

USABILITY EVALUATION

OF MOBILE SYSTEMS AND PROVIDING FEEDBACK TO DESIGNERS



A P P E N D I X B

GROUP E3-211 - JUNE 2004
AALBORG UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE

AALBORG UNIVERSITY, DENMARK.



INFORMATICS - 10TH SEMESTER – MASTER THESIS.

TITLE:

Usability Evaluation of Mobile
Systems and Providing Feedback
To Designers

PROJECT PERIOD:

5th of February, 2004 –
10th of June, 2004

GROUP E3-211:

Christian Monrad Nielsen
Michael Overgaard
Michael Bach Pedersen
Sigge Stenild

SUPERVISOR:

Jan Stage

NUMBER OF PAGES:

Summary &
Appendix A: 62
Appendix B: 182

NUMBER OF COPIES:

8

ABSTRACT:

This master thesis investigates if any applicable methods for evaluating mobile systems exist, and how feedback to the designers of the system can be provided. This is done by conducting a literature review of past research within the area of Human-Computer Interaction with focus on user-based usability evaluations. The study provides an overview of tendencies in this area, and the results from the study are utilized in the design and conduction of two usability evaluations; one in a laboratory and the other one in field settings. The purpose of these evaluations is to investigate how, and in which settings, evaluations of a mobile system can be conducted. Two usability reports, documenting the usability problems identified in the respective evaluations, are used as a foundation for providing feedback to the designers of the mobile system. The result of the thesis is a presentation of an applied method for evaluating mobile systems, and insight into how usability reports can be used as a mean for providing feedback to designers.

SUMMARY

Two overall topics are addressed in this master thesis. The first topic is how to evaluate mobile systems and how to manage some of the challenges relating to the evaluation of such systems. The second topic is concerned with a different aspect, relating to all types of usability evaluations, mobile systems as well as traditional desktop systems, namely how the result from an evaluation can be used to provide feedback to the designers of a system.

To investigate how others have addressed the challenges of evaluating mobile systems, we perform an extensive study of published papers within the area of HCI. The results of this study, combined with literature on usability evaluations in general, were used to come up with a method for evaluating a specific mobile system. The mobile system is used for registering e.g. materials, time, and mileage and a number of different fields of work. The Danish software company Net-Mill, located in Aars, is in the process of developing the system and was interested in having the mobile part of their product usability evaluated. Two of the developers from the company also agreed to participate in an experiment, which allowed us to study how feedback can be given to the designers based on our usability evaluations.

The results from the usability evaluation is also to be used by a Ph.D. student at the Department of Building Technology and Structural Engineering at Aalborg University, who is working on a project describing how information technology can be implemented on construction sites and in the companies working there. Through this Ph.D. student, contact was established to a school in Horsens, which were very interested in cooperating with us by providing the necessary number of participants for the usability evaluation. During the entire process of writing this thesis, cooperation with external partners have been paramount, since this allowed us to gain insight into how the result from a usability evaluation can be used in real-life software development projects.

The results of this thesis fall within three areas. The first results are an overview of relevant approaches on how to perform user-based usability evaluation of mobile systems found in published papers. The second results are experiences gained during a practical usability evaluation and a comparison between two evaluations of the same system. The third results are a number of lessons learned from a concrete experiment on how to provide usable feedback to the designers of a system.

This thesis has not been documented through a traditional report, but instead it consists of three individual research papers and a summary. The papers are can be read in any

SUMMARY

order, and as stand-alone papers, but the intend is that the summary, in hand, elaborates and explains the overall relation between the papers and presents the essential results. Through this process of elaboration, three subordinate research questions are addressed, which in the end serves as a foundation for answering a more general research question.

PREFACE

This master thesis deals with usability evaluation of mobile systems and how to provide feedback to the designers, based on the results of the evaluation. The thesis consists of this summary and three individual research papers.

The elaboration of this thesis would not have been possible if it was not for the inputs, comments, and support that we received from the many people involved. First and foremost we would like to thank our supervisor Jan Stage for always providing detailed and constructive feedback on both theoretical and practical issues concerning the content and structure of this thesis. Furthermore, we would also like to thank Mikael B. Skov for reviewing and commenting the methodological approach of our evaluations, and Rune T. Høegh for providing technical assistance in relation to the usability laboratory and the mini-camera used in our evaluations. A great acknowledgement is given to Mads Carlsen, Department of Building Technology and Structural Engineering, for his involvement in creating contact with Net-Mill International A/S and Vitus Bering CEU. Appreciation should also be given to the employees at Net-Mill International A/S for their positive cooperation and their participation in the feedback sessions. Ole Math from Vitus Bering CEU should also be thanked for his involvement in acquiring test participants and setting up the evaluations. In addition to this, we thank the participants for their willingness to participate in the evaluations. Rolf Molich should be thanked for explaining his view on feedback through usability reports. Lastly, Aage Nielsen, Department of Mathematical Science, is thanked for his advice and explanations about performing statistical analysis.

Bibliographical reference format:

All references are enclosed in [...]. References, such as Jensen [2000] or [Jensen, 2000], are used for referencing articles or books in References at the end of the thesis. Appendix A is located in this thesis, while Appendix B is available as a separate report.

Christian Monrad Nielsen

Michael Overgaard

Michael Bach Pedersen

Sigge Stenild

CONTENTS

1	INTRODUCTION	1
1.1	USABILITY EVALUATION METHODS	1
1.2	RESEARCH QUESTIONS	3
1.2.1	<i>Research Question 1</i>	3
1.2.2	<i>Research Question 2</i>	3
1.2.3	<i>Research Question 3</i>	4
2	RESEARCH PAPERS	5
2.1	RESEARCH PAPER 1	
	A REVIEW OF LITERATURE ON USABILITY EVALUATION METHODS FOR MOBILE SYSTEMS	6
2.2	RESEARCH PAPER 2	
	USABILITY EVALUATION OF A MOBILE SYSTEM: COMPARISON OF A LABORATORY AND A FIELD EVALUATION	7
2.3	RESEARCH PAPER 3	
	PROVIDING FEEDBACK TO DESIGNERS: ARE USABILITY REPORTS ANY GOOD?	8
3	RESEARCH METHODOLOGY	11
3.1	RESEARCH QUESTION 1	11
3.2	RESEARCH QUESTION 2	12
3.3	RESEARCH QUESTION 3	12
4	CONCLUSION	15
4.1	RESEARCH QUESTIONS	15
4.2	LIMITATIONS	17
4.3	FUTURE WORK	17
	REFERENCES.....	19
	APPENDIX A.....	23
	RESEARCH PAPER 1	
	A REVIEW OF LITERATURE ON USABILITY EVALUATION METHODS FOR MOBILE SYSTEMS	23
	RESEARCH PAPER 2	
	USABILITY EVALUATION OF A MOBILE SYSTEM: COMPARISON OF A LABORATORY AND A FIELD EVALUATION	41
	RESEARCH PAPER 3	
	PROVIDING FEEDBACK TO DESIGNERS: ARE USABILITY REPORTS ANY GOOD?	53

1 INTRODUCTION

The concept of usability originally emerged as a result of intense research into the use of advanced technology during the Second World War. Researchers realized that the adaptation of machines to the human operator would increase human-machine reaction, performance and speed [Dix et al., 1998:2]. It also became apparent that machines could aid human cognition, and ideas emerged on ways in which all sorts of information could be displayed on screens [Bush, 1945] [Engelbart, 1962]. These ideas would later form the basis of today's computer interfaces [Card et al., 1983] and usability has, for several years, been the main subject of interest in the field of Human-Computer Interaction (HCI) [Carroll, 2001].

Usability is a combination of several factors that can be used to measure the quality of a user's experience when interacting with a product or system. It is defined as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use [ISO, 1998]. Usability is important, from the user's perspective, since it can make the difference between performing a task accurately or not, and to enjoy the process or being frustrated. From the developer's point of view, emphasizing on usability can be the decisive factor between the success or failure of a system. From a management perspective, productivity among the employees might decrease immensely, if the applied software has poor usability. Common to all of these perspectives are that the lack of usability can cost both time and effort, and can have great influence on the success or failure of a system [Karat, 1994] [Mayhew, 1999:449-482].

1.1 USABILITY EVALUATION METHODS

The purpose of usability evaluation methods are to evaluate the interaction of the human with the computer with the objective of identifying aspects of this interaction that can be improved to increase usability [Gray & Salzman, 1998]. Several methods for performing usability evaluation have been presented in the literature; heuristic evaluation [Nielsen & Molich, 1990], cognitive walkthrough [Lewis et al., 1990] [Wharton et al., 1992], which are expert-based evaluation techniques, while think-aloud evaluation [Rubin, 1994] [Molich, 2000] and observation [Nielsen, 1993] [Molich, 2000] are evaluation techniques based on user participation. Studies show that the two approaches identify different kinds and numbers of usability problems [Doubleday et al., 1997] [Kjeldskov & Skov, 2003], and that expert evaluations tend to discover more problems than through user-based testing, but problems discovered during user-based testing is more likely to be true

usability problems [Bailey, 1992]. The reason could be that experts have difficulties in predicting which problems the users might encounter [Norman, 2002:157], since users often have domain specific knowledge that the experts do not possess [Croft, 1986]. Although user participation in the evaluation of a system can be both time-consuming and costly, involving the users in the system development process can lead to increased user satisfaction and increased system usage [Baroudi, 1986].

In general, usability evaluations are often conducted as a part of an iterative approach in the development process, where evaluations are used to progressively refine the design of a system [Dix et al., 1998:187-189]. Usability evaluations enable the designer to incorporate changes in the system, based on the feedback from the evaluators, and thereby increase the level of usability in the system [Mayhew, 1999:229-230].

Reporting usability problems is often done through a usability report [Dumas & Redish, 1993] [Rubin, 1994], which constitutes the feedback from the evaluation to the designers. Investigating this part of the usability engineering lifecycle, illustrated as “Feedback” in Figure 1, can be seen as a kind of meta-usability evaluation of the usability report itself, and the way it is presented to the developers. What good is a report if it can only be understood by the HCI-professionals, who wrote the report? After all it is the developers that have to understand and fix the problems eventually. Therefore, it is important to explore how the evaluators, who performed the evaluation, can explain and convince the developers that the problems discovered in the evaluation and described in the usability reports, are in fact real problems experienced by real users.

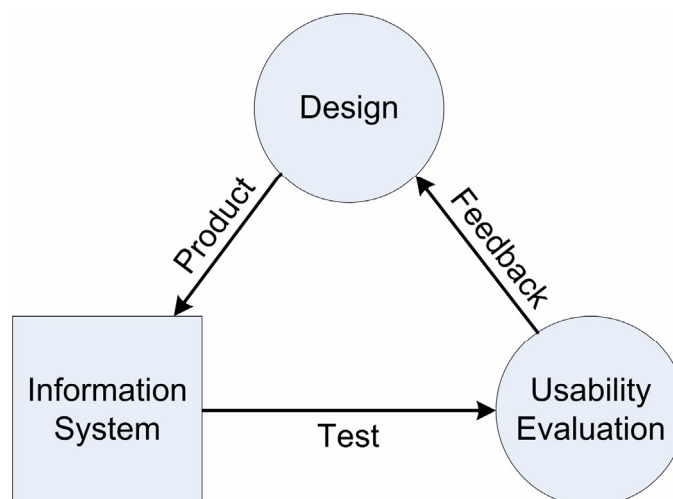


Figure 1: The simplified usability engineering lifecycle.

Usability evaluation of stationary systems is a well-established discipline within the area of HCI. In the 1980's, laboratory usability testing was the primary usability evaluation method to examine interfaces, because of the possibility to create a controlled environ-

ment allowing the evaluators to collect different kinds of data [Hartson et al. 2001], e.g. user performance and user preference [Rubin, 1994:104-106]. The spreading and increased popularity of mobile systems has introduced a number of challenges for performing usability evaluations of such systems. Mobile systems are often used in highly dynamic contexts [Kjeldskov & Stage, 2004], which makes it enticing and relevant to conduct field-based evaluations, but performing field-based usability evaluations is not an easy task [Nielsen, 1998] [Brewster, 2002]. Data collection and applying evaluation techniques, such as think-aloud and observation, is far from trivial in a field-based evaluation, since the test participant is physically moving around in the environment [Kjeldskov & Stage, 2004].

In Nielsen et al. [2004] we described the design, implementation, and evaluation of a mobile system used for communication and collaboration in a safety-critical domain. The evaluation were based on two different approaches; a heuristic inspection and a think-aloud evaluation, involving the end-users, and conducted in field settings. Through this project, we were faced with the problems and challenges of performing usability evaluation of mobile systems and how to use the results to improve future designs.

1.2 RESEARCH QUESTIONS

The overall research question in this thesis is to discover, whether any applicable methods for evaluating mobile systems exist, which can provide usable results for the designers. In order to answer this, we present the following three research questions.

1.2.1 RESEARCH QUESTION 1

Usability evaluation of mobile systems has been an area of research for some years, but the research is not yet well-established on a methodological level. In order to benefit from the lessons learned by other researchers and create an overview of the options available, the first question of this thesis is:

What has been published on user-based usability evaluations of mobile systems in key HCI journals and conference proceedings?

1.2.2 RESEARCH QUESTION 2

In order to study the practical application of methods for evaluating mobile systems, we need to conduct a usability evaluation of such a system, analyze the data, and finally pro-

duce usability reports based on the evaluations. This justifies the second research question of this thesis:

How and in what setting can a usability evaluation of a mobile system be conducted?

1.2.3 RESEARCH QUESTION 3

If the results from a usability evaluation are to be used in the further development of a system, it is important that the designers understand and acknowledge the usability problems discovered during the evaluation. This issue is addressed in the third research question:

How can usable feedback from a usability evaluation be provided to the designers of a system?

The three research questions will be addressed in the three research papers summarized in chapter 2.

As mentioned previously, many types of usability evaluation methods exist, but in this thesis we focus on methods involving user participation. Furthermore, the evaluation will be performed in different settings, and usability reports will be produced, one for each of the evaluations. This is done in order to compare the two types of evaluations, and to discover, whether the recipients of the two different reports have specific preferences towards either of the evaluation approaches and their results. Our focus is not merely to evaluate a mobile system, but also to explore how the problems found, is successfully reported back to the designers. As a part of this, we want to discover whether the designers understand the nature of the problems, and whether they consider a usability report as an asset in the further development.

2 RESEARCH PAPERS

This chapter presents the three individual research papers of the thesis, which are included in Appendix A. The relation between the three research papers is illustrated in Figure 2.

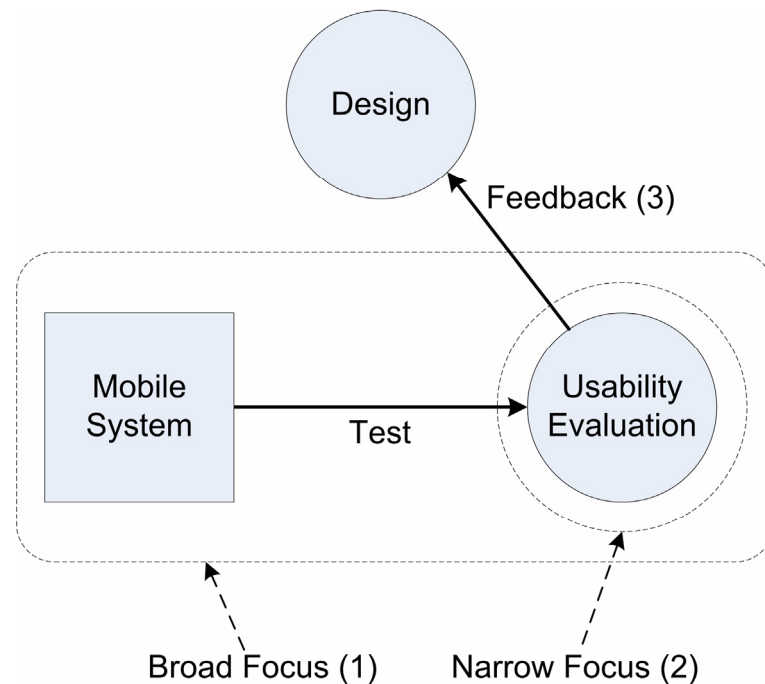


Figure 2: The relation between the individual research papers.

The papers explore user-based usability evaluations of mobile systems based on a literature review (Broad Focus), the process of performing usability evaluations of a mobile system (Narrow Focus), and the process of providing feedback to the designers of the mobile system (Feedback). The three research papers relate to the three research questions, described in section 1.2, accordingly.

- 1) Nielsen, C. M., Overgaard, M., Pedersen, M. B. & Stenild, S. (2004). A Review of Literature on Usability Evaluation Methods for Mobile Systems. Department of Computer Science, Aalborg University, 2004.
- 2) Nielsen, C. M., Overgaard, M., Pedersen, M. B. & Stenild, S. (2004). Usability Evaluation of a Mobile System: Comparison of a Laboratory and a Field Evaluation. Department of Computer Science, Aalborg University, 2004.
- 3) Nielsen, C. M., Overgaard, M., Pedersen, M. B. & Stenild, S. (2004). Providing Feedback to Designers: Are Usability Reports Any Good? Department of Computer Science, Aalborg University, 2004.

2.1 RESEARCH PAPER 1

A REVIEW OF LITERATURE ON USABILITY EVALUATION METHODS FOR MOBILE SYSTEMS

The first paper reviews literature on usability evaluation methods for mobile systems. Out of 1826 papers from key journals and conference proceedings, 58 papers were identified, which described a user-based usability evaluation of a mobile system. The procedure of identifying the papers, relevant to our study, was done in a number of steps after which the relevant papers were categorized according to 12 categories. These categories provide an overview of the papers and their content.

