Direct Integration: Training Software Developers and Designers to Conduct Usability Evaluations

Mikael B. Skov and Jan Stage
Aalborg University
Department of Computer Science
DK-9220 Aalborg East
Denmark
{dubois,jans}@cs.aau.dk

ABSTRACT
Many improvements of 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 approaches involve a complete division of work between developers and evaluators, which is an undesirable complexity for many software development projects. This paper takes a different approach by exploring to what extent software developers and designers can be trained to carry out their own usability evaluations. The paper 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 in a simple approach for user-based website usability testing that was taught in a 40 hour course. This approach supported them in planning, conducting, and interpreting the results of a usability evaluation of an interactive website. They gained good competence in conducting the evaluation, defining task assignments and producing a usability report, while they were less successful in acquiring skills for identifying and describing usability problems.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Graphical user interfaces (GUI), Theory and methods.*

General Terms
Measurement, Experimentation, Human Factors.

Keywords
Usability evaluation, 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 [6].

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 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 [15] 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 [7] for an overview.

Compared to both of these approaches, website development is, however, 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. Due to this, website developers must accommodate a massive variety of user preferences and capabilities.

Many contemporary websites suffer from problems with low usability, e.g. an early investigation of content accessibility found that 29 of 50 popular websites were either inaccessible or only partly accessible [17]. 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 [16].

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 [15]. 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 duration of one month and efforts amounting to around 150 person-hours are not unusual [11][12][13]. These figures are simply not feasible for many website projects. The projects do not have this amount of

I-USED'08, September 24, 2008, Pisa, Italy

resources, and they cannot wait for the usability evaluators to conduct the evaluation and provide relevant feedback.

The two approaches that were emphasized above share a key characteristic, as they involve a complete division of work between developers and evaluators. The software is produced by the developers, and its usability is assessed by the evaluators. This division of work is undesirable or impossible in many fast-paced projects. The division of work necessitates handovers between the two groups, and this will increase project complexity and tend to lengthen development time. Thus the division between developers and evaluators is a main obstacle for integrating usability evaluation into most website development projects.

This paper 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 designers and software developers who had received a 40 hour training course could conduct a usability evaluation of a reasonable quality. In the following 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

The idea of reducing the gap between software development and usability evaluation by broadening the skills of software developers is not new. It has been suggested that education of software developers in usability engineering could contribute to reduce the problems with the usability that characterize many software products. This suggestion focused on a general awareness of usability issues and on the early activities in a development project [9].

It has also been discussed on a more general level how development teams can best be 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 [8].

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 [9]. Unfortunately, they are not complemented by the research studies of cost and effects that were also requested [8]. In fact, we have only been able to find very few systematic studies of efforts to train software developers in key topics from usability engineering.

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 [3]. This study measured the effect of a training course and also provided improved insight to the way experts work in this area.

3. METHOD

We have conducted an empirical study of a training course that is intended to teach software developers and designers to conduct usability evaluations. The aim of the study was to provide the participants with skills in formative usability evaluation.

Table 1. The 10 class meetings of the training course

#	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. The teams choose the website themselves. Experience with conducting tests and the results achieved are discussed afterwards.
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	
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. The usability evaluations are conducted at the university in assigned rooms for each team. After the usability test sessions, the teams analyze the empirical data and make a usability report that describes the identified usability problems.
6	Describing the usability testing method and how to collect and analyze empirical data	
7	Other usability evaluation methods and how to conduct a full-scale usability test session	
8	Website structures, information search and web surfing	
9	Guidelines for website design and principles for orientation and navigation	
10	Principles for visual design and different interaction styles	

3.1 Training Course

We studied the training course in a first year university curriculum. The course included ten class meetings, cf. Table 1, each lasting four hours that was divided evenly between two hours of lecture, and two hours of exercises in smaller teams. The course required no specific skills in information technology which is the reason why class meeting number one and five included introductions to technological issues. The purpose of the exercises was to practice selected techniques from the lectures. In the first four class meetings, the exercises made the students conduct small usability pilot tests in order to train and practice their practical skills 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 [14][15]. The second method was based on questionnaires that test subjects fill in after completing each task and after completion of the entire test [16]. The students were

also introduced to additional methods such as interviewing, heuristic inspection, cognitive walkthroughs, etc.

