understand what the causes are and then how they problem can be resolved. This understanding of causes and generation of ideas for improvement have been studied, and it has been concluded that such redesign proposals are more useful for a developer than the mere description of the problem (Hornbæk and Frøkjær 2005). However, in our study, we only required the students to identify the usability problems. In an educational setting, generation of redesign proposals would be the next step.

6. Conclusion

This article has presented the results from a study of a course that was employed to train software developers and designers in conducting usability evaluations of a website. The idea behind this effort was that if developers can be trained to conduct usability evaluations, the gap between usability evaluation and software design will be reduced. The aim behind these efforts is to improve software development by supporting faster development–evaluation cycles and by increasing the effectiveness of usability evaluations.

The course was based on a simple approach to usability testing that quickly teaches fundamental usability skills. Whether this approach is effective has been explored through a large empirical study where 36 student teams from the first year of software development and design-oriented educations were trained in and applied the approach to evaluate the usability of the Hotmail website.

The overall conclusion is that the student teams were able to conduct usability evaluations and produce usability reports of a reasonable quality and with relevant results. However, when compared to professional evaluator teams, there were clear differences. The student teams performed well in defining good tasks for the test subjects, and the data material in their reports was significantly better than the professionals. They were less successful on several of the other variables, and they performed clearly worse when it came to the identification of problems, which is a main purpose of a usability test. It was also difficult for them to express the problems found in a manner that was relevant to a software developer working in practice.

The aim of the training course we have presented in this article is to enable software developers and designers to conduct their own website usability evaluations. The students who were trained in the approach gained a significant step towards the level of expert evaluators. However, they still lacked competence in some of the key areas. Thus, we see the training course as a relevant complement to classical usability testing conducted in a formalised manner in advanced laboratories by highly specialised experts.

Our study is limited in a number of ways. First, the environment in which the evaluations were conducted was in many ways not optimal for the usability test sessions. In some cases, the students were faced with slow Internet access that might have influenced the results. Second, motivation and stress factors could prove important in this study. The student teams were not required to produce the reports but did it voluntarily, and none of them received any payment or other kind of compensation, so their motivation may have been limited. Yet, their incentives were comparable to the expert teams from the professional usability laboratories who also participated voluntarily. Third, the demographics of the test subjects are not varied with respect to age and education. Most test subjects were $\sim$21 years of age with approximately the same school background and recently started on an IT or design-oriented education. Fourth, the usability evaluations by the students and professionals were carried out at different points in time. This might have introduced a difference, but it is only minor as the problem lists were very similar. Fifth, a main factor for assessing the performance was the number of usability problems identified. However, it has been argued that the relevance of the usability problems is more relevant. Finally, the students evaluated a website that they had not developed themselves. The purpose of training developers in usability evaluation is to enable them to evaluate their own products. This dimension was not included in the study presented here.

The use of university students as a substitute for real software developers and designers working in practice has often, and rightly, been criticised. Yet in this case, it is less questionable. With a group of software developers from practice, it would be difficult to distinguish between their experience and the effect of the training course. With students who have basic knowledge about software development but no practical experience, that empirical problem vanishes.

Having said that, it would still be very interesting to conduct a similar study with real website developers and designers. This might be combined with a longitudinal study of the long-term effect on the quality of the websites developed. The main shortcoming that came up in our analysis was the students' lack of skill in identifying and describing usability problems. A different study could be based on a training course that was changed to focus directly on identification of usability problems. On the overall level, we made an interesting observation the year after. Some of the students who participated here were two semesters later mixed with computer science students. The computer science students were first trained in programming and construction rather than evaluation. Our observation was that the students who started with evaluation seemed to perform better in a complete development project compared to those who started with programming. Thus, it would be interesting to test this observation in a controlled experiment.

Acknowledgements

This research was supported in part by the Danish Research Councils under grant number 2106-04-0022. The authors thank the students for their participation in the study. They

are also grateful to the comments from the anonymous reviewers.

References

Anderson, R.I., 2000. Making an e-business conceptualization and design process more ''user''-centered. *Interactions*, 7 (4), 27–30.

Baskerville, R. and Pries-Heje, J., 2001. Racing the e-bomb: how the internet is redefining information systems development methodology. *In*: B. Fitzgerald, N. Russo, and J. DeGross, eds. *Realigning research and practice in information systems development*. Kluwer, 49–68.