Statistical tests showed a very significant difference between the numbers of papers performing either laboratory-based evaluations or field-based evaluations, with field evaluation as the most commonly applied type of evaluation. Additionally a minor part of the papers performing laboratory evaluation tried to recreate a field-like experience in the laboratory. Initially, the study showed no significant difference between the use of data collection techniques between laboratory and field evaluations, with the exception of interviews, which were conducted more frequently in the field-based evaluations. On closer inspection, we discovered that video was used to record different aspects in laboratory and field evaluations respectively. The use of video in laboratory settings focused on the screen of the device and the user's interaction, while recordings in field evaluations captured the user's movement and interaction with other individuals. The study also found that when analyzing the data collected in the usability evaluations, 32% conducted various types of statistical analysis and 22% performed some kind of time measurement analysis, such as task completion time. Qualitative approaches, such as conversations analysis, interaction analysis, transcriptions, and socio technical approaches, were utilized in only 16% of the papers. Finally, it was discovered that in all of the 58 papers, the designers and the evaluators were the same individuals, which is presumably why none of the papers described how the evaluators provided feedback to the designers about the results of the evaluations.

The paper concludes that there is a significant bias towards performing the usability evaluation in realistic settings, with field-based evaluation as the preferred type of evaluation. Concerning the data collections methods, interview was more frequently used in field evaluations, and those using video recordings did so with different purposes, when comparing laboratory and field evaluations. The study showed no difference between laboratory and field evaluations concerning other data collection methods. Furthermore, quantitative data analysis methods are more often used than qualitative methods.

2.2 RESEARCH PAPER 2

USABILITY EVALUATION OF A MOBILE SYSTEM: COMPARISON OF A LABORATORY AND A FIELD EVALUATION

The second paper explores how to conduct usability evaluations of a mobile system. The reported results originate from two user-based usability evaluations, in which a mobile system was evaluated, in both laboratory and field settings. The two evaluations were conducted by two separate evaluation teams to ensure independence between the evaluations. The test participants were two different groups of apprentices. The evaluation design was the same in both evaluations in order for the evaluators to minimize influence on the result from differences in design. A problem list was produced for each of the evaluations and the problems were categorized according to severity and themes. Furthermore, usability aspects, described in the ISO 9241-11 definition, were utilized when comparing the laboratory and field evaluations.

The reluctance towards field-based evaluations, have been the lack of control and the complicated data collection, however data collection during our usability evaluations demonstrated that by using a mini camera mounted on the mobile phone, it was possible to get a good view of the screen and the users' interaction with the mobile phone. This was particularly successful in the field evaluation, where the small screen of a mobile phone, in addition to the free movement of the users, presents evaluators with difficult data collection conditions.

The problem lists from the evaluations revealed that the field-based evaluation uncovered the most usability problems, as it identified 60 problems compared to the 48 problems found in the laboratory evaluation. Furthermore, the field evaluation identified problems within two themes; 'cognitive load' and 'interaction style', which did not appear in the laboratory evaluation. The laboratory evaluation did not discover any problems within themes not also discovered in the field evaluation. The comparison of the overall usability, according to the ISO 9241-11 definition, showed that the system got a lower overall usability rating in the field-based evaluation. From a joint list of problems (see Appendix B), based on the two evaluations, it can be seen that 58% of the problems were unique, which indicate that both evaluations are important, if a broad and varied measurement of the usability problems of mobile systems is to be obtained. Through a combination of severity, distribution of unique problems, and problems found by both evaluations, the comparison showed that the more severe a problem is, it is more likely to be identified in both evaluations.

The paper concludes that it can be beneficiary to conduct evaluations of a mobile system in field settings, based on the amount of problems found, although critical problems are

likely to be identified in both types of evaluations. Furthermore, the paper concludes that it is possible, when conducting evaluations of mobile systems, to reduce the problems of complicated data collection that small screens of mobile phones and field evaluations impose.

2.3 RESEARCH PAPER 3

PROVIDING FEEDBACK TO DESIGNERS: ARE USABILITY REPORTS ANY GOOD?

The third paper explores usability reports as a mean for providing feedback to the designers about usability problems identified in a user-based usability evaluation. In order to do this, two separate usability reports were used, which can be found in Appendix B. A feedback experiment was conducted in order to observe any changes in the developers' understanding of the usability problems, as they read and reviewed the reports. The feedback session was conducted in cooperation with two developers from a software company, who had developed the system evaluated in the reports.

Both developers used the same approach when reading the usability reports, which was straightforward from the beginning to the end, while using the appendices, and the logs to clarify details, when something was unclear. After each of the three feedback sessions, the developers were asked to describe the advantages and disadvantages of the system, and through these descriptions, it was observed that the developer perception of both advantages and disadvantages of the system changed, as they read the usability reports.

In the interviews, both developers mentioned the problem list, rated by theme, severity, and the number of users, and the elaborating descriptions of the problems as very important for their understanding of the problems. They stated that these parts of the usability reports were very useful in their future work on the system. The developers mentioned that the log files were important for providing further insight into problems, although they could not be used directly to resolve the problems. Furthermore, they mentioned the log files, as important in respect to how they rate the validity of the usability evaluations. The developers did not value the general assessments of the system much, as they mentioned that the summary and the conclusion were amusing to read, but not very useful. The designers found that the NASA-TLX data was difficult to understand, and they were uncertain how the results should be interpreted, because it was not put into context.

The study concludes that the problem list, along with the detailed descriptions of the usability problems, is important and essential for the designers when trying to understand the usability problems described in a usability report. Log files, the description of the

evaluation design, and the test participants' subjective opinions add further insight to the problems in the report, while general assessments of the system and NASA-TLX results were less useful to the designers.

3 RESEARCH METHODOLOGY

This chapter presents and elaborates on the methodological approaches and epistemological considerations pertaining to the three research questions.

No single correct research method can be said to exist for any of the three questions addressed in this report. According to Galliers & Land [1987], what is important, is to understand and recognize the potentials and limitations associated with each method. Research within Information Systems spans many disciplines [Galliers & Land, 1987], but upon closer inspection of the focus in this thesis, some types of research methods may be more applicable than others. Table 1 provides an overview of the research methods in relation to the three research questions. The separation of *method* from *object* in the table is inspired by Galliers & Land [1987] and Kjeldskov [2003]. Following the table, research methods in relation to each research question are discussed.

Research				
Question	Object	Method	Purpose	Setting
#1	Literature	Survey research	Describe, Understand	Environment independent
#2	Mobile System, Usability evaluation method	Laboratory experiments, Applied research	Explore, Compare	Artificial setting, Natural setting
#3	Designers, Usability reports	Case study, Applied research	Hypothesis development, Describe	Natural setting, Environment independent

Table 1: Research methods in response to the three research question.

3.1 RESEARCH QUESTION 1

In response to research question 1, about what other researchers have done in order to evaluate mobile systems, we conduct a literature survey of past research on user-based usability evaluation of mobile systems. By examining and classifying relevant research papers, through a number of predefined steps, an overview of tendencies in current research is achieved. Hence, the purpose of the research in paper 1 is to describe and understand past research, which according to Galliers [1992] ensures that future research is built on past endeavours.

The epistemological foundation behind this paper is exploratory quantitative research. We assume that it is possible to classify and categorize papers objectively, but acknowledge that some articles require a degree of interpretation due to missing information or vague descriptions. One of the disadvantages of the survey research method is the un-

known that the researchers have when analyzing, sorting, and categorizing the data. To counter these issues to the extent possible, the analyzing and sorting process is done in two steps, and the categorisation was done in pairs of two persons. On the basis of aspects of interpretation, we have strived to make the limitations to the survey explicit.

3.2 RESEARCH QUESTION 2

The second research question has a twofold purpose. First of all, it is to identify and document usability problems related to a specific mobile system. Secondly, the goal is to compare the results of conducting user-based usability evaluations in both laboratory and field settings. In order to explore and gain insight into which problems the intended users might experience, evaluations of the system have to be conducted. This process falls within the category of applied research, as the evaluation of the mobile system involves exploring novel data collection techniques to gain experience when using these new techniques [Wynekoop & Conger, 1990]. The experiences are documented in both research paper 2 and the two usability reports in Appendix B.

We are aware that, by being an active part of the evaluations, we can no longer be considered being objective, which contradicts the positivistic assumption of the objective observer [Myers, 1997] [Dahlbom & Mathiassen, 2000:209-210]. On the other hand, our data collection technique is mainly quantitative in order to obtain measures, as objectively as possible, for us to be able to compare the evaluations [Straub et al., 2004]. This is done through e.g. statistical calculations. Additionally, it is clear that in the identification and the severity rating of usability problems, a high degree of interpretation takes place. As a mean to counter this subjective influence, each evaluation session is interpreted by two persons, who, in cooperation, work out a problem list on the basis of all the sessions. A joint problem list is elaborated from the problem lists from both evaluations, which makes that a comparison is possible. Joining the lists requires some interpretation of the problems, which is why this is a joint effort by a person from each evaluation team. The joint problem list can therefore be viewed as qualitative data [Myers, 1997]. Finally, we triangulate the statistical measurements and the results of the joint problem list, since measurements alone are not sufficient in describing the evaluation situation [Straub et al., 2004]. Hence, we believe that we obtain more robust results and a fuller picture of the evaluations through triangulating [Kaplan & Duchon, 1988].

3.3 RESEARCH QUESTION 3

In research question 3, we perform a case study in order to gain insight into how two usability reports are used for feedback purposes to designers of a mobile system.

Through interviews, we collect rich data that could explain the casual relations [Wynekoop & Conger] between the reports, the designers, and the feedback. However, since this qualitative data is very rich on details, it is necessary to process it, which was done by a transcription. Figure 3 illustrates this process and the corresponding levels of detail. One drawback with this approach is that it implies loss of some of the information in the original data. This is due to that transcriptions cannot be considered completely objective and that transcriptions of audio recordings are inherently less rich than the original recordings [Alrø & Dirckinck-Holmfeld, 1997:81-83].

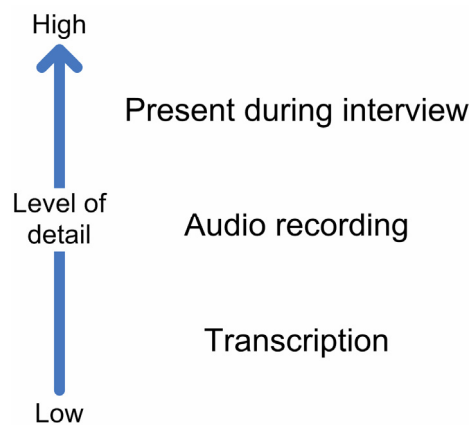


Figure 3: Type of data.

Qualitative data collected needs interpretation, and therefore cannot be judged using the same criteria as traditional positivistic quantitative research [Markus & Lee, 1999]. On the other hand this type of data can be used to develop new hypothesis [Wynekoop & Conger, 1990] on the structure and content of usability reports.

4 CONCLUSION

This master thesis deals with HCI, with focus on user-based usability evaluation of mobile systems, conducted in both laboratory and field settings, and how to provide effective feedback to the designers of the system. Firstly, this chapter summarizes the conclusions of the research conducted in accordance with the three research questions, stated in section 1.2. Secondly, limitations of the approach and the results are discussed and avenues for further research are proposed.

4.1 RESEARCH QUESTIONS

The evaluation of mobile systems involves the activities of designing and conducting the evaluation, and how to use the results in the further development of the system. We find that these activities introduce a number of challenges for the evaluators, since mobile systems are often used in different and highly dynamic contexts, which should be taken into account during the evaluation. Additional challenges are related to getting the designer to understand the problems, experienced by the users, are in fact real usability problems. The three research questions are addressed by the three research papers accordingly and the primary results of this thesis are summarized in the subsequent paragraphs.

Research question 1: What has been published on user-based usability evaluations of mobile systems in key HCI journals and conference proceedings?

The first research question revealed a significant bias towards performing usability evaluations in field settings. This tendency indicates that most researchers acknowledge the importance of field-based usability evaluations for evaluating mobile systems in their context of use. Additionally, a minor part of the papers performing laboratory evaluations tried to recreate a field-like experience. This further stresses the preconceived notion of using realistic settings for conducting evaluations of mobile systems. Interview was more often used for data collection in field evaluations than in laboratory evaluations, and the use of video recordings had different purposes in laboratory evaluations compared to field evaluations. Apart from this, there was no difference between the two types of evaluations, concerning the use of methods for data collection. In relation to methods for data analysis presented in the papers, the study revealed a bias towards using quantitative data analysis methods, relative to qualitative methods.

Research question 2: How and in what setting can a usability evaluation of a mobile system be conducted?

The second research question revealed that the field-based evaluation uncovered more usability problems than the laboratory-based evaluation, and the system got a lower overall usability rating in the field evaluation. These findings indicate that the field evaluation is better suited to find usability problems in a mobile system, than a laboratory evaluation, when considering the amount of problems found. 58% of the problems were unique, laboratory and field evaluations combined, and the study demonstrated that the more severe a problem was, it was more likely to be identified in both evaluations. Furthermore, categorizing the identified problems, according to usability themes, was helpful in determining which areas of the system the problems revolved around.

Research question 3: How can usable feedback from a usability evaluation be provided to the designers of a system?

In the third paper the designers emphasized on the list of usability problems and the related detailed descriptions of individual problems, as being most useful in their further development of the system. Furthermore, the two different kinds of log files, based on the video recordings and system events, were important if information about the cause and circumstances of a problem was to be obtained. On the other hand, the designers had difficulties in interpreting the standalone results of the NASA TLX test, because they were not explained and elaborated on in the two usability reports, to the extent needed. In addition to this, the general assessments, e.g. the executive summary and the conclusion, had little value to the designers in relation to understanding of the problems discovered in both evaluations. This leads to the conclusion that it was possible to provide usable feedback to the designers, through the use of a problem list and the detailed descriptions of the distinct problem.

The overall research question of this thesis was to discover, whether any applicable methods for evaluating mobile systems exist, which can provide usable results for the designers. To answer this question, we have identified a number of research papers, which, combined with general literature on how to design and perform usability evaluations, enabled us to perform a user-based evaluation of a specific mobile system. The specific approach involved evaluation in both laboratory and field settings, using the same quantitative and qualitative methods for data collection. The results of our feedback experiment showed that usability reports, based on user-based usability evaluations, can provide usable feedback to the designers. Combined with the lessons we have learned in relation to the content and structure of usability reports, we believe that the method we

have presented are applicable for evaluating mobile systems, and that it can provide the designers with usable feedback in the form of a usability report.

4.2 LIMITATIONS

The first limitation relates to the notion of laboratory settings versus field settings. Our evaluations are performed in a dedicated laboratory and in a warehouse. The intention with the warehouse was to perform the evaluations in a context closer to the real world. It can be argued that both evaluations should be considered as laboratory evaluations, but we find that the distinction between laboratory and field is not necessarily either/or, but that many different types of field-settings exist with varying degrees of realism. The warehouse did undoubtedly provide a more field-like experience, than in the usability laboratory, but it should be taken into consideration that it is unknown what the results of a field evaluation would have been, if the context had been even more realistic.

The second limitation is concerned with the fact that all of our experiments are subject to limited generalizability due to the limitations of involving only two designers, one system, and two usability reports. More generalizable results can only be achieved through a number of similar experiments, which can either confirm or dismiss the results presented in this thesis.

The third limitation is related to the choice of usability evaluation approach and the type of feedback. The overall research question of the thesis was to discover, whether any applicable methods for evaluating mobile systems exist, which can provide usable results for the designers. It was necessary to limit the scope, which was done by choosing a user-based usability evaluation approach, which means that the result is only valid within the scope of user-based usability evaluations. Furthermore, the second part of the overall research question was limited by choosing usability reports as the type of feedback. This implies that the results concerning feedback have unknown applicability to other ways of providing feedback than usability reports.

4.3 FUTURE WORK

The first and somewhat obvious suggestion for future work, if a higher degree of generalizability was to be achieved, would be to perform a number of similar experiments. The second suggestion is to examine whether the feedback presented to the designers results in a system with a higher degree of usability. One way of doing so is through action research into how descriptions in a usability report can be combined with the type of development methods used in the studied software company.

The third suggestion for future research is to apply the same overall experimental design in relation to e.g. heuristic inspection, to examine its ability to identify usability problems in mobile systems, and whether the result can be used to provide useful feedback to designers. The ultimate purpose of such an experiment might then be to compare user-based and expert-based evaluations methods and their ability to generate effective feedback.

REFERENCES

- Alrø, H. & Dirckinck-Holmfeld, L. (Ed.) (1997). *Videoobservation*. Aalborg Universitetsforlag.
- Bailey, R.W., Allan, R.W. and Raiello, P. (1992). Usability testing vs. heuristic evaluation: A head-to-head comparison. *Proceedings of the Human Factors Society 36th Annual Meeting* (1992).
- Baroudi, J. J., Olson, M. H: & Ives, B. (1986). An empirical study of the impact of user involvement on system usage and information satisfaction. *Communications of the ACM*, Volume 29 Issue 3.
- Brewster, S. (2002). Overcoming the Lack of Screen Space on Mobile Computers. *Personal and Ubiquitous Computing*, 6: 188-205.
- Bush, V. (1945). *As We May Think*. July 1945 issue of *The Atlantic Monthly*.
- Card, S. K., Moran, T. P. & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Carroll, J. (2001). *Human-Computer Interaction in the New Millennium*. Published by Addison Wesley Professional, ACM Press.
- Croft, W. B. (1986). User-specified domain knowledge for document retrieval. *Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval*.
- Dahlbom, B. & Mathiassen, L. (2000). *Computers in Context – The Philosophy and Practice of Systems Design*. Blackwell Publishers Inc.
- Kaplan, B. & Duchon, D. (1988) *Combining Qualitative and Quantitative Methods in Information Systems Research: A Case Study*, *MIS Quarterly*, Vol. 12, Issue 4.
- Dix, A., Finlay, J., Abowd, G. & Beale, R. (1998). *Human-Computer Interaction*. Prentice Hall Europe, Second Edition, 1998.
- Doubleday, A., Ryan, M., Springett, M. & Sutcliffe, A. (1997). A comparison of usability techniques for evaluating design. *Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques*, ACM Press, New York, NY, USA.
- Dumas, J. S. & Redish, J. C. (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex Publishing.
- Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework*. Summary Report, Stanford Research Institute, on Contract AF 49(638)-1024, October 1962.

- Galliers, R. D. & Land, F. F: (1987). Choosing Appropriate Information Systems Research Methodologies. *Communications of the ACM*, Volume 30, Number 11, pp. 900-902, November, 1987.
- Galliers, R. D. (1992). Choosing Information Systems Research Approaches. In Galliers R. D. (Ed.) 1992, *Information Systems Research: Issues, Methods and Practical Guidelines*, Blackwells Scientific Publications, pp. 144-162, Boston, MA.
- Gray, W. D. & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, 13(3), 203-261.
- Hartson, H. R., Andre, T. S. & Williges, R. C. (2001). Criteria For Evaluating Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4), 373-410, Lawrence Erlbaum Associates, Inc.
- ISO (1998). The international Organization for Standardization, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. (ISO 9241-11).
- Karat, C.-M. (1994). A business case approach to usability cost justification. In *Cost-justifying usability*, Academic Press, Inc., Orlando, FL, USA.
- Kjeldskov, J. & Skov, M. B. (2003). Evaluating the Usability of a Mobile Collaborative System: Exploring Two Different Laboratory Approaches. In *Proceedings of the 4th International Symposium on Collaborative Technologies and Systems 2003*, Orlando, Florida, SCS press.
- Kjeldskov, J. & Stage, J. (2004). New Techniques for Usability Evaluation of Mobile Systems. Accepted for publications in *International Journal of Human-Computer Studies*, Elsevier (forthcoming 2004).
- Kjeldskov, J. (2003). *Human-Computer Interaction Design for Emerging Technologies: Virtual Reality, Augmented Reality and Mobile Computer Systems*. Ph.D. Thesis, Department of Computer Science, Aalborg University.
- Lewis, C., Polson, P., Wharton, C. & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. *Proceedings of CHI 90*, 235-242. New York, NY: ACM.
- Markus, M. L. & Lee, A. S. (1999). Special Issue on Intensive Research in Information Systems: Using Qualitative Interpretive, and Case Methods to Study Information Technology – Foreword. *MIS Quarterly*, Vol. 23, No. 1, pp. 35-38, March, 1999.
- Mayhew, D. J. (1999). *The Usability Engineering Lifecycle – A Practitioner’s Handbook for User Interface Design*. Morgan Kaufman Publishers, Inc., 1999.

- Molich, R. (2000). *Brugervenlige edb-systemer*. Teknisk Forlag.
- Myers, M. D. (1997). Qualitative Research in Information Systems, *MIS Quarterly*, Vol. 21, Issue 2, pp. 241-242. MISQ Discovery, archival version, June 1997, <http://www.misq.org/misqd961/isworld/>, updated version, May 12, 2004 available at: <http://www.auckland.ac.nz/msis/isworld/>.
- Nielsen, C. (1998). Testing in the Field. In Werner, B. (Ed.), *Proceedings of the third Asia Pacific Computer Human Interaction Conference*, pp. 285-290. IEEE Computer Society.
- Nielsen, C. M., Overgaard, M., Pedersen, M. B. & Stenild, S. (2004). *The Development of a Mobile System for Communicating and Collaborating – An Object-Oriented HCI Approach*. Department of Computer Science, Aalborg University, 2004.
- Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of CHI 90*, 249-256. New York, NY: ACM.
- Nielsen, J. (1993). *Usability Engineering*. Academic Press, Boston.
- Norman, D. A. (2002). *The Design of Everyday Things*. Basic Books (Perseus), New York, 2002.
- Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York, NY: John Wiley & Sons.
- Straub, D., Gefen D. & Boudreau, M. (2004). Qualitative Research in Information Systems. In: *PhD Supervisors and Students Handbook for Information Systems Research, IFIP TC8 Supervisors Workshop, May 2004 (Forthcoming)*.
- Wharton, C., Bradford, J., Jeffries, R. & Franzke, M. (1992). Applying cognitive walk-throughs to more complex user interfaces: experiences, issues, and recommendations. *Proceedings of the SIGCHI conference on Human factors in computing systems, 1992*.
- Wynekoop, J. L. & Conger, S. A. (1990). A review of computer aided software engineering research methods, in *Information systems research: Contemporary approaches and emergent traditions*. In Nissen H-E., H. K. Klein and R. Hirschheim (eds.), *Information systems research*; p. 301-325, Elsevier Science Publishers B.V, 1991.

APPENDIX A

RESEARCH PAPER 1

A REVIEW OF LITERATURE ON USABILITY EVALUATION METHODS FOR MOBILE SYSTEMS

RESEARCH PAPER 2

USABILITY EVALUATION OF A MOBILE SYSTEM: COMPARISON OF A LABORATORY AND A FIELD EVALUATION

RESEARCH PAPER 3

PROVIDING FEEDBACK TO DESIGNERS: ARE USABILITY REPORTS ANY GOOD?

A Review of Literature on Usability Evaluation Methods for Mobile Systems

Christian Monrad Nielsen, Michael Overgaard, Michael Bach Pedersen & Sigge Stenild
Department of Computer Science, Aalborg University, Denmark
{monrad, mio, mbp, stardust}@cs.auc.dk

Abstract

In this paper we explore and examine 1826 papers from key conference proceedings and journals for approaches on how to perform user-based usability evaluation of mobile systems in laboratory and field settings respectively. During the study, we identified 58 papers, which covered a usability evaluation of a mobile system. These papers were then categorized in relation to relevant issues and activities when conducting user-based usability evaluations of mobile systems in either laboratory- or field-based settings. The result of the review showed that despite the challenges of performing field-based evaluations, it was the most frequently applied evaluation approach. The review also revealed that interviews were more often used for data collection in the field evaluations compared to laboratory evaluations. Furthermore, video recordings were used for different purposes between the two types of evaluations, because field conditions complicate the use of video. Finally, statistical quantitative methods, such as ANOVA, were the most prevalent applied methods for data analysis.

1. Introduction

The study of usability evaluation methods is a well-established research area, but researchers are still faced with several difficult issues when designing their experiments, such as the degree of realism and how to collect data. This is even more prominent when conducting evaluations of mobile devices, since they are often used in different and highly dynamic contexts [Kjeldskov & Stage, 2004]. This particular research area is relatively young and therefore has little knowledge on a methodological level [Kjeldskov & Graham, 2003]. Conducting a review of literature within this research area can provide insight into tendencies and approaches utilized in other studies and development projects.

Wynekoop & Conger [1990] classifies research approaches in computer-aided software engineering (CASE) by reviewing and classifying research papers to provide an overview of current research methods and purposes within the information systems area. Using the same overall approach, Kjeldskov & Graham [2003] reviews what research methods are dominant

within the area of mobile HCI. Their study reveals a bias towards engineering systems using applied approaches, and those performing evaluation do so in a laboratory setting.

Mohamedally et al. [2003] presents a review, based on three years of publications from the Mobile HCI conferences, to highlight the areas, where most research have been focussed, but also to draw attention to areas lacking research. They discover that empirical user-based studies and ethnographical analysis in user needs requirements are strongly promoted in the mobile HCI community, while expert evaluations were less popular. Their study also shows that context- and location-aware systems are becoming more prominent research areas. Hansen et al. [2002:21-34] expands the same selection of papers with two years and explores how user needs and requirements are identified in the development of mobile systems. They discover that papers conducting usability evaluations span all over the conference's five years, but less than a quarter of these papers describes how they perform the evaluation. A majority of the papers apply classic evaluation methods, both qualitative and quantitative.

All of the above papers conduct literature studies in order to explore and examine trends and tendencies concerning either research methods or usability evaluation methods (UEMs). In this paper we apply a similar approach in our review of literature, by studying and categorizing several papers to obtain an overview of user-based usability evaluation methods presented within the HCI research area. Although our approach is similar, the amount and variety of papers are considerably larger. A total of 1826 papers, from key journals and conference proceedings within the HCI research area, are explored in this paper. The results of the study can be used as an inspiration to produce ideas on how both laboratory- and field-based evaluations of mobile systems can be conducted.

The purpose of this paper is to explore, categorize, and present, which methods and techniques for user-based usability evaluation of mobile systems are presented in key HCI literature.

More specifically, this paper will focus on the evaluation of mobile systems and how evaluations are conducted in either laboratory or field settings. This includes examining how data are collected during these

evaluations, because several difficulties in collecting data are introduced in field-based evaluations [Kjeldskov & Stage, 2004]. Afterwards, we examine how the data is analyzed, in order to discover tendencies in data analysis approaches in both qualitative and quantitative data collection methods.

In section 2 we describe what constitutes a usability evaluation method and what characterizes different approaches to usability evaluation, whereas section 3 describes specific concerns when evaluating mobile systems. In section 4 we will go through the process in the literature review and describe how we categorized the relevant papers. Section 5 and 6 presents and discusses the results, which provides an indication of the most commonly used approaches and aspects of user-based usability evaluation of mobile systems. Finally, section 8 provides the conclusion.

2. Usability Evaluation Methods

Before searching through the HCI literature, we will describe what characterizes a UEM, in order to clarify what we will be looking for in our review of literature.

The International Organization for Standardization (ISO) defines usability as the “extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [ISO, 1998]. Furthermore, the key terms in this definition are described as [ISO, 1998]:

- *Effectiveness*: Accuracy and completeness with which users achieve specified goals.
- *Efficiency*: Resources expended in relation to the accuracy and completeness with which users achieve goals.
- *Satisfaction*: Freedom from discomfort, and positive attitudes towards the use of the product.
- *Context of use*: Users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used.

The activity of evaluating the usability is any activity, either analytical or empirical, aimed at assessing or understanding the usability of an interactive system or prototype [Rosson & Carroll, 2002]. Several methods for evaluating the usability have been created and presented. These UEMs are often based on common sense, case studies, lessons learned and collected from various organizations, or on the results of experimental studies designed to compare two or more UEMs [Gray & Salzman, 1998]. UEMs are used to evaluate the interaction of the human with the computer for the purpose of identifying aspects of this interaction that can be

improved to increase usability [Gray & Salzman, 1998].

UEMs can be divided into two different categories, namely *formative evaluation* and *summative evaluation* [Hartson et al., 2003]. Formative evaluation is performed at different stages during the development of a system and the purpose is to find usability problems so that they can be fixed and the usability of the system can be improved. Formative evaluation often focuses on qualitative data [Hartson et al., 2003] and if used in combination with the quantitative data, it can add “flavour to the formative process because it is used to assess the level of usability” [Hartson et al., 2000]. On the other hand, summative evaluation focus on assessing the level of usability achieved in a system for it to be compared to other systems and/or usability metrics, often through quantitative data [Travis, 2003].

The focus of this paper is formative empirical UEMs, which all involve users when evaluating the usability of a system. Based on this, we exclude evaluation methods based on experts like heuristic evaluation [Nielsen, 1993][Molich, 2000] and cognitive walk-through [Wharton et al., 1994] [Dix et al., 1998]. These can also be described as analytical methods [Gray & Salzman, 1998] [Hix & Hartson, 1993]. Empirical methods, on the other hand, focus on examining the system while it is used by the target users [Gray & Salzman, 1998] [Hix & Hartson, 1993], and these are the methods of interest in this paper. Therefore, formative empirical UEMs, which include user participation, are the approaches we are looking for in the literature review. In the rest of this paper they will be referred to as user-based usability evaluation methods.

3. Evaluation of Mobile Systems

A mobile system is characterised by being used while moving, in different locations and situations (changing contexts), and they often have very small visual displays [Kakihara & Sørensen, 2002] [Dey & Abowd, 2001]. Considering usability evaluation of mobile systems, the location needs to be taken into account [Kjeldskov & Stage, 2004], since the use of mobile systems is closely connected to three factors: *Environment*, *application*, and *modalities*, which affect how usability evaluation of mobile systems is performed [Kristoffersen & Ljungberg, 1999].

The requirements of mobile applications will be very different from the stationary setting that until recently has been the dominating one in the area of HCI [Holmquist et al., 2002]. A number of challenges and opportunities for the design and evaluation of mobile applications have been identified, such as context sensitivity and ergonomics, which shows that there is a need for a general and coherent design and evaluation approach

that addresses properties that are unique for mobile applications [Holmquist et al., 2002].

Conducting evaluations in laboratory settings provides the opportunity of experimental control and collection of high quality data [Kjeldskov & Stage, 2004], and several authors have published extensive guidelines on how tests in laboratory settings should be conducted [Dumas & Redish, 1993] [Nielsen, 1993] [Rubin, 1994]. Evaluating the usability of mobile systems constitutes a potential challenge, since the use of such systems is typically closely related to activities in their physical surroundings and often requires a high level of domain-specific knowledge [Nielsen, 1998]. If this is the case, usability evaluation of mobile systems can potentially benefit from performing the evaluation in the field, but it is not an easy task [Nielsen, 1998] [Brewster, 2002]. The difficulties of field evaluation can be summed up as [Beck et al., 2003]:

- It is complicated to establish realistic studies.
- It is non-trivial applying evaluation techniques like observation and think-aloud in the field.
- Data collection is complicated and the control is limited when users are physically moving around.

One possible solution to the problems faced in field testing is using a dedicated simulator [Kjeldskov & Stage, 2003b] or try to and recreate a part of the field context in the laboratory, but it can be difficult to recreate realistic physical settings in the laboratory [Kjeldskov & Stage, 2004].

4. Method

The papers we have studied originate from the following conferences and journals from the last five years¹:

- International Conference on Human-Computer Interaction (INTERACT).
- Symposium on Human-Computer Interaction with Mobile Devices (Mobile HCI).
- Conference on Computer-Human Interaction (CHI).
- Conference on Computer-Supported Cooperative Work (CSCW).
- Behaviour and Information Technology (BIT).
- International Journal of Human-Computer Interaction (IJHCI).

- International Journal of Human-Computer Studies (IJHCS).
- Personal and Ubiquitous Computing (PUC).
- Transactions on Computer-Human Interaction (TOCHI).

The complete listing and distribution of papers across the respective years, journals, and conferences are shown in Table 1:

Conferences	1999	2000	2001	2002	2003	Total
INTERACT	145		152		170	467
Mobile HCI	17		23	50	52	142
CHI	78	72	69	61	75	355
CSCW		52		39		91
Journals	1999	2000	2001	2002	2003	Total
BIT	38	40	44	40	34	196
IJHCI	11	26	25	27	32	121
IJHCS		45	75	44	77	241
PUC		12	41	49	39	141
TOCHI	13	19	12	14	14	72
Total	302	266	441	324	493	1826

Table 1: The number and distribution of the explored papers.

Table 1 shows that a total of 1826 papers were explored. The reason for the blank cells in the table are that INTERACT and CSCW are biannual conferences and Mobile HCI were held in conjunction with INTERACT in 1999 and 2001. Among the journals, papers from IJHCS and PUC² in 1999 were not available. The numbers enclosed in brackets in the remaining parts of this paper refers to the papers studied in the review, which can be found in the last part of the reference list.

The process of selecting the papers, relevant to our study, consisted of the following three steps:

1. Reading abstracts and pre-selecting candidates
2. Producing summaries
3. Narrowing down the candidates

4.1. Reading Abstracts

In this step, we distributed all of the 1826 papers among the four authors. The abstracts of each paper were read in order to determine if any papers were relevant according to the following criteria. The papers had to:

¹ Complete listing of all the papers can be found at: <http://www.cs.auc.dk/~miov/inf8/literature.html>

² The journal "Personal and Ubiquitous Computing" was previously to 2000 published under the name "Personal Technologies".

1. Involve a mobile system and a user-based usability evaluation.
2. Describe the evaluation approach chosen.

The first criterion is our *subject* of interest and originates from the fact that mobile systems are inherently different from desktop systems. We have also stated that we focus on usability evaluations involving users, hence the criterion of user-based evaluation. In order for us to understand and use any of the approaches described in the papers or parts of them, we required some level of detail in the description of their evaluation approach. If the papers provided little information on the *method* applied, it might be difficult to comprehend their approach.

Since abstracts sometimes were not entirely representative for the whole content of a paper, it can be argued that our approach in the first round of the selection process was not “thorough enough”. But reading 1826 papers in their entirety were not possible to do in just a few weeks, and it might not provide a considerably better result. Furthermore, because of the large amount of papers considered in the first round, only one person read the abstracts of each paper.

In order to ensure that the papers indeed covered these aspects, we chose *all* papers that mentioned *anything* on usability evaluation of mobile systems in either the title or the abstract. Usability evaluation are often referred to using words like “test”, “trial”, and “evaluation”, hence we focused on all of these words. If reading the abstract did not provide a clear indication of its relevance, it was skimmed through to reveal whether the content was relevant. This left us with 99 papers distributed as shown in Table 2:

Conferences	1999	2000	2001	2002	2003	Total
INTERACT	5		10		12	27
Mobile HCI	3		3	9	12	27
CHI	0	2	1	3	6	12
CSCW		1		3		4
Journals	1999	2000	2001	2002	2003	Total
BIT	1	1	1	1	0	4
IJHCI	0	0	1	1	0	2
IJHCS		1	3	3	3	10
PUC		0	0	7	2	9
TOCHI	0	2	1	1	0	4
Total	9	7	20	28	35	99

Table 2: The remaining papers after the first round of selection.

4.2. Producing Summaries

During this step, the 99 papers were split up in four parts and distributed among the authors, which then read the papers in their entirety. After reading each paper, a focused summary of the paper was produced,

focusing on issues relevant to the two criteria. To provide an insight into our approach, an example of a summary is presented below, based on Ward & Tsukahara (2003) [56]:

“The paper is about developing a tool for taking class-notes during a lecture emphasising on improving the note-taking process. They describe different input devices and their advantages and consider different design principles and choices. Following this, they conducted a user study in order to discover whether their system (NoteTaker) is better than paper and pencil for taking notes. Four users familiarized themselves with the system before the evaluation and they were asked whether they would like to continue using the system. In addition to this, they performed a semi-controlled laboratory test, video-taping four different lectures, where the students used the system. Finally they present the results of the evaluation.”

The purpose of writing these summaries for each paper was not to use them as the basis for selection, since they did not cover the entire scope of the paper. They were produced in order to create an overview of the 99 papers and as a method for quickly gaining insight into what the paper described in relation to the two criteria.

4.3. Narrowing Down the Candidates

Afterwards, two persons reviewed all of the summaries and papers in cooperation. Through this process, we agreed that 39 did not cover the predefined criteria for selection, which indicate that we were not excessively selective in the first round. After removing the 39 papers we ended up with 60 papers that fulfilled the criteria and their distribution are shown in Table 3:

Conferences	1999	2000	2001	2002	2003	Total
INTERACT	1		6		4	11
Mobile HCI	2		2	7	10	21
CHI	0	2	1	3	6	12
CSCW		1		1		2
Journals	1999	2000	2001	2002	2003	Total
BIT	0	0	1	1	0	2
IJHCI	0	0	1	0	0	1
IJHCS		0	0	0	2	2
PUC		0	0	5	2	7
TOCHI	0	1	0	1	0	2
Total	3	4	11	18	24	60

Table 3: The remaining papers after the second round of selection.

As the next step, we looked for redundant papers, which were any papers that were essentially the same, but had been published in different outlets. We found two pairs of papers, which covered the same two projects, and this left us with a total of 58 unique papers. The oldest from the two pairs of papers were disre-

garded in relation to our findings. The reason for this was that basically identical papers should only count once in our findings, since they did not provide any new information. The papers disregarded were: Chittaro & Dal Cin (2001) [18] and Kaikkonen & Roto (2002) [31], since they are represented by: Chittaro & Dal Cin (2002) [19] and Kaikkonen & Roto (2003) [32]. After discarding these papers, we end up with the number of papers presented in Table 4:

Conferences	1999	2000	2001	2002	2003	Total
INTERACT	1		6		4	11
Mobile HCI	2		1	6	10	19
CHI	0	2	1	3	6	12
CSCW		1		1		2
Journals	1999	2000	2001	2002	2003	
BIT	0	0	1	1	0	2
IJHCI	0	0	1	0	0	1
IJHCS		0	0	0	2	2
PUC		0	0	5	2	7
TOCHI	0	1	0	1	0	2
Total	3	4	11	18	24	58

Table 4: The final selection of papers after redundant papers were disregarded.

4.4. Categories

In order to obtain an overview of the distinct characteristics of the different papers, a number of issues relevant to a usability evaluation process were determined. The results from the categorization of the papers are shown in the tables in Appendix A. In this section, the different categories, their purpose, and their justification will be described.

Device: The type of device(s) that is evaluated, e.g. mobile telephones, PDA's, Tablet PC's and hybrids. Numerous examples of papers, describing the *device* being evaluated, exist, since the device determines what kind of input and output modalities that is available and the ergonomics of the device in general [Dix, 1998:110].

Technology: Any auxiliary technology in the device used by the system being evaluated. *Technology* can provide both possibilities and limitations in relation to the design of a system [Kaikkonen & Roto, 2003].

System: The overall functionality of the system being evaluated. A system may operate on different devices, e.g. PDAs, laptops or systems in cars. The *system* description provides information on the intention, functionality and the application domain of the system [Mathiassen et al., 2000].

User Description: Indicates whether a description of the evaluation participants is presented in the paper or not. Selecting and acquiring the appropriate test participants are a crucial part of testing process. Test re-

sults can only be considered valid if the participants are typical end-users [Nielsen, 1993:175-179] [Rubin, 1994:119-139] [Molich, 2000:104-105].

Number of Participants: Describes the number of participants used in the respective evaluations. A reasonable *number of participants* are required in order to produce statistically valid results and limit biasing conditions [Nielsen, 1993:173-174] [Rubin, 1994:128] [Molich, 2000:104-105].

Laboratory and Field: Covers the range of information, which relate to the physical setup and the location of the evaluation. The location and context influence the evaluation, and benefits and drawbacks have been described for both *laboratory*- and *field*-based evaluations [Nielsen, 1998] [Pascoe et al., 2000] [Als et al., 2003] [Kjeldskov et al. 2004] [Kjeldskov & Stage, 2004].

Data Collection: The method of data collection before, during and after the evaluation, e.g. logs, interviews, questionnaires, or observations. Both qualitative and quantitative approaches for *data collection* exist [Nielsen, 1993:175-179] [Rubin, 1994: 156-169] [Molich, 2000:108-109], but some of them are not easily applied in field-based evaluations [Kjeldskov & Stage, 2004] [Kjeldskov et al., 2004].

Tasks: Indicates whether specific assignments were utilized during the evaluation. The evaluation might require the participants to perform specified *tasks*, which should be both realistic and relate to the application domain [Rubin, 1994:179-184] [Molich, 2000:102-104]. Performing general usage of the system is not considered as performing specific tasks.

Data Analysis: Indicates how the data were analyzed. *Analysis of data* should be conducted in a systematic and scientific reproducible fashion [Rubin, 1994:257-283] [Molich, 2000:108-110].

Results: Describes whether the paper presents the results of the usability evaluation and data analysis. The presentation of the *results* is important in order to determine the applicability and usefulness of the evaluation as a whole. Different ways to communicate the test results have been described [Rubin, 1994: 283-293] [Molich, 2000:110-112].

Comparison: Indicates whether a paper describes comparisons of setups, experiences, results or similar aspects between evaluations in field and/or laboratory settings. *Comparisons* can be used to improve usability evaluation methods, although they should be done with care in order to produce statistically valid results [Gray & Salzman, 1998] [Hartson et al., 2001].

4.5. Categorization of Papers

In order to categorize the papers, each paper was read by two persons, who performed a categorization indi-

vidually and afterwards they compared their findings for each paper. If the two persons' findings were not consistent concerning a specific category, we noted that a difference had occurred. A difference occurred if the two persons had noted different things or issues about one of the categories covered in a paper. The number of occurred differences for each category is shown in Figure 1 and is based on the 58 papers.

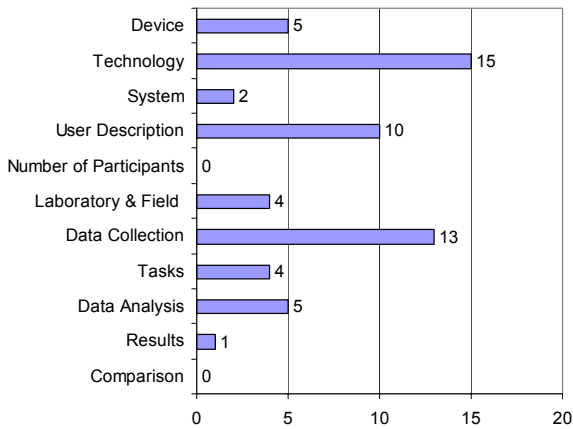


Figure 1: The number of differences in relation to the categories.

If a difference had occurred, the paper was then examined once again, and the content was discussed until consensus was reached about the specific categorization, which was done in order to ensure consistency. As it can be seen in Figure 1, technology, user description, and data collection had the most differences, which might be due to the following reasons:

- It is difficult to determine the primary technology used by a system, since the descriptions about the technology are often very short.
- It is difficult to determine how much information is needed in order to decide, when it is a proper user description.
- Some data collection methods might cover similar approaches, but have dissimilar names and descriptions.

4.6. Categorization Table

The results of the categorization of the 58 papers are shown in the four tables in Appendix A. Table 7 shows the papers from the first two conference proceedings and the first seven categories, while Table 8 shows the last five categories according to these papers. Table 9 shows the papers from the remaining conference proceedings and journals along with the first seven categories, while Table 10 shows the last five categories ac-

ording to the remaining papers. Some of the cells in all four tables are left blank, which means that the particular article did not provide any information within the respective category. The relationship between the tables mentioned above is shown in Figure 2.

Table 7	Table 8
Table 9	Table 10

Figure 2: The relationship between the different tables in the comparison table.

5. Results

This section presents the results from our literature study described above. All of the results are based on the categorization table in Appendix A (Table 7 - 10). The study showed that out of a total 1826 papers, we identified 58 papers, which presented a user-based usability evaluation of a mobile system and also described the evaluation procedure.

The study revealed that 55% (32 of 58 papers) evaluate systems running on PDA's while 21% (12 of 58 papers) evaluated systems on mobile phones. This makes PDA's and mobile phones the most commonly used devices in the presented usability evaluations.

Considering the technology used by the system, 22% (13 of 58 papers) used Wireless LAN, 10% (6 of 58 papers) used WAP technology, and 9% (5 of 58 papers) used GPS coordinates as the primary technology in their system. This makes Wireless LAN the most used technology and in 10 out of the 13 papers (77%), Wireless LAN is used together with a PDA.

When considering the type of system evaluated in the papers, guide and navigation systems were the most represented systems evaluated with 26% (15 of 58 papers), and in 93% (14 of 15 papers) of the cases, these systems were evaluated in the field. A distant number two were instant messaging systems with only 5% (3 of 58 papers). This indicates a clear bias towards evaluating traditional mobile devices running typical mobile systems. Apart from these traditional mobile systems, we found some papers, who explored novel systems and new domains for applying mobile systems, such as:

- Paramedic Information System [3]
- Programmable Building Blocks [58].
- Nonverbal Calls on Mobile Phones [42].
- Counting giraffes in Africa [47].

We also examined, whether the papers had a detailed description of the participants, and we noted the num-

ber of participants in each evaluation. 86% (50 of 58 papers) provided more or less detailed descriptions about the participants, but in several cases, it was difficult to interpret from these descriptions, whether it was the future users of the system, who participated in the evaluation, or anybody else available. The number of participants in laboratory evaluations varied from 6 to 48 participants, and in field evaluations they varied from 1 to 60 participants. 10% (6 out of 58) of the papers conducted long-term evaluations, and four of them explicitly wrote for how long time the evaluations lasted. They ranged from 2 weeks [29] to 2 months [47] and they were all performed in the field. None of them supplemented with short term laboratory or field evaluations.

The type of evaluation is divided into two subcategories; laboratory evaluation and field evaluation. A two-tailed, large sample test for population proportion showed a very significant difference between performing laboratory evaluations only and field evaluations only ($z=3.01$, $p=0.003$) with field evaluations as the most commonly applied evaluation method. Only 9% (5 of 58 papers) conducted both a laboratory and field evaluation. The distribution of the articles is shown in Table 5. The reason why the enumeration of papers goes up to 60 in Table 5 is because the two redundant articles, [18] and [31], found, as described in section 4.3, is not included in this table, so the total number of papers is still 58.

	Laboratory Evaluation Only	Field Evaluation Only	Both Laboratory and Field Evaluation
Articles	10, 13, 17, 22, 23, 24, 26, 32, 33, 35, 36, 37, 39, 42, 44, 45, 50, 52, 54, 57, 59	1, 2, 3, 4, 6, 7, 8, 9, 12, 14, 15, 16, 19, 20, 21, 25, 27, 28, 29, 34, 38, 40, 41, 43, 46, 47, 48, 51, 53, 55, 58, 60	5, 11, 30, 49, 56
Total	21	32	5
% of 58	36%	55%	9%

Table 5: The papers distributed on the type of evaluation.

In the papers conducting laboratory evaluations only, there is a clear bias towards performing traditional laboratory tests. Conducting a large sample test for the difference between two population proportions shows a very significant difference between papers using participants, who were sitting at a table while evaluating the system, and the papers, who tried to create field-like settings in the laboratory ($z=6.07$, $p<0.001$). The field-like experience within the laboratory was established in various ways, e.g. through the use of a stairmaster [49] or having the participants use a driving simulator [26] [36]. In 81% (26 of 32 papers) of the field-based evaluations, they explicitly described how they have established realistic settings during the test,

e.g. downtown Stockholm in a car [14], the participants workplace [34] and in a retail environment [43]. These are also examples of very realistic settings, while others simulate such settings, e.g. using a mock-up of a shopping mall while performing role-play [7].

Data collection methods were primarily based on traditional approaches, such as interview by 53% (31 of 58 papers), questionnaire by 43% (25 of 58 papers), observation by 47% (27 of 58 papers), video recording by 34% (20 of 58 papers), and different kinds of logs by 29% (17 of 58 papers). Considering data collection approaches in a laboratory evaluation compared to field evaluation and those who performed both types of evaluations, we got the results presented in Figure 3:

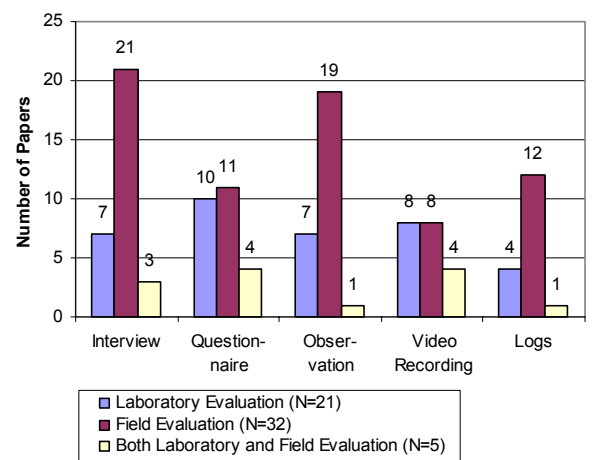


Figure 3: The number of papers by type of evaluation relative to the data collection methods.

Based on Figure 3, the distribution of data collection methods, within each of the evaluation approaches, is illustrated in Table 6:

	Laboratory Evaluation Only	Field Evaluation Only	Both Laboratory and Field Evaluation
Interview	33%	66%	60%
Questionnaire	48%	34%	80%
Observation	33%	59%	20%
Video Recording	38%	25%	80%
Logs	19%	38%	20%

Table 6: The distribution of data collection methods within each of the different evaluation approaches.

A two-tailed large sample test for the difference between two population proportions showed a significant difference between using interview in laboratory and field evaluations ($z=2.30$, $p=0.02$). Figure 3 and Table 6 give the impression that there is a considerable difference between the two types of evaluation methods concerning observation, however the difference is only

marginally significant ($z=1.85$, $p=0.06$). Although no significant difference was found, concerning the use of video recordings between laboratory and field evaluations ($z=1.01$, $p=0.31$), video recordings had different purposes in the two types of evaluations. In the laboratory evaluations, video were used to record a close up view of the screen and the user's interaction with the device [13] [33] [44] [45], as well as the user's head and eye movements [36]. In the field evaluations, video were used to capture the user's movement [5] [49], behavior [12], and interaction with other individuals when using the system [12] [25] [51]. The reason for the different focus in field evaluation were due to the complicated conditions for video recording the device's screen, as well as the user's interaction with the device, in these settings [5] [49]. In order to overcome some of the difficulties of video recording the user's interaction with devices with small screens, one article reported using a mini-camera attached to the mobile phone [32]. The camera captured only the phone's screen and button area. The test monitor was then able to observe the user's interaction from the LCD screen on a video camera [Nyyssönen et al., 2004]. No significant difference existed between laboratory and field evaluations concerning the use of questionnaires ($z=0.96$, $p=0.34$) and logs ($z=1.43$, $p=0.15$). Logs were often used as a supplement to other data collection methods, mainly qualitative methods, and we did not find any papers, where logs were the only method for data collection. One paper [16] used only audio recordings during a think-aloud evaluation, and the interface of the device was observed by shadowing the participants. This approach was most likely possible, since the system was running on a Tablet PC, which has a relatively large screen.

Having the participants perform specific tasks during the evaluation was utilized in 69% (40 of 58 papers) of the papers, although the level of detail in the descriptions of the different tasks varied a great deal. A large part of the remaining 31%, who did not utilize specific tasks during the evaluation, had the participants perform some kind of general usage of the evaluated system. In the papers describing long-term evaluations, only 17% (1 of 6 papers) described the tasks performed by the participants during this evaluation [46].

When analyzing the data collected in the usability evaluation, 32% (19 out of 58 papers) conducted various types of statistical analysis, and of these ANOVA was the most commonly used by 58% (11 out of 19 papers). Additionally, 22% (13 out of 58 papers) conducted some kind of time measurement analysis, such as task completion time. Qualitative approaches, such as conversation analysis, interaction analysis, transcriptions, and socio technical approaches, were utilized in 16% (9 of 58 papers) of the papers.

7% (4 of 58 papers) of the papers conducted evaluations in both laboratory and field settings and afterwards compared the results of these tests.

6. Discussion

The purpose of the literature survey is to obtain an overview of the most common approaches when conducting user-based usability evaluations of mobile systems, and arranging the papers in different categories according to their content seemed both appropriate and relevant.

The categories *device*, *technology* and *system* are used to determine the type of mobile system evaluated in the papers. *User description*, *number of participants*, *laboratory* and *field* categories are relevant when wanting to achieve information about the framework of the user-based usability evaluations. The remaining categories; *data collection*, *tasks*, *data analysis*, *results* and *comparison* are relevant in order to obtain an overview of the papers' evaluation approaches, which is how the data are collected, analyzed, and presented afterwards. Although the categorization provided the desired overview of the papers and their content, it was sometimes difficult to describe the papers' approaches concerning all of the categories.

In section 3 we describe some of the difficulties of performing field evaluation. Despite the complexity of establishing realistic studies we have found several papers describing field evaluation, although they vary greatly in the degree to which they perform truly realistic field evaluations. Different types of systems, devices, and technology might also require different degrees of realism in order to produce satisfactory results. Some types of systems are simply impossible to evaluate in a 100% real environment, such as systems operating in safety-critical domains, e.g. air traffic control systems [Fields et al., 1999] and control and monitoring systems for large container vessels [Kjeldskov & Stage, 2003a].

All of the 58 papers found in our review present evaluations performed by the designers of the system themselves. This might not be the appropriate approach for identifying usability problems, since the designers are biased towards the system, and might not be able to evaluate the system as objectively as external evaluators would. The lack of independence between those who design and those who evaluate may pose a threat to objective evaluation [Bachrach & Newcomer, 2002]. If the designer and evaluator is the same person, they are very likely to enhance the reputation and prestige of the designer, as well as the credibility of the theories and methods they advocate [Bachrach & Newcomer, 2002]. On the other hand, this independency can have some disadvantages, such as the evaluator being unfamiliar with the application domain and having inade-

quate knowledge about conventions, functionality and design issues in the system [Hartson et al., 2004].

7. Conclusion

In this paper we have examined usability evaluation methods within the field of HCI by reviewing 1826 papers. We have identified 58 papers that described user-based usability evaluation of mobile systems, and we discovered trends and tendencies in several aspects of the evaluation procedure.

The study showed a significant bias towards performing the evaluation in field settings despite the challenges of both performing and collection data during a usability evaluation in field-based settings. PDA's and mobile phones were the most often applied devices, and navigation and guide systems were the most common systems in the review. There was a clear bias towards evaluating these systems in field settings, which might be due to the functionality of such geographical location-dependent systems, are difficult to test properly in demarcated laboratory settings. Creating realistic settings was also used in the laboratory-based evaluations, since a minor part of those, who performed laboratory testing, tried to recreate a field-like experience within the laboratory in various ways.

Interview was more frequently used in field evaluations, which also applied for observation to some degree, because the difference between laboratory and field evaluations concerning the latter data collection method was only marginal. The study revealed no significant difference between the laboratory and field evaluations concerning the use of video recordings, questionnaires and logs for data collection, but a closer inspection revealed that the use video recordings had different purposes between laboratory and field evaluations, because of the complicated conditions for video recording in field settings. Furthermore, quantitative data analysis methods, especially ANOVA, are more often used than qualitative methods.

The presented review of user-based usability evaluation methods has some limitations. If field evaluations are conducted in a controlled environment, it is difficult to determine how much they actually differ from laboratory evaluations. This might make it inappropriate to have such a strict separation between the two categories, since they do not provide additional and useful information, such as degree of realism, about the evaluation other than the location of the test. However it is still interesting in relation to how data collection were performed, since it is one of the fundamental challenges of field-based usability evaluation.

Some research papers provide little information on some of the categories, which makes it difficult to describe their approach concerning all the categories. Perhaps the information, we require in relation to our

categories, were not relevant for the scientific focus and purpose of the specific paper, although the authors of these papers might have had important information on activities or techniques, if they were asked.

Different opinions of the reviewers might also have influenced the selection of papers in the first stages of the literature review. During the last rounds of selection, the influence of different opinions is diminished, since two persons read each paper in their entirety and afterwards discussed their findings.

The findings in this paper provide methodological insight into research tendencies in the area of user-based usability evaluation methods in relation to mobile systems. Further research would be needed to explore the different approaches, and examine how prominent they are in identifying usability problems.

References

Literature referenced in the paper:

- Als, B. S., Høegh, R. T., Kjeldskov, J., Skov, M. B. & Stage, J. (2003). *Comparing Usability Evaluations of Mobile System*, in Proceedings of the 3rd Danish Human-Computer Interaction Research Symposium. Roskilde University, Denmark.
- Bachrach, C. & Newcomer, S. F. (2002). *Addressing Bias in Intervention Research*, in Journal of Adolescent Health, Volume 31, Number 4, 2002.
- Beck, E., Christiansen, M., Kjeldskov, J., Kolbe, N. & Stage, J. (2003). *Experimental Evaluation of Techniques for Usability Testing of Mobile Systems in a Laboratory Setting*, in Proceedings of OzCHI 2003, Brisbane, Australia.
- Brewster, S. (2002). *Overcoming the Lack of Screen Space on Mobile Computers*, in Personal and Ubiquitous Computing, 6: 188-205.
- Dey, A. K. & Abowd, G. D., (2000). *Towards a Better Understanding of Context and Context-Awareness*, Conference on Human Factors in Computing Systems (CHI 2000), The Hague, The Netherlands, 2000.
- Dix, A., Finlay, J., Abowd, G. & Beale, R. (1998). *Human-Computer Interaction*, Prentice Hall Europe, Second Edition, 1998.
- Dumas, J. S. & Redish, J. C. (1993). *A Practical Guide to Usability Testing*, 1st edition, Greenwood Publishing Group Inc., Westport, CT, USA.
- Fields, R., Paterno, F., Santoro, C. & Tahmassebi, S. (1999). *Comparing Design Options for Allocating Communication Media in Cooperative Safety-Critical Contexts: A Method and a Case Study*, in ACM Transactions on Computer-Human Interaction, Vol. 6, No. 4, December 1999, Pages 370-398.
- Gray, W. D. & Salzman, M. C. (1998). *Damaged merchandise? A review of experiments that compare usability evaluation methods*, Human-Computer Interaction, 13(3), 203-261.
- Hansen, K. K., Høegh, R. T. & Lauritsen, S. (2002). *Fra behov til krav i en mobil kontekst*, Department of Computer Science, Aalborg University, 2002.

- Hartson, H. R., Shivakumar, P. & Pérez-Quiñones, M. A. (2004). *Usability Inspection of Digital Libraries: A Case Study*, accepted for publication in the Special Issue of Journal of Digital Libraries on Usability.
- Hartson, H. R., Andre, T. S. & Williges, R. C. (2001). *Criteria for evaluating usability evaluation methods*. International Journal of Human-Computer Interaction, 13, 4, 373-410.
- Hartson, H. R., Andre, T. S. & Williges, R. C. (2003). *Criteria For Evaluating Usability Evaluation Methods*, International Journal of Human-Computer Interaction, February 2003, Vol. 15, No. 1, Pages 145-181.
- Hix, D. & Hartson, H. R. (1993). *Formative Evaluation: Ensuring Usability in User Interfaces*, in L. Bass & P. Dewan (Eds.), Trends in Software, Volume 1: User Interface Software. New York: Wiley, 1-30.
- Holmquist, L. E., Höök, K., Juhlin, O. & Persson, P. (2002). *Challenges and Opportunities for the Design and Evaluation of Mobile Applications*, presented at the workshop Main issues in designing interactive mobile services, at Mobile HCI'02.
- ISO (1998). *The international Organization for Standardization, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability*, (ISO 9241-11).
- Kaikkonen, A. & Roto, V. (2003). *Navigating in a mobile XHTML application*, CHI '03, Ft. Lauderdale, Florida, USA, 2003.
- Kakahara, M. & Sørensen, C. (2002). *Mobility: An Extended Perspective*, in 35th Hawaii International Conference on System Sciences (HICSS-35), edited by Sprague Jr., R. IEEE, 2002.
- Kjeldskov, J. & Graham, C. (2003). *A Review of MobileHCI Research Methods*, in Proceedings of the 5th International Mobile HCI 2003 conference, Udine, Italy.
- Kjeldskov, J. & Stage, J. (2003a). *The Process of Developing a Mobile Device for Communication in a Safety-Critical Domain*, in Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction, Interact 2003. Zürich, Switzerland.
- Kjeldskov, J. & Stage, J. (2003b). *Designing the Handheld Maritime Communicator*, in Proceedings of the 1st Conference on Designing User Experiences, DUX 2003. San Francisco, CA, USA.
- Kjeldskov, J. & Stage, J. (2004). *New Techniques for Usability Evaluation of Mobile Systems*, accepted for publications in International Journal of Human-Computer Studies, Elsevier (forthcoming 2004).
- Kjeldskov, J., Skov, M. B., Als, B. S. & Høegh, R. T. (2004). *Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field*, accepted for publication in Proceedings of the 6th International Mobile HCI 2004 conference, Glasgow, Scotland.
- Kristoffersen, S. & Ljungberg, F. (1999). *Mobile Use of IT*, in Proceedings of The 19th Information Systems Research Seminar in Scandinavia, edited by T. Käkölä, Jyväskylä, Finland.
- Mathiassen, L., Munk-Madsen, A., Nielsen, P. A. & Stage, J. (2000). *Object Oriented Analysis & Design*, Marko Publishing ApS, Aalborg, Denmark, 2000.
- Mohamedally, D., Zaphiris, P. & Petrie, H. L. (2003). *Recent research in mobile computing: A review and taxonomy of HCI issues*, in Proceedings of HCI International 2003, Crete.
- Molich, R. (2000). *Brugervenlige edb-systemer*, Teknisk Forlag.
- Nielsen, C. M., Overgaard, M., Pedersen, M. B. & Stenild, S. (2004). *The Development of a Mobile System for Communicating and Collaborating – An Object-Oriented HCI Approach*, Department of Computer Science, Aalborg University, 2004.
- Nielsen, J. (1993). *Usability Engineering*, Academic Press, Boston.
- Nielsen, C. (1998). *Testing in the Field*, in Proceedings of the Third Asia Pacific Computer Human Interaction Conference (APCHI 98), Werner, B. (ed.), IEEE Computer Society, California, 1998, p. 285-290.
- Nyysönen, T., Roto, V. & Kaikkonen, A. (2004). *Mini-Camera for Usability Tests and Demonstration*, Nokia Research Center, Finland, 2004.
- Pascoe, J., Ryan, N. & Morse, D. (2000). *Using while moving: HCI issues in fieldwork environments*, ACM Transactions on Computer-Human Interaction, Volume 7, Issues 1-4, 2000.
- Rosson, M. B. & Carroll, J. M. (2002). *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*, Academic Press, 2002.
- Rubin, J. (1994). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, John Wiley & Sons, Inc., New York, NY, USA.
- Travis, D. (2003). *Standards Update: Usability Test Reporting*, found 07/05/04: <http://www.userfocus.co.uk/articles/cif.html>.
- Wharton, C., Rieman, J., Lewis, C. & Polson P. (1994). *The Cognitive Walkthrough: A practitioner's guide*, in Jakob Nielsen and Robert L. Mack, editors, Usability Inspection Methods. John Wiley and Sons, Inc. 1994.
- Wynekoop, J. L. & Conger, S. A. (1990). *A review of computer aided software engineering research methods*, in Information systems research: Contemporary approaches and emergent traditions, In Nissen H-E., H. K. Klein and R. Hirschheim (eds.), Information systems research; p. 301-325. 1991, Elsevier Science Publishers B.V.

Literature studied in the review:

1. Aittola, M., Ryhänen, T., Ojala, T. (2003). *SmartLibrary - Location-Aware Mobile Library Service*, Mobile HCI 03, Udine, Italy, 2003.
2. Åkesson, K. & Nilsson, A. (2002). *Designing Leisure Applications for the Mundane Car-Commute*, Personal and Ubiquitous Computing, Volume 6, Number 1-6, 2002.
3. Baber, C., Harris, T. & Harrison, B. (1999). *Demonstrating the Concept of Physical Hyperspace for an Art Gallery*, INTERACT '99, Edinburgh, Scotland, 1999.
4. Baber, C., Bristow, H., Cheng, S., Hedley, A., Kuriyama, Y., Lien, M., Pollard, J., & Sorrell, P. (2001).


- Augmenting Museums and Art Galleries*, INTERACT '01, Tokyo, Japan, 2001.
5. Baillie, L. (2003). *Future Telecommunications: Exploring Actual Use*, INTERACT '03, Zürich, Switzerland, 2003.
 6. Bjork, S., Falk, J., Hansson, R. & Ljungstrand, P. (2001). *Pirates! Using the Physical World as a Game Board*, INTERACT '01, Tokyo, Japan, 2001.
 7. Bohnenberger, T., Jameson, A., Krüger, A. & Butz, A. (2002). *Location-Aware Shopping Assistance: Evaluation of a Decision-Theoretic Approach*, Mobile HCI 02, Pisa, Italy, 2002.
 8. Bornträger, C., Cheverst, K., Davies, N., Dix, A., Friday, A. & Seitz, J. (2003). *Experiments with Multi-modal Interfaces in a Context-Aware City Guide*, Mobile HCI 03, Udine, Italy, 2003.
 9. Bosman, S., Groenendaal, B., Findlater, J. W., Visser, T., de Graaf, M. & Markopoulos, M. (2003). *GentleGuide: An Exploration of Haptic Output for Indoors Pedestrian Guidance*, Mobile HCI 03, Udine, Italy, 2003.
 10. Brewster, S., Lumsden, J., Bell, M., Hall, M. & Tasker, S. (2003). *Multimodal 'eyes-free' interaction techniques for wearable devices*, CHI '03, Ft. Lauderdale, Florida, USA, 2003.
 11. Brewster, S. (2002). *Overcoming the Lack of Screen Space on Mobile Computers*, Personal and Ubiquitous Computing, Volume 6, Number 1-6, 2002.
 12. Brown, B., MacColl, I., Chalmers, M., Galani, A., Randell, C. & Steed, A. (2003). *Lessons from the lighthouse: collaboration in a shared mixed reality system*, CHI '03, Ft. Lauderdale, Florida, USA, 2003.
 13. Bruijn, O. D., Spence, R. & Chong, M. Y. (2001). *RSVP Browser: Web Browsing on Small Screen Devices*, Mobile HCI 01, Lille, France, 2001.
 14. Brunnberg, L. & Juhlin, O. (2003). *Motion and Spatiality in a Gaming Situation - Enhancing Mobile Computer Games with the Highway Experience*, INTERACT '03, Zürich, Switzerland, 2003.
 15. Cheverst, K., Davies, N., Mitchell, K., Friday, A. & Efstathiou, C. (2000). *Developing a context-aware electronic tourist guide: some issues and experiences*, CHI '00, The Hague, The Netherlands, 2000.
 16. Cheverst, K., Mitchell, K. & Davies, N. (2002). *Exploring Context-aware Information Push*, Personal and Ubiquitous Computing, Volume 6, Number 1-6, 2002.
 17. Chittaro, L. & Camaggio, A. (2002). *Visualizing Bar Charts on WAP Phones*, Mobile HCI 02, Pisa, Italy, 2002.
 18. Chittaro, L. & Cin, P. D. (2001). *Evaluating Interface Design Choices on WAP Phones: Single-choice List Selection and Navigation among Cards*, Mobile HCI 01, Lille, France, 2001.
 19. Chittaro, L. & Cin, P. D. (2002). *Evaluating Interface Design Choices on WAP Phones: Navigation and Selection*, Personal and Ubiquitous Computing, Volume 6, Number 1-6, 2002.
 20. Esbjörnsson, M., Juhlin, O. & Östergren, M. (2003). *Motorcyclists Using Hocman - Field Trials on Mobile Interaction*, Mobile HCI 03, Udine, Italy, 2003.
 21. Fithian, R., Iachello, G., Moghazy, J., Pousman, Z. & Stasko, J. T. (2003). *The Design and Evaluation of a Mobile Location-Aware Handheld Event Planner*, Mobile HCI 03, Udine, Italy, 2003.
 22. Giller, V., Melcher, R., Schrammel, J., Sefelin, R. & Tscheligi, M. (2003). *Usability Evaluations for Multi-device Application Development Three Example Studies*, Mobile HCI 03, Udine, Italy, 2003.
 23. Goldstein, M., Nyberg, M. & Anneroth, M. (2003). *Providing proper affordances when transferring source metaphors from information appliances to a 3G mobile multipurpose handset*, Personal and Ubiquitous Computing, Volume 7, Number 1-6, 2003.
 24. Graham, R. & Carter, C. (1999). *Comparison of speech input and manual control of in-car devices while on the move*, Mobile HCI 99, Edinburgh, Scotland, 1999.
 25. Grinter, R. E., Aoki, P. M., Szymanski, M. H., Thornton, J. D., Woodruff, A. & Hurst, A. (2002). *Revisiting the visit: Understanding how technology can shape the museum visit*, CSCW '02, New Orleans, Louisiana, USA, 2002.
 26. Gupta, N., Bisantz, A. M. & Singh, T. (2002). *The effects of adverse condition warning system characteristics on driver performance: an investigation of alarm signal type and threshold level*, Behaviour & Information Technology, Volume 21, Number 1-6, 2002.
 27. Hibino, S. & Mockus, A. (2002). *handiMessenger: Awareness-Enhanced Universal Communication for Mobile Users*, Mobile HCI 02, Pisa, Italy, 2002.
 28. Isaacs, E., Walendowski, A. & Ranganathan, D. (2002). *Hubbub: a sound-enhanced mobile instant messenger that supports awareness and opportunistic interactions*, CHI '02, Minneapolis, Minnesota, USA, 2002.
 29. Jones, M., Jain, P., Buchanan, G. & Marsden, G. (2003). *Using a Mobile Device to Vary the Pace of Search*, Mobile HCI 03, Udine, Italy, 2003.
 30. Kaasinen, E. (2003). *User needs for location-aware mobile services*, Personal and Ubiquitous Computing, Volume 7, Number 1-6, 2003.
 31. Kaikkonen, A. & Roto, V. (2002). *XHTML in Mobile Application Development*, Mobile HCI 02, Pisa, Italy, 2002.
 32. Kaikkonen, A. & Roto, V. (2003). *Navigating in a mobile XHTML application*, CHI '03, Ft. Lauderdale, Florida, USA, 2003.
 33. Kjeldskov, J. & Stage, J. (2003). *The Process of Developing a Mobile Device for Communication in a Safety-critical Domain*, INTERACT '03, Zürich, Switzerland, 2003.
 34. Klemmer, S. R., Graham, J., Wolff, G. J. & Landay, J. A. (2003). *Books with voices: paper transcripts as a physical interface to oral histories*, CHI '03, Ft. Lauderdale, Florida, USA, 2003.
 35. Koppinen, A. (1999). *Design challenges of an In-Car Communication System UI*, Mobile HCI 99, Edinburgh, Scotland, 1999.
 36. Labiale, G. (2001). *Visual search and preferences concerning different types of guidance displays*, Behaviour & Information Technology, Volume 20, Number 1-6, 2001.

37. Lamberts, H. (2002). *Case Study: A PDA Example of User Centered Design*, Mobile HCI 02, Pisa, Italy, 2002.
38. Lehtikoinen, J. (2001). *An Evaluation of Augmented Reality Navigational Maps in Head-Worn Displays*, INTERACT '01, Tokyo, Japan, 2001.
39. Ma, X., Maglio, P. P. & Su, H. (2003). *Multimodal Menu Interface for Mobile Web Browsing*, INTERACT '03, Zürich, Switzerland, 2003.
40. McClard, A. & Somers, P. (2000). *Unleashed: Web tablet integration into the home*, CHI '00, The Hague, The Netherlands, 2000.
41. Milewski, A. E. & Smith, T. M. (2000). *Providing presence cues to telephone user*, CSCW '00, Philadelphia, Pennsylvania, United States, 2000.
42. Nelson, L., Bly, S. & Sokoler, T. (2001). *Quiet calls: talking silently on mobile phones*, CHI '01, Seattle, Washington, United States, 2001.
43. Newcomb, E., Pashley, T. & Stasko, J. (2003). *Mobile computing in the retail arena*, CHI '03, Ft. Lauderdale, Florida, USA, 2003.
44. Nyberg, M., Bjork, S., Goldstein, M. & Redstrom, J. (2001). *Handheld Applications Design: Merging Information Appliances without Affecting Usability*, INTERACT '01, Tokyo, Japan, 2001.
45. Öquist, G. & Goldstein, M. (2002). *Towards an Improved Readability on Mobile Devices: Evaluating Adaptive Rapid Serial Visual Presentation*, Mobile HCI 02, Pisa, Italy, 2002.
46. Palen, L. & Salzman, M. (2002). *Beyond the handset: designing for wireless communications usability*, ACM Transactions on Computer-Human Interaction, Volume 9, Issues 1-4, 2002.
47. Pascoe, J., Ryan, N. & Morse, D. (2000). *Using while moving: HCI issues in fieldwork environments*, ACM Transactions on Computer-Human Interaction, Volume 7, Issues 1-4, 2000.
48. Peltonen, J., Ollila, M. & Ojala, T. (2003). *TimeMachine Oulu -- Dynamic Creation of Cultural-Spatio-Temporal Models as a Mobile Service*, Mobile HCI 03, Udine, Italy, 2003.
49. Pirhonen, A., Brewster, S. & Holguin, C. (2002). *Gestural and audio metaphors as a means of control for mobile devices*, CHI '02, Minneapolis, Minnesota, USA, 2002.
50. Sazawal, V., Want, R. & Borriello, G. (2002). *The Uni-gesture Approach: One-Handed Text Entry for Small Devices*, Mobile HCI 02, Pisa, Italy, 2002.
51. Sharples, M., Corlett, D. & Westmancott, O. (2002). *The Design and Implementation of a Mobile Learning Resource*, Personal and Ubiquitous Computing, Volume 6, Number 1-6, 2002.
52. Silfverberg, M. (2003). *Using Mobile Keypads with Limited Visual Feedback: Implications to Handheld and Wearable Devices*, Mobile HCI 03, Udine, Italy, 2003.
53. Spain, K. A., Phipps, C. A., Rogers, M. E. & Chaparro, B. S. (2001). *Data Collection in the Palm of Your Hand: A Case Study*, International Journal of Human-Computer Interaction, Vol. 13, No. 2, 2001.
54. Terveen, L., McMackin, J., Amento, B. & Hill, W. (2002). *Specifying preferences based on user history*, CHI '02, Minneapolis, Minnesota, USA, 2002.
55. Van de Sluis, R., Eggen, B., Jansen, J. & Kohar, H. (2001). *User Interface for an In-Home Environment*, INTERACT '01, Tokyo, Japan, 2001.
56. Ward, N. & Tsukahara, W. (2003). *A study in responsiveness in spoken dialog*, International Journal of Human-Computer Studies, Volume 59, Issues 1-6, 2003.
57. Watters, C., Duffy, J. & Duffy, K. (2003). *Using large tables on small display devices*, International Journal of Human-Computer Studies, Volume 58, Issues 1-6, 2003.
58. Wyeth, P. & Wyeth, G. (2001). *Electronic Blocks: Tangible Programming Elements for Preschoolers*, INTERACT '01, Tokyo, Japan, 2001.
59. Yee, K. (2003). *Peephole displays: pen interaction on spatially aware handheld computers*, CHI '03, Ft. Lauderdale, Florida, USA, 2003.
60. Zurita, G., Nussbaum, M. & Shaples, M. (2003). *Encouraging Face-to-Face Collaborative Learning through the Use of Handheld Computers in the Classroom*, Mobile HCI 03, Udine, Italy, 2003.


(This page is intentionally left blank – Appendix A are located on the next page)

Appendix A


CONFERENCES	Device	Technology	System	User Description	Number of Participants	Laboratory	Field
INTERACT (12)	Baber et al. (1999) [3]	Wearable Computer, HMD	Speech Recognition	Paramedic IS	Yes		Paramedic Training Center
	Baber et al. (2001) [4]	HMD, PDA, Tablet PC	IR	Museum Guide	Yes	28	Comparison of Three Different Prototypes
	Björk et al. (2001) [6]	PDA	WLAN, RF Sensors	Game	No	13	At Conference
	Lehikoinen (2001) [38]	Wearable Computer, HMD	GPS	Navigation Map, AR	Yes	10-12	Campus, Moving between buildings
	Nyberg et al. (2001) [44]	PDA, Mobile Phone, Hybrid		Information & Call Handling Systems	Yes	18	Ericsson Lab
	van de Sluis et al. (2001) [55]	"Token" (control device)	IR, RF Sensors	Media Control	No	24	Household
	Wyeth & Wyeth (2001) [58]	Electronic Blocks		Programmable Building Blocks (Toy)	Yes	28	Indoor Play Area, Investigator participate actively
	Baillie (2003) [5]	Mobile Phone		Multimodal Route Finder	Yes	12	Yes
	Brunnberg & Juhlin (2003) [14]	PDA	GPS	Kids Game	Yes	4	Downtown Stockholm in car
	Kjeldskov & Stage (2003) [33]	PDA	WLAN	Safety Critical Communication System	Yes	6	Simulated Context, Think-aloud
	Ma et al. (2003) [39]	Simulated PDA on Laptop		Speech Recognition Interface	Yes	14	Compare two variants system
Mobile HCI (21)	Graham & Carter (1999) [24]	Mobile Phone, ICE for Jaguars	Speech Recognition	In-Car Entertainment System	No, No	48, 30	Car Simulator (Game)
	Koppinen (1999) [35]	Handset, Touchscreen		In-Car Com. System (Safety Critical)	No		PC Simulator
	Bruijn et al. (2001) [13]	Laptop	WAP	Mobile Web Browser, RSVP	Yes	30	Yes
	Bohnenberger et al. (2002) [7]	PDA	IR	Context Aware Shopping Guide	Yes	20	Mockup of a Shopping Mall, Roleplay
	Chittaro & Camaglio (2002) [17]	Mobile Phone	WAP	Visualizing Bar Charts	Yes	20	Yes
	Hibino & Mockus (2002) [27]	PDA	WLAN	Mobile Messenger	No, No	10, 6	Yes
	Lamberts (2002) [37]	PDA		Connectivity UI	No		Yes
	Öquist & Goldstein (2002) [45]	PDA		RSVP	Yes	16	Dedicated Usability Lab
	Sazawal et al. (2002) [50]	PC, Handheld Device		Tilting UI for Writing	Yes, Yes	12, 4	Yes
	Aittola et al. (2003) [11]	PDA	WLAN	Library Guide	Yes	32	Yes
	Bornträger et al. (2003) [8]	PDA	GPS	Tour Guide	Yes	16	City of Lancaster, UK
	Bosman et al. (2003) [9]	Wearable Computer	Haptic Stimulator (Vibrator)	Direction Guide System	Yes	16	Campus
	Esbjörnsson et al. (2003) [20]	PDA	WLAN	Greeting System (Motorcyclists)	Yes	6	Driving route on motorcycle
	Fithian et al. (2003) [21]	Hybrid		Event & Meeting Planner	Yes	9	Campus, Talk-aloud, Scenario-based evaluation
	Giller et al. (2003) [22]	PDA, Mobile Phone			Yes, Yes, Yes	10, 10, 10	Yes
	Jones et al. (2003) [29]	PDA		Search System	No	3	Yes (2 weeks)
	Peltonen et al. (2003) [48]	PDA	WLAN	Tour Guide	Yes	10	City of Oulu
	Silfverberg (2003) [52]	Mobile Phone		Keyboard (Tactile)	Yes	12	Yes (Sitting at table)
	Zurita et al. (2003) [60]	PDA	WLAN	Collaborative Learning	Yes	11	Class Room, Discount Usability Engineering

Table 7: The complete listing of the chosen papers and their categorisations (1 of 4) 

CONFERENCES	Data Collection	Tasks	Data Analysis	Results	Comparison	
INTERACT (12)	Baber et al. (1999) [3]	Time Measures, Discussion, Observation	Yes	Performance measures	Yes	
	Baber et al. (2001) [4]	Observation, Questionnaire, Interviews	Yes	ANOVA, Performance measures	Yes	
	Björk et al. (2001) [6]	Logs, Interviews	No		Yes	
	Lehikoinen (2001) [38]	Interviews, Observation	Yes		Yes	
	Nyberg et al. (2001) [44]	Benchmark, Video, Questionnaire (mental workload, subjective satisfaction)	Yes		Yes	
	van de Sluis et al. (2001) [55]		No		Yes	
	Wyeth & Wyeth (2001) [58]	Video, Observation	No	Level of enjoyment, attention, and interest.	Yes	
	Baillie (2003) [5]	Questionnaire (Paradise Method), Interview, Input Data, Video, Notes	Yes	Conversational Analysis, Note Incidents	Yes	Yes
	Brunnberg & Juhlin (2003) [14]	Video	No	Facial expression, appearance, movement of device, gaze, and spontaneous comments	Yes	
	Kjeldskov & Stage (2003) [33]	Video, Video Logs, Observation, Group Interview	Yes	Transscription of video	Yes	
	Ma et al. (2003) [39]	Questionnaire, Interview, Logs, Subjective Ratings	Yes	Time	Yes	
	Mobile HCI (21)	Graham & Carter (1999) [24]	Questionnaire, NASA-TLX	Yes		Yes
Koppinen (1999) [35]		Interview, Observation	No		Yes	
Bruijn et al. (2001) [13]		Video	No	Time analysis, ANOVA	Yes	
Bohnenberger et al. (2002) [7]		Observation, Questionnaire, Interviews	Yes	Performance, T-test	Yes	
Chittaro & Camaggio (2002) [17]			Yes	Time for completion, corectness, Wilcoxon-test, T-test	Yes	
Hibino & Mockus (2002) [27]		Log, Questionnaire	No	Statistical calculations	Yes	
Lamberts (2002) [37]			Yes		No	
Öquist & Goldstein (2002) [45]		NASA TLX, Audio & Video Recording, Benchmark	Yes	Words per minute, Various time aspects	Yes	
Sazawal et al. (2002) [50]		Observation, Interview	Yes	ANOVA	Yes	
Aittola et al. (2003) [1]		Interview, Observation, Questionnaire	Yes	Degree of labour required	Yes	
Bornträger et al. (2003) [8]		Observation, Log, Interview	No	Use of audio and movement, Effect of group size, Use of headphones, T-test	Yes	
Bosman et al. (2003) [9]		Observation, Interviews, Likert-scale	Yes	Time, Errors, T-test, Two-Tailed Wilkoxon signal rank test	Yes	
Esbjörnsson et al. (2003) [20]		Semi Structured Interview	No		Yes	
Fithian et al. (2003) [21]		Interview, Observation, Questionnaires, Benchmark, Likert-scale	Yes	Demographic data, Completed tasks, Time for completion	Yes	
Giller et al. (2003) [22]		Interviews, Questionnaires	Yes	ANOVA	Yes	
Jones et al. (2003) [29]		Interview, Diary	No		Yes	
Peltonen et al. (2003) [48]		Questionnaire, Observation	Yes	Subjective user scales	Yes	
Silfverberg (2003) [52]		Time Log, Subjective Ratings, Key presses, Error rate	Yes	Statistical calculations, Key presses, Error rate	Yes	
Zurita et al. (2003) [60]		Interview, Observation Recorded on Special Forms.	Yes	Compare against usability heuristics	Yes	

Table 8: The complete listing of the chosen papers and their categorisations (2 of 4) 

CONFERENCES	Device	Technology	System	User Description	Number of Participants	Laboratory	Field	
CHI (12)	Cheverst et al. (2000) [15]	Tablet PC	WLAN	Tourist Guide	Yes	60		Talk-aloud
	McClard & Somers (2000) [40]	Tablet PC	WLAN	WWW Usage	Yes	28		Household (Long term)
	Nelson et al. (2001) [42]	PDA, Mobile Phone		Mobile Phone Non-Verbal Calls	Yes	9	Yes	
	Isaacs et al. (2002) [28]	PDA	Modem	Mobile Messenger (Text & Sound)	Yes	28		Long term
	Pirhonen et al. (2002) [49]	PDA		Multimodal Audioplayer	Yes, Yes	15, 6	Stair master	Corridor (University)
	Terveen et al. (2002) [54]	Mobile Phone		Setting User Preferences	Yes	24	Yes	
	Brewster et al. (2003) [10]	PDA, Wearable Computer		Multimodal Audio Feedback	Yes, Yes	18, 20	Room (University)	
	Brown et al. (2003) [12]	PDA	Ultra Sonic Tracking System	Mixed Reality Museum Guide	Yes	30		Yes
	Kaikkonen & Roto (2003) [32]	Mobile Phone	GPRS, WAP	Browser/XHTML Application	Yes, Yes	20, 10	Think-aloud	
	Klemmer et al. (2003) [34]	PDA	Barcode Scanner	Digital Video Interviews (Books with voices)	Yes	13		Participants' Workplace
	Newcomb et al. (2003) [43]	PDA		Shopping Guide	Yes	5		Retail Environment, Think-aloud
Yee (2003) [59]	PDA	Mouse Based Tracking Device	Peephole Display	Yes	24	Yes		
CSCW (2)	Milewski & Smith (2000) [41]	PDA, PC	WLAN	Wireless Address Book	Yes	15		Yes (6 weeks)
	Grinter et al. (2002) [25]	PDA	WLAN	Museum Guide	Yes	47		Historical House in Woodside, California
JOURNALS	Device	Technology	System	User Description	Number of Participants	Laboratory	Field	
BIT (3)	Labiale (2001) [36]	Small Screen		Road Guidance Information	Yes	32	Car, Environmental simulator	
	Gupta et al. (2002) [26]	Simulated Car	Force Feedback	Car Adverse Condition Warning System	Yes	25	Simulates Car	
IJHCI (1)	Spain et al. (2001) [53]	PDA		Data Collection in Field	Yes	8		Yes
IJHCS (2)	Ward & Tatsukawa (2003) [56]	Laptop, Touchscreen		Note Taking	Yes, Yes	4, 10	Yes	University Lectures
	Watters et al. (2003) [57]	PDA, Simulated Handheld	WAP	Presenting Tables on Small Displays	Yes	84	Yes	
PUC (7)	Brewster (2002) [11]	PDA		Sound Supported Mobile Systems	Yes	44	Sitting at table	Walk along specified route
	Cheverst et al. (2002) [16]	PDA	WLAN	City Guide	No	20		Lancaster Castle, Talk-aloud
	Chittaro & Dal (2002) [19]	Mobile Phone	WAP	Movie Reservation	Yes	40		Home Environment
	Sharples et al. (2002) [51]	Tablet PC	USB Camera, WLAN, PCMCIA Cardphone	Collaborative Learning (Children)	Yes	38		Yes
	Åkesson & Nilsson (2002) [2]	In-car Box With Input-wheel	Speech Recognition	In-Car Music Selection	Yes	3		Users own car
	Goldstein et al. (2003) [23]	Multipurpose Mobile Handset	Camera, WAP	Multi Purpose	Yes	14	Erison Lab, Talk-aloud	
	Kaasinen (2003) [30]	PDA, Mobile Phone	GPS, WAP, Bluetooth	Yellow Pages (Guide System)	Yes	55	Yes	Yes
TOCHI (2)	Pascoe (2000) [47]	PDA	GPS	Data Collection (Girafs)	Yes	1		Kenya (2 months)
	Palen & Salzman (2002) [46]	Wireless Phone, Mobile Phone			Yes	19		Yes (6 weeks)

Table 9: The complete listing of the chosen papers and their categorisations (3 of 4) 

CONFERENCES		Data Collection	Tasks	Data Analysis	Results	Comparison
CHI (12)	Cheverst et al. (2000) [15]	Observations, Audio Recording, Time Log, Semi-structured Interview.	No		Yes	
	McClard & Somers (2000) [40]	Log, Pre & Post Interviews, Observation	No		Yes	
	Nelson et al. (2001) [42]	Observation, Video, Logs, Discussion, Open Ended Questions.	Yes	Written summaries, Transcription	Yes	
	Isaacs et al. (2002) [28]	Log, Interviews, Informal Conversations, Video, Email Feedback.	No	Statistical Calculations	Yes	
	Pirhonen et al. (2002) [49]	Observation, Interview, Log, Questionnaire, Video, NASA TLX, User drawings.	Yes	PWS, T-test, Time, Errors, Completion, Annoyance	Yes	Yes
	Terveen et al. (2002) [54]	Interview	Yes	Time, ANOVA, One Factor	Yes	
	Brewster et al. (2003) [10]	NASA TLX, PWS, CRS, Observation	No	PWS, CRS, Comfort, Annoyance, Tukey HSD, ANOVA	Yes	
	Brown et al. (2003) [12]	Video, Observation, Log, Interview, Audio Recording	Yes	Transcript, Interactional Analysis	Yes	
	Kaikkonen & Roto (2003) [32]	Video (Mini camera mounted on device), Video & Audio Recording, Observation, Questionnaire, Subjective ratings	Yes	Task execution time	Yes	
	Klemmer et al. (2003) [34]	Video, Questionnaire, Observation	Yes	Time, Access statistic, Usage style	Yes	
	Newcomb et al. (2003) [43]	Interview, Questionnaire, Audio Recording, Observation	Yes	Shopping behaviour while holding the PDA	Yes	
Yee (2003) [59]	Questionnaire, Observation	Yes	Task time	Yes		
CSCW (2)	Milewski & Smith (2000) [41]	Log, Questionnaire, Interview	No		Yes	
	Grinter et al. (2002) [25]	Audio Recording, Interview, Log, Video	No		Yes	
JOURNALS		Data Collection	Tasks	Data Analysis	Results	Comparison
BIT (3)	Labiale (2001) [36]	Video (4 Cameras), Questionnaire	Yes	Time, Focus, Error rate (Glance data software & Kronos software), ANOVA	Yes	
	Gupta et al. (2002) [26]	Questionnaire	Yes	ANOVA	Yes	
IJHCI (1)	Spain et al. (2001) [53]	Observation (Notes by Test Administrator)	Yes		Yes	
IJHCS (2)	Ward & Tatsukawa (2003) [56]	Questionnaire, Video	Yes		Yes	Yes
	Watters et al. (2003) [57]	Questionnaire, 7 Hypothesis they test	Yes	MANOVA, ANOVA	Yes	
PUC (7)	Brewster (2002) [11]	NASA-TLX	Yes	ANOVA, Tukey HSD	Yes	Yes
	Cheverst et al. (2002) [16]	Interview, Observation	No		Yes	
	Chittaro & Dal (2002) [19]	Video, Questionnaire, Log	Yes	Time for completion, NAC-Wilcoxon test, SCC-Mann-Whitney test, Subjective evaluation, ANOVA	Yes	
	Sharples et al. (2002) [51]	Questionnaire, Video, BBC Camera Crew	Yes		Yes	
	Åkesson & Nilsson (2002) [2]	Observation, Notes, Interview	Yes		Yes	
	Goldstein et al. (2003) [23]	Video, Questionnaire, Benchmark, Satisfaction-rating	Yes	Efficiency, Effectiveness	Yes	
TOCHI (2)	Kaasinen (2003) [30]	Questionnaire, Interview, Video & Audio Recording, Photos	Yes		Yes	
	Pascoe (2000) [47]	Log	Yes	Amount of successful observation	Yes	
	Palen & Salzman (2002) [46]	Interview, Voice Mail Diaries, Phone Records (Log), Diaries	No	Socio technical approach	Yes	

Table 10: The complete listing of the chosen papers and their categorisations (4 of 4)



Usability Evaluation of a Mobile System: Comparison of a Laboratory and a Field Evaluation

Christian Monrad Nielsen, Michael Overgaard, Michael Bach Pedersen & Sigge Stenild
Department of Computer Science, Aalborg University, Denmark
{monrad, mio, mbp, stardust}@cs.auc.dk

Abstract

This article presents two user-based usability evaluations of a mobile system, which are conducted in laboratory and in a field-based setting respectively. The purpose of the evaluations is to obtain experience regarding similar data collection methods in both approaches, and to compare the laboratory and the field-based approach on problems identified and on the overall usability rating of the mobile system. The conclusion is that it is worthwhile to conduct evaluations in field settings as it identifies significantly more usability problems and reveal problems related to themes not otherwise identified in the laboratory evaluation. Furthermore, it concludes that it is possible to diminish the problems of complicated data collection, which small screens of mobile phones and field evaluations inflict, through the use of a mini camera mounted on the mobile phone.

1. Introduction

Research in usability evaluation of stationary systems is a well-established research area [Karat et al., 1992] [Gray & Salzman, 1998], but research within usability evaluation of mobile systems field is not yet as extensive [Pedell et al., 2003]. Therefore, a set of usability evaluation methods and data collection techniques has not yet been established within the field of mobile HCI [Kjeldskov et al. 2004] [Pirhonen et al., 2002].

It is important that systems for mobile devices are tested in realistic settings, since testing in a conventional usability laboratory is not likely to find all problems that would occur in real mobile usage [Johnson, 1998]. However, usability evaluation in the field is time consuming and subject to the problem of complicated data collection and limited control [Johnson, 1998] [Baillie, 2003] [Kjeldskov & Stage, 2004] [Kjeldskov et al., 2004]. Usability evaluations in laboratory settings are not troubled with these problems, but instead they lack the realism of the real life context [Kjeldskov & Stage, 2004].

The importance of field-based usability evaluation of mobile systems is also in focus in Baillie [2003] and Pirhonen et al. [2002]. Both of these papers focus on

the similarities and differences of the two approaches when testing mobile systems. Based on their observations they identified different interaction behaviors in the laboratory and in the field settings. Baillie [2003] concludes that it is worthwhile carrying out studies in the field, even though it is problematic due to difficulties in capturing the events on the screen, and the interaction between the user and the mobile device. However, the dissimilarity in collecting data in field and laboratory evaluations can have an impact on the results, as Pirhonen et al. [2002] describe. They conclude that both of the studies are needed to get a reliable overall view of usability, but they also describe the dissimilarity in their results between laboratory and field evaluation as a consequence of the differences in quantitative and qualitative data collection techniques [Pirhonen et al., 2002].

Nielsen et al. [2004] presents a literature study on user-based usability evaluation of mobile systems, going back five years and choosing a wide selection of papers from key journals and conference proceedings within the HCI research area. They found that 55% of the papers performed field based evaluations and that 9% performed both laboratory and field evaluations. This study indicates that most researchers acknowledge the importance of field usability evaluation for mobile systems.

Not all studies show that usability evaluation in the field is worth the extra effort [Kjeldskov et al., 2004] [Kjeldskov & Stage, 2004]. Kjeldskov & Stage [2004] present and evaluate six different techniques for evaluating the usability of a mobile system in laboratory settings. These techniques are compared to a field evaluation. Their tests reveal that the simplest technique, where the user is sitting at a table, was better than any of the other techniques, when focusing on identifying usability problems. Furthermore, Kjeldskov et al. [2004] describe that when setting up a realistic laboratory evaluation, the field evaluation achieves very little added value.

This paper has two purposes. Firstly, it is to present our experiences in conducting two similar user-based usability evaluation in both laboratory and field-based settings, using the exact same data collection techniques; video recordings, observations, and interaction logs. Secondly, we want to compare the two evalua-

tions, based on problem lists and on measurements of usability, according to the ISO 9241-11 definition [ISO, 1998].

In the next section, we describe the system that is evaluated. In section 3 and ,4 the method and procedure for the evaluations are described, while section 5 presents the results of both evaluations. The results consist of a joint problem list and usability assessments according to the ISO 9241-11 definition. Section 6 is a discussion of the findings, and section 7 provides the conclusion.

2. System Description

The evaluated system is used for registering the use of equipment, materials, mileage, and working hours for workmen. The system runs on a regular Sony Ericsson T68i mobile phone, with an AirClic barcode scanner attached and uses GPRS for transmitting data.



Figure 1: Using the barcode scanner for executing a system command.

The system is a part of a larger administrative system that also includes a web-based part, which is not covered in our evaluation. In order to use the system the user has to use a sheet of paper, which contains all necessary barcodes. This sheet has barcodes for system commands, tools, equipment, and materials. When a user needs to register some kind of information, he has to scan the appropriate barcode, which provides access to menus in the system. Figure 1 pictures the system.



Figure 2: A typical menu screen from the system.

Figure 2 shows a typical menu screen from the system. The screen contains two elements that can be edited by the user. Selection is shown in inverted colours. The user accepts and registers information in the system by selecting “Ok!” in the menus.

3. Evaluation Method

Two user-based usability evaluations of the system were conducted, one in a usability laboratory and one in field-based settings. The method for both evaluations was based on Rubin’s [1994] guidelines on how to setup and conduct usability tests. The purpose of the two evaluations was to evaluate the system through the performance and preference by the intended users. It was done by recording the number of tasks completed, time of completion, furthermore we noted the number of errors and difficulties in using the mobile barcode scanner and the users’ attitude towards the system.

3.1. Design

A technical teacher at Vitus Bering CEU, a technical high school described the initial task proposals, which were then modified to fit the purposes of the evaluations, and this resulted in nine specific tasks. Basically the tasks were identical for the laboratory and the field-based evaluation, but were different in a single task where the field evaluation included a physical aspect in order to complete the task.

In addition to the tasks, a pre-questionnaire was made to gather data of the participant’s experience with different types of information technology. As a session follow-up a NASA-TLX test [Hart & Staveland, 1988] was performed alongside a post-questionnaire. The purpose of the post-questionnaire was to reveal the participant’s subjective opinion about the evaluation, the system, and the usage of it. This was done by rating different issues on scales from one to seven, with seven being the best.

Two separate teams composed of a test monitor and a logger conducted the two evaluations, and each team conducted a pilot-evaluation prior to the respective evaluations.

3.2. Participants

The participants were all from Vitus Bering CEU. In age they ranged from 16 to 36, and were all apprentices in the field of earthwork-engineering. A total of 14 participants’ took part, divided into two groups of seven. Each group consisted of four participants from the basic stage of the apprenticeship and three from later stages. The majority of the participants rated themselves as having daily experience with mobile phones, but low familiarity with WAP services. Most of the participants had no or little experience with bar-

code scanners, but three were experienced in using the technology. Table 1 provides an overview of the participants experience with related technology. Their experience is rated from none (0) to very experienced (5).

	Participant	Age	Mobile Phones	WAP-services	Barcode Scanners
Laboratory	1	20	4	2	2
	2	20	4	0	2
	3	18	3	0	0
	4	18	3	1	2
	5	35	4	1	1
	6	26	4	0	2
	7	16	5	2	5
Field	1	36	3	0	4
	2	19	4	4	2
	3	35	3	0	4
	4	17	4	1	0
	5	17	4	0	2
	6	18	3	0	2
	7	17	3	1	2

Table 1: An overview of the participants' age and experience with different technologies.

A day before the laboratory evaluation the participants received two hours of introduction to the mobile system, where they were introduced to the functionality and got hands-on experience in using the barcode scanner.

3.3. Data Collection

With mobile systems in the field, capturing screen and interaction can present a challenge [Esbjörnsson et al., 2003] [Johnson, 1998] [Kjeldskov & Stage, 2004]. To accommodate this, Nokia has developed a mini-camera that can be mounted on the mobile phone [Nyyssönen et al., 2004].

A similar wireless device has been developed by the HCI Department at Aalborg University. The camera itself is mounted on a flexible wire-arm in order for it to bend into different positions. The camera is attached to the device evaluated using Velcro tape. The camera transmits a wireless video signal to a receiver that records it on digital video. As the camera focus on the device, the picture is steady and allows for closer examinations, see Figure 3.

This mini-camera was used as the main data collecting technique alongside with observation. Additionally, a system logs was used, which recorded the commands executed along with timestamps for each user. These data collection techniques were utilized in the same way in both evaluations.



Figure 3: The mini-camera with the mobile barcode scanner system attached.

4. Procedure

Before the evaluation session the participant answered the pre-questionnaire. Afterwards the test monitor gave an introduction to the evaluation. The participant then worked through as many of the nine written tasks as possible, which was handed to him one by one. During the evaluation the participant was encouraged to *think-aloud*. If the test monitor observed that the participant was helplessly stuck, the evaluation was continued from the proceeding task, even though the current task was not completed. The participant was encouraged to verbally indicate when he felt he had completed a task. Each session was limited to 40 minutes. After the session the participant was debriefed about the session. He was then taken to another location to answer the NASA-TLX scorecards. They were given to him in randomly order to avoid order effect [Frøkjær et al., 2000]. Lastly the participant answered the post-questionnaire.

4.1. Laboratory

The laboratory evaluation took place in a state-of-the-art usability laboratory at Aalborg University, where the participant was placed at a table with the test monitor behind him to his right side. The logger was placed in an adjacent control room behind a one-way mirror.

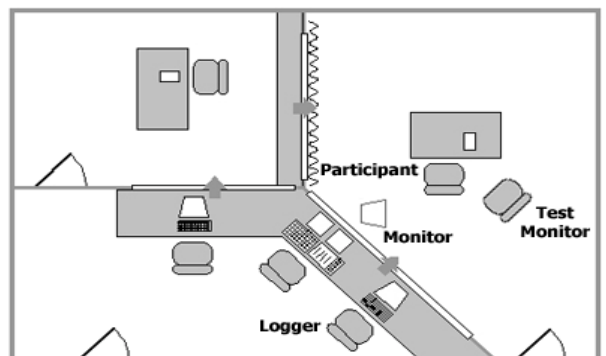


Figure 4: An overview of the state-of-the-art usability laboratory.

Four cameras recorded the session; one in front of the participant and the test monitor, one angled from above, a close-up of the table and the mounted mini-camera, see Figure 5. The images from the mini-camera were visible to the test monitor via a monitor screen placed behind the participant, see Figure 4. A high fidelity microphone recorded the sound.



Figure 5: The combined camera recordings.

4.2. Field-based Approach

The field-based evaluations were conducted in a warehouse at Vitus Bering CEU. The warehouse is designed to accommodate practical learning in the construction business and its interior reflects real working environments, which made it ideal for evaluation purposes. The participant was placed at a specified working area with the test monitor beside him. During the session, the logger was close by, primarily in order to observe the evaluation and take notes, and secondly to operate the recording equipment. The session was recorded by microphone attached on the user, and the mini-camera mounted on the mobile phone. Figure 6. show a participant during the evaluation.

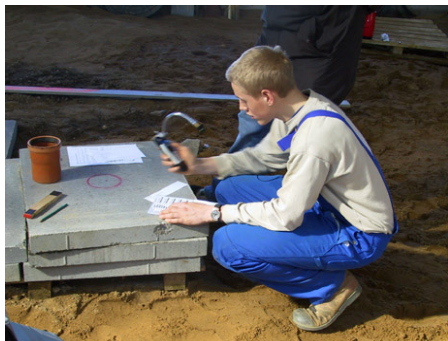


Figure 6: One of the participants solving a task during the field-based evaluation.

5. Data Analysis

The evaluator effect is not negligible in think-aloud tests and has an impact on both the problems identified and their severity rates [Hertzum & Jacobsen, 2001]. In order for us to minimize this effect and to create more valid results, each member, in either the laboratory or the field-based evaluation team, evaluated the respective test data. Figure 7 illustrates the process of working out the problem list for each evaluation.

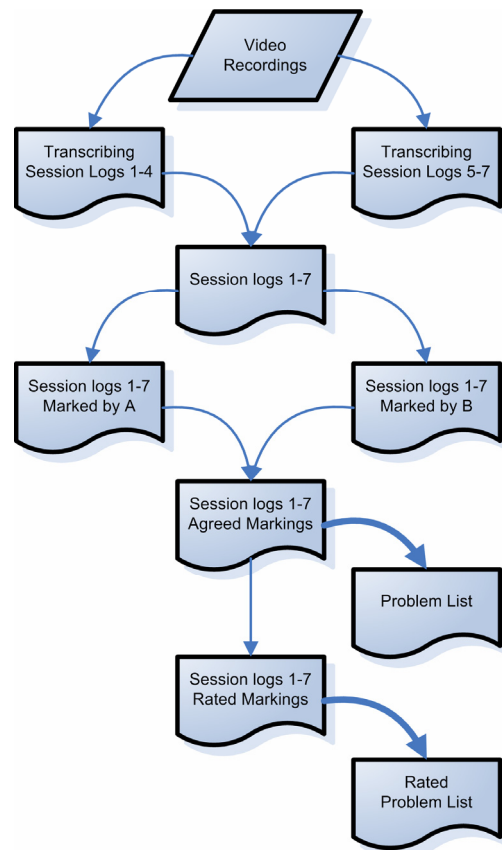


Figure 7: An overview of the process of elaborating the problem list for each of the evaluations.

Each team divided the data between the team members and wrote a session log for each participant. All the logs were then given to each member who read each of them and marked places with usability problems. No ordering or severity rating was done at this stage. Afterwards the two team members compared session logs and discussed each marking until consensus was reached. This resulted in a problem list containing the problems and in which sessions they occurred. Each marking in each session was then severity rated by the team members together according to the severity ratings proposed by Molich [2000], and the highest rating of each problem was noted resulting in a severity rated problem list.

5.1. Joint Problem List

In order to compare the evaluations, a combined problem list was needed. Figure 8 illustrates how the two problem lists from the laboratory and the field-based evaluation were combined into a joint list.

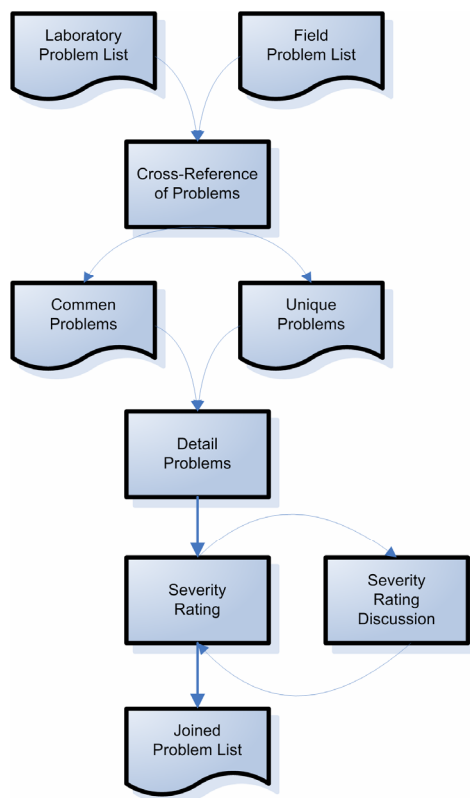


Figure 8: The process of combining the two problem lists into a joint list.

After the completion of each problem list, one member from each team reviewed the two lists, and made cross-references between the problems in order to find common and unique problems. These problems were then discussed and elaborated if needed. If overlap between problems was found, the overlap was seen as one problem and the remaining parts became independent problems. After detailing the problems the severity ranking of each problem was reviewed and severity was up- or downgraded if needed. All team members discussed the ratings until consensus was reached. The result was a joint problem list for both evaluations

5.2. Usability Themes

The next step was categorizing the usability problems by theme. The categorizations were done in order to identify what the main usability problems revolved around. Furthermore, it was done in order to compare the two usability evaluations and what types of usability problems that were discovered by them. The themes

originate from Nielsen et al. [2004] were a similar categorization was done. Below is a brief definition of each of the themes.

Ergonomics, relates to the physical characteristics of interaction [Dix et al., 1998].

Task Flow, is about the sequence of steps of which tasks should be conducted [Dix et al., 1998].

Feedback, concerns how the system sends information back to the user about what action has been done [Norman, 1990] and system notifications in relation to system events.

Consistency, relates to consistency in command naming, labels across different screens and consistency in the structure of commands [Dix et al., 1998].

Interaction Styles, covers the design strategy and determines how the system's interactive resources are organized [Newman and Lamming, 1995].

Cognitive Load, concerns the amount of cognitive resources needed to use the system [Pedell et al., 2003].

Information, regards how and what information is presented by the system at a certain time [Pedell et al., 2003].

Navigation, is about how the user navigates through the screens of the system [Pedell et al., 2003].

User's Mental Model, The user's model is the mental model developed through interaction with the system [Norman, 1990].

Affordance, refers to problems on how the user perceives the properties of an object, and what the actual properties of that object are [Norman, 1990].

Mapping, relates to how controls and displays should exploit natural mappings, which take advantage of physical analogies and cultural standards [Norman, 1990].

Visibility, concerns which controls are available in the user interface at a specific time [Norman, 1990].

5.3. Overall Usability

As a part of the comparison of the two tests we investigated if a difference in overall usability was present between the two evaluation approaches. When the usability of a system is to be assessed, Frøkjær et al. [2000] states that it is important to investigate all the distinct aspects of usability as defined by ISO's definition, namely *effectiveness*, *efficiency*, and *satisfaction* [ISO, 1998]. If only one or two aspects of this usability definition are considered, an unreliable conclusion of the overall usability may be drawn. On the other hand, the evaluation measures should fit the specific situation and rely on a firm understanding of how tasks, users,

and technologies constitute use situations in a specific use context. The evaluations in this paper assess the usability of the system by:

- *Efficiency*, by task completion time.
- *Effectiveness*, by the number of tasks completed.
- *Satisfaction*, through subjective user rating on a satisfaction scale.

Task completion time made it possible to investigate if the users used more time in one of the evaluations. Comparison of the number of completed tasks gave an indication of how well the users, in either the laboratory or the field-based approach, were in achieving the given tasks by utilizing the system. The satisfaction rating showed how well the users liked the system more after the respective tests had been conducted.

6. Results

This section provides an overview of the problems identified in the two usability evaluations. It starts by outlining the general usability problems according to the laboratory or the field-based evaluation, then the severity ratings, and finally the problem themes. Next, it investigates the overall usability as defined by ISO.

6.1. The Joint Problem List

The usability evaluations identified 76 different usability problems altogether. 27 usability problems were categorized as critical, 30 problems as severe, and 19 as cosmetic, see Table 2.

Critical	Severe	Cosmetic	Total
27	30	19	76

Table 2: Number of total identified usability problems, and the distribution of these in the severity categories.

6.1.1. Evaluation Type and Severity

The laboratory evaluation identified 104 occurrences of usability problems and in the field-based evaluation 123 instances were uncovered. A t-test shows no significant difference, between the two evaluations ($t_{12}=0.83$, $p>0.1$) on this matter. Removing multiple occurrences of the same usability problem, leaves 48 unique problems identified in the laboratory evaluation and 60 unique problems identified in the field evaluation, see Figure 9. A two tailed large sample test for population proportions shows a significant difference in the amount of usability problems identified in the laboratory- and in the field-based evaluation ($z=2.85$, $p=0.006$).

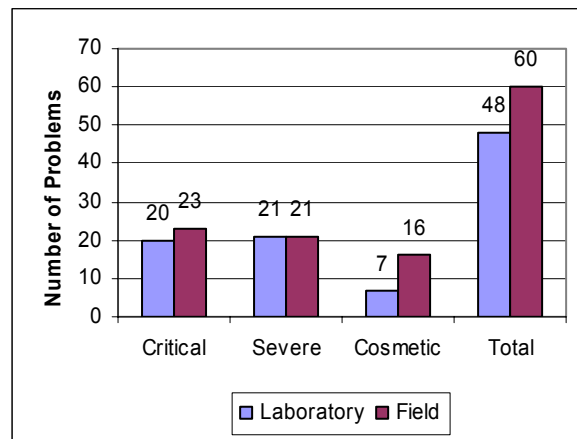


Figure 9: Number of usability problems found in the laboratory and the field evaluation, and how these are distributed amongst severity categories.

6.1.2. Unique Problems and Severity Ratings

42% (32 out of 76) of the usability problems were identified in both evaluations, which means that the remaining 58% (44 out of 76) of the problems were unique for either the laboratory or the field evaluation. This result suggests that it might be important to conduct both evaluations, as Pirhonen et al. [2002] describe, in order to find the most usability problems. On the other hand the result could indicate that different evaluators identify different problems, as pointed out by Hertzum & Jacobsen [2001] and Molich et al. [2004].

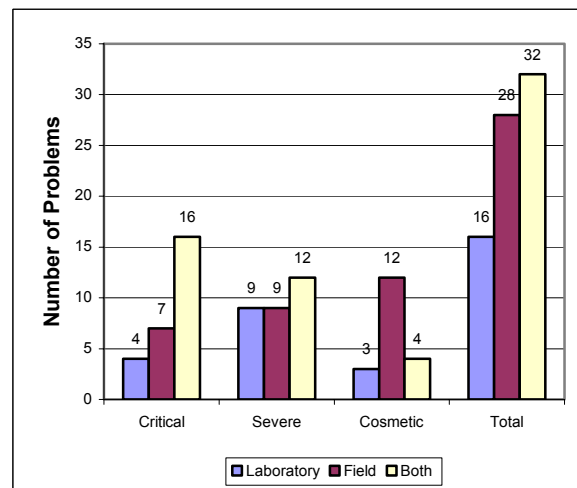


Figure 10: Number of problems, which are unique or found by both evaluations, combined with severity.

Figure 10 shows that a total of 11 of all the uniquely found usability problems were critical, 18 were severe, and 15 were cosmetic. A two tailed large sample test for population proportions shows that there is a significant difference in the critical category ($z=1.96$, $p=0.05$) and the severe category ($z=2.24$, $p=0.025$), when com-

paring number of problems identified in only one evaluations and problems identified by both. In the cosmetic category, the difference is very significant ($z=6.19$, $p=0.001$). This indicates that the more severe a problem is, the more likely it is to be identified in both evaluations.

6.1.3. Evaluation Type, Themes, and Severity

The usability evaluations identify problems that are similar, which result in overlap between the problems of the two evaluations, see Table 3. *Distinct Problems* in Table 3 express the number of distinct problems for both usability evaluations.

In both evaluations, the themes *feedback* and *information*, contains the largest part of the usability problems, in total 43.5% (33 out of 76). This clearly indicates to which parts the problems in the system were related. *Feedback* accounts for 18 of these problems, while 15 problems were related to *information*. A comparison of the laboratory and the field-based evaluation shows that there is no significant difference in the amount of problems identified in these two themes.

Two other themes have a high problem rate, *affordance* and *task flow*, each with 8 occurrences, account for 21.0% of the problems identified. Looking at the distribution of these problems between the two evaluations, it can be seen that problems related to *affordance* are equally found in the both evaluations, while more *task flow* related problems are apparent in the field-based approach. However the difference it is not significant between the two evaluations ($z=1.40$, $p>0.05$).

The remaining problems account for 35.5% (27 out of 76) of the total usability problems and are distributed between the last eight themes. If a comparison of laboratory and field is made, see Table 3. It shows that the themes *cognitive load* and *interaction style* is identified only in the field-based evaluation. The reason *cognitive load* only is present in the field evaluation, might be as Baillie [2003] also describes, that in realistic settings, the users are easier to become frustrated and thereby increasing the cognitive load. *Interaction style* can be explained by the more realistic context of use that exists in the field evaluation. This means that the user has to balance mobile phone and barcodes in the hands, and additionally that he sometimes has to squad down.

Comparing themes with severity categories, it can be seen that *feedback* and *information* accounts for most of the critical problems, see Table 3, which corresponds with that the biggest amount of problems identified are found within these categories. Other themes of interest are *navigation* and *consistency*. All of the problems related to *navigation* are categorized as critical, and *consistency* has a significant part of its problems categorized as critical ($z=5.4$, $p<0.001$).

	Laboratory				Field				Distinct Problems
	Critical	Severe	Cosmetic	Total	Critical	Severe	Cosmetic	Total	
Affordance		6		6		4	2	6	8
Cognitive Load						2		2	2
Consistency	3			3	3		1	4	4
Ergonomics	1		1	2	1		3	4	5
Feedback	6	6	1	13	4	7	1	12	18
Information	6	3	2	11	7	4	2	13	15
Interaction Style					1		3	4	4
Mapping		1	1	2			2	2	3
Navigation	2			2	2			2	2
Task Flow			2	2	4	1	2	7	8
User's Mental Model	2	1		3	1	1		2	3
Visibility		4		4		2		2	4
Total	20	21	7	48	23	21	16	60	76

Table 3: The problems from the two usability evaluations combined with themes and severity.

6.1.4. Unique Problems, Themes and Severity

Looking at the unique problems of each evaluation and combining them with themes and severity, we see that both laboratory and field identify several unique *feedback* problems. This is not surprising since *feedback* is one of the themes with most related usability problems.

Furthermore, the field evaluation identifies four unique critical *task flow* problems and four unique *interaction style* problems, see Table 4. *Interaction style* has been mentioned before, see 6.1.3, but *task flow* is a new finding. The presence of critical *task flow* problems can be explained by the more realistic context of use, which the field-based evaluation provides.

	Laboratory				Field			
	Critical	Severe	Cosmetic	Total	Critical	Severe	Cosmetic	Total
Affordance		2		2			2	2
Cognitive Load						2		2
Consistency							1	1
Ergonomics			1	1			3	3
Feedback	3	2	1	6	1	3	1	5
Information		2		2	1	3		4
Interaction Style					1		3	4
Mapping		1		1			1	1
Navigation								
Task Flow			1	1	4	1	1	6
User's Mental Model	1			1				
Visibility		2		2				
Total	4	9	3	16	7	9	12	28

Table 4: The unique problems from the two usability evaluations combined with themes and severity.

6.2. ISO 9241-11

In assessing the overall usability as defined by ISO 9241-11, the criterion for the tests were that each participant should be able to complete the nine tasks within the 40 minutes time-scope of each session.

6.2.1. Efficiency

The overall completion time for each task is based on the completed instances of a task. Tasks not completed are not included.

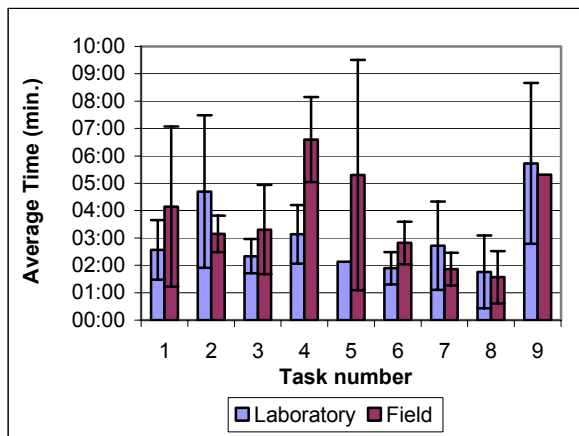


Figure 11: The average completion time of tasks in the laboratory and field-based evaluation. The black lines indicate the standard deviations.

Figure 11 shows a comparison between the average time used for each task, in either the laboratory or the field-based evaluation, along with their standard deviations. A t-test show that the difference in completion time for task 4 was very significant ($t_{12}=4.62$, $p<0.005$). This can be explained by the fact that the participants in the field-based evaluation had an extra aspect to the task, which was measuring of a flag.

Furthermore it can be seen that there exists a difference in completion time concerning task 6, which is significant ($t_{12}=2.56$, $p=0.025$), despite no difference in task description existed. Task 5 in the laboratory evaluation was only completed by one participant, which explains the absence of an indication of standard deviation.

6.2.2. Effectiveness

A task was categorized as complete if the end result was equal to a predefined solution. On the contrary a task was not complete; if the end result differed from the solution, if the task was interrupted by the test monitor, or not applied due to the limited time-scope of each session.

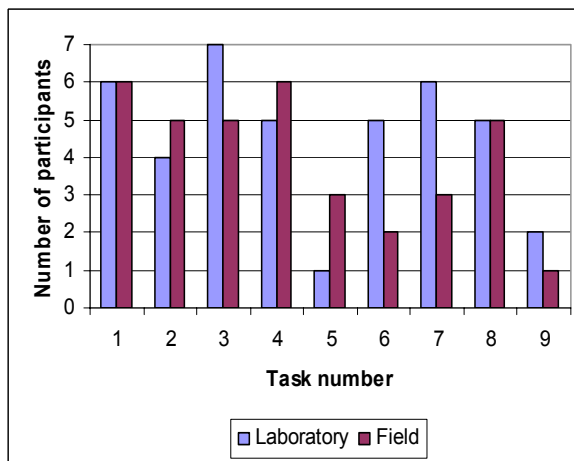


Figure 12: The number of participants that completed each task.

A significant difference in number of completed tasks is only present in task 7 ($z=1.67$, $p=0.048$). This indicates that no great distinction exists between the two evaluation approaches, when looking at the ability to complete the tasks. Figure 12 also illustrates that the least completed task was number 9, which was only completed by 21% (3 out of 14) of the participants. An explanation to this can be the complexity of the task and the time-scope of the evaluation.

6.2.3. Satisfaction

The participant's satisfaction was measured after the evaluation session by letting them rate their overall satisfaction with the system on a scale from one to seven, where seven was the best, see Table 5.

	Satisfaction	
	Laboratory	Field
Mean	5.29	5.00
Std. Dev.	1.28	0.93

Table 5: Satisfaction ratings of the system.

The difference between the average rating in each evaluation is not significant ($t_{12}=0.50$, $p>0.1$). This indicates that the participants' opinion of the system is the same, regardless of the evaluation approach.

6.3. Workload

A measurement of the workload of each evaluation was done through the NASA-TLX results. The average workload for the participants in the laboratory approach is 52.9 out of maximum a score of 100, while the average for the field-based evaluation is 58.4, see Figure 13. A t-test showed that the difference was not significant ($t_{12}=0.63$, $p>0.1$), which indicates that the participants, though being in more realistic settings, did not experience an increased workload.

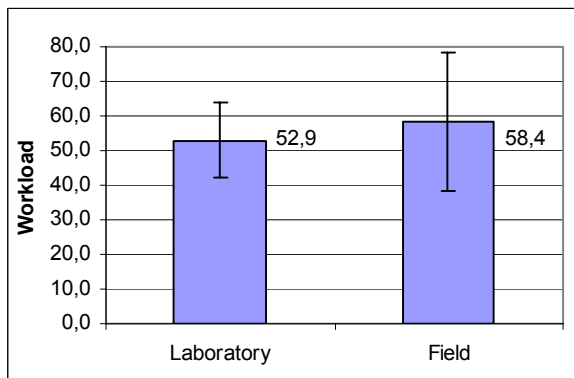


Figure 13: The average workload measured through NASA-TLX, and their standard deviations.

In the post-questionnaire the participants also rated the physical and mental demands needed in their evaluation sessions. There is no significant difference in the physical aspects of using the system ($t_{12}=0.63$, $p>0.1$). It can be argued that the reason for this is that the tasks performed only differ because of the physical aspect in task 4, and that the participants were standing, rather than sitting in the field-based evaluation.

A t-test of the post-questionnaires showed a very significant difference ($t_{12}=4.19$, $p<0.005$) in mental demands and a significant difference in frustration level ($t_{12}=2.04$, $p=0.05$), where both aspects were highest in the laboratory evaluation. This result contradicts the NASA-TLX result and can be an indication of that the laboratory evaluation for some reason results in greater demands for mental resources and leads to more frustration when using the system. Therefore, the results indicate that an unfamiliar environment and a less naturalistic setup have influenced the test participants, but to what degree are uncertain.

7. Discussion

In this paper we present our experiences in conducting two similar user-based usability evaluations in laboratory and field-based settings using the same data collection techniques. Secondly, we compare the two evaluations, based on the problem lists and the ISO definition about usability.

7.1. Data Collection

In the laboratory evaluation we combined the mini-camera recordings with recordings from three other cameras and presented them simultaneously in a quadro-view, see Figure 5. When the recordings were reviewed, it was difficult to see the screen of the mobile phone in detail. Therefore, it should be considered which view that contributes most in illustrating an evaluation situation, and make it the main focus on the screen. In the field-based evaluation, only a full-screen

view was available from the mini camera. This provided a good picture of the screen, but made it impossible to properly see interaction with objects in the environment, such as the barcodes. A second camera recording of the session in the field settings could have eased the difficulties in capturing the interaction.

We experienced that the wireless mini-camera recordings flickered during the laboratory evaluation. This might be due to Wi-Fi and other wireless services present at the premises for the evaluation. It should be considered if a wireless approach is indeed necessary for an evaluation, or if a cable version of a mini-camera, which provides more clear pictures, is sufficient.

7.2. Data Analysis

Differences existed between the problem lists from the evaluations. It can be argued that different data lies behind each list, and that this is one of the reasons for that differences exist in the identified problems, even though both evaluations