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 emphasized that the problems identified should be categorized, at least in terms of 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 due to its vast popularity. Finally, it has 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 below under Data Analysis).

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

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 four groups of students participated together in the course described in this paper. None of the participants had any experience with usability evaluation prior to the study.

36 teams involving 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. 129 (55%) of the students acted as test subjects, 69 (30%) as loggers, and 36 (15%) as test monitors, cf. [15]. 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). 42 (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 summarized in Table 2.

3.4 Setting

Due to 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.

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 emphasized 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. [15]. 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.

3.6 Data Collection

The main data collected in the study was 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, which had an average length of 9.14 pages (SD=5.02). 30 (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 minutes (SD=15.32 minutes).

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.

3.7 Data Analysis

All reports were analyzed, evaluated, and marked by the two authors of this paper according to the following three steps.

Step 1. We designed a scheme for the evaluation of the 36 reports by analyzing, evaluating and marking five 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 three 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 two-page marking guide.

Step 2. We worked individually and marked each of the 36 reports in terms of the 17 variables by using the 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=outstanding answer.

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 categorization 12) Practical relevance of usability problems 13) Qualitative results overview 14) Quantitative results overview 15) Use of literature 16) Conclusion 17) Test procedure evaluation

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 realizing their objectives. Each time a report would described such an obstacle or delay, we would count that as a usability problem. 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>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. For 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.

To examine the overall performance of the students, we included two additional sets of data in the study. Firstly, we compared the student reports to usability reports produced by teams from professional laboratories. These reports were selected from a pool of usability reports produced in another research study where nine usability laboratories received the same scenario as we used and conducted similar usability tests of www.hotmail.com, cf. [10][11]. Of the nine usability reports, we discarded one because it was only based on heuristic inspection, which was different from our focus on user-based evaluation. The remaining eight usability reports were analyzed, assessed, and marked through the same procedure as the student reports. Secondly, we calculated 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. This was done 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. 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.

Table 4. Results for conducting the evaluations. Boldface numbers indicate significant differences between the student and professional teams.

	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)

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 ($z=0.02$, $p=.984$). 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.

Table 5. Results for the usability reports. Boldface numbers indicate significant differences between the student and professional teams.

	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)

Table 6. Results for the outcome of the usability evaluations. Boldface numbers indicate significant differences between the student and professional teams.

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

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.

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 categorization 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 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 subject 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 minutes to 225 minutes (mean=113.26 minutes,

SD=65.59 minutes). 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.

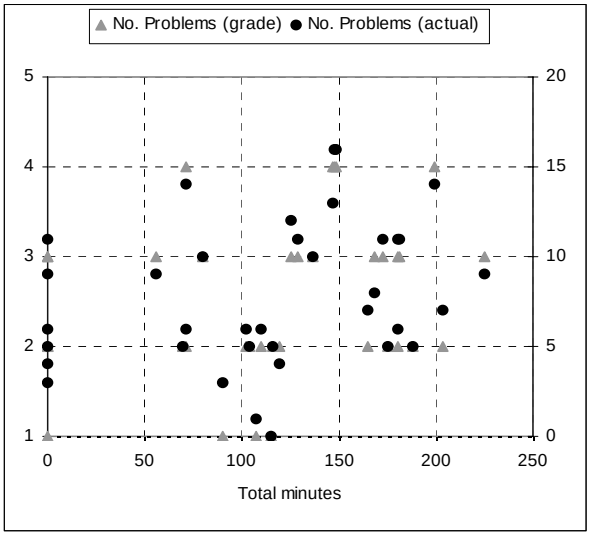

Figure 1. Correlation between the length of all tests in the 36 teams and the number of identified usability problems (reported as grading 1-5). Six teams did not report the time spent on the tests.

As a complementary perspective, we analyzed 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.

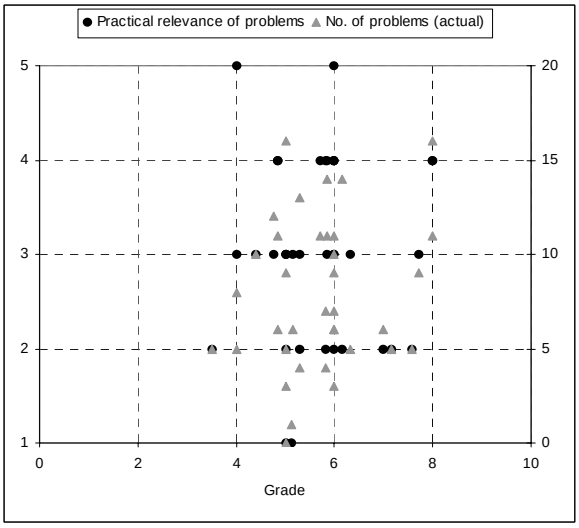

Figure 2. Correlation between the team grading (reported as zero to nine) and the number of identified usability problems (reported as grading 1-5 and the actual number identified).

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 emphasized in the introduction, several studies have found that many websites suffer from low usability [17]. The purpose of our study was to explore to what extent 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. If that was possible, such a training programme could help designers and developers face the challenges of and reduce the amount of usability problems on the websites they produce.

One of our key findings concerns identification and categorization 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 since 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 [5]. 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 seven problems. The latter is actually rather disappointing for a professional team. We have analyzed 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.

Another variable that 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 6 out of 8 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.

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 conduct their own usability evaluations, the gap between usability evaluation and software design will disappear.

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 would be relevant to a software developer working in practice.

Time pressure is a key reason why established knowledge and methodologies are ignored in many website development projects [2]. 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 [1][4]. The aim of the training course we have presented in this paper 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 formalized manner in advanced laboratories by highly specialized 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. None of the teams volunteered for the course and the study, and none of them received any payment or other kind of compensation. All teams participated in the course because it was a mandatory part of their curriculum, but they did not have to pass an exam in the course itself. This implies that students did not have the same kinds of incentives for conducting the usability test sessions as the evaluators from the professional usability laboratories. Thirdly, the demographics of the test subjects are not varied with respect to age and education. Most test subjects were approximately 21 years of age with approximately the same school background and recently started on an IT or design-oriented education.

The use of university students as a substitute for real software developers and designers working in practice has often, and rightly, been criticized. 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 could 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.

7. ACKNOWLEDGMENTS

We would like to thank the students for their participation in the study.

8. REFERENCES

[1] Anderson, R. I. (2000) Making an E-Business Conceptualization and Design Process More "User"-Centered. *interactions*, 7(4) (July-August):27-30.

[2] Baskerville, R., and Pries-Heje, J. (2001) Racing the E-Bomb: How the Internet is Redefining Information Systems Development Methodology. In N. Russo et al. (eds.), *Realigning Research and Practice in Information Systems Development*, Kluwer, 49-68.

[3] 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.

[4] Broadbent, S., and Cara, F. (2000) A Narrative Approach to User Requirements for Web Design. *interactions*, 7(6):31-35 (November-December).

[5] 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.

[6] 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.

[7] Høegh, R. T., Nielsen, C. M., Overgaard, M., Pedersen, M. B. and Stage, J. (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.

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

[9] 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)*, 162-169. ACM Press.

[10] Molich, R. (undated) Comparative Usability Evaluation Reports. Available at http://www.dialogdesign.dk/cue.html.

[11] Molich, R., Ede, M. R., Kaasgaard, K. and Karyukin, B. (2004) Comparative Usability Evaluation. *Behaviour & Information Technology*, 23(1):65-74.

[12] Molich, R., and Nielsen, J. (1990) Improving a Human-Computer Dialogue. *Communications of the ACM*, 33(3): 338-348.

[13] Nielsen, J. (1992) Finding Usability Problems Through Heuristic Evaluation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'92)*, 373-380. ACM Press.

[14] Nielsen, J. (1993) *Usability Engineering*. Morgan Kaufmann Publishers.

[15] Rubin, J. (1994) *Handbook of Usability Testing. How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons.

[16] Spool, J. M., Scanlon, T., Schroeder, W., Snyder, C., and DeAngelo, T. (1999) *Website Usability. A Designer's Guide*. Morgan Kaufmann Publishers.

[17] 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*, 139-144. ACM Press.