Blandford, A., Buckingham Shum, S.J., and Young, R.M., 1998. Training software engineers in a novel usability evaluation technique. *International Journal of Human–Computer Studies*, 49 (3), 245–279.

Broadbent, S. and Cara, F., 2000. A narrative approach to user requirements for web design. *Interactions*, 7 (6), 31–35.

Brown, W., Rahman, M., and Hacker, T., 2006. Home page usability and credibility. A comparison of the fastest growing companies to the Fortune 30 and the implications to IT governance. *Information Management and Computer Security*, 14 (3), 252–269.

Følstad, A., 2007. Work-domain experts as evaluators: usability inspection of domain-specific work-support systems. *International Journal of Human–Computer Interaction*, 22 (3), 217–245.

Häkli, A., 2005. *Introducing user-centred design in a small-size software development organization*. Thesis (Master's). Helsinki University of Technology.

Hertzum, M. and Jacobsen, N.E., 2003. The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human–Computer Interaction*, 15 (1), 183–204.

Hornbæk, K. and Frøkjær, E., 2005. Comparing usability problems and redesign proposals as input to practical systems development. *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'05)*, New York: ACM Press, 391–400.

Hornbæk, K. and Frøkjær, E., 2008. Making use of business goals in usability evaluation: an experiment with novice evaluators. *Proceeding of the SIGCHI conference on human factors in computing systems (CHI'08)*, New York: ACM Press, 903–912.

Hornbæk, K. and Stage, J., 2006. The interplay between usability evaluation and user interaction design. *International Journal of Human–Computer Interaction*, 21 (2), 117–124.

Høegh, R.T., *et al*., 2006. The impact of usability reports and user test observations on developers' understanding of usability data: an exploratory study. *International Journal of Human–Computer Interaction*, 21 (2), 173–196.

Howarth, J.R., 2007. *Supporting novice usability practitioners with usability engineering tools*. Thesis (PhD). Virginia Polytechnic Institute and State University.

John, B.E., 1996. Evaluating usability evaluation techniques. *ACM Computing Surveys*, 28 (4es), 139.

Karat, J. and Dayton, T., 1995. Practical education for improving software usability. *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'95)*, New York: ACM, Press, 162–169.

Law, E.L. and Hvannberg, E.T., 2008. Consolidating usability problems with novice evaluators. *Proceedings of the nordic conference on human–computer interaction (NordiCHI'08)*, New York: ACM Press, 495–498.

Maswera, T., Dawson, R., and Edwards, J., 2005. Analysis of usability and accessibility errors of E-commerce websites of tourist organisations in four African countries. *Proceedings of the International Conference on Information and Communication Technologies in Tourism*, Vienna: Springer, 531–542.

Molich, R., undated. *Comparative usability evaluation reports* [online]. Available from: http://www.dialogdesign.dk/CUE-2.htm [Accessed 30 September 2008].

Molich, R., 2007. *Usable web design*. Copenhagen: Nyt Teknisk Forlag.

Molich, R., *et al*., 2004. Comparative usability evaluation. *Behaviour and Information Technology*, 23 (1), 65–74.

Molich, R. and Nielsen, J., 1990. Improving a human–computer dialogue. *Communications of the ACM*, 33 (3), 338–348.

Nielsen, J., 1992. Finding usability problems through heuristic evaluation. *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'92)*, San Francisco, New York: ACM Press, 373–380.

Nielsen, J., 1993. *Usability engineering*. Morgan Kaufmann Publishers.

Rubin, J., 1994. *Handbook of usability testing. How to plan, design, and conduct effective tests*. New York: John Wiley & Sons.

Spool, J.M., *et al*., 1999. *Website usability. A designer's guide*. San Francisco: Morgan Kaufmann Publishers.

Sullivan, T. and Matson, R., 2000. Barriers to use: usability and content accessibility on the web's most popular sites. *Proceedings of conference on universal usability*, New York: ACM Press, 139–144.

Zaphiris, P. and Ellis, R.D., 2001. Website usability and content accessibility of the top USA universities. *Proceedings of WebNet 2001 World Conference on the WWW and Internet (WebNet)*, Orlando, FL: AACE, 1380–1385.

Zaphiris, P., Kurniawan, S.H., and Ellis, R.D., 2001. Usability and accessibility comparison of governmental, organizational, educational and commercial aging/health-related web sites. *WebNet Journal*, 3 (3), 45–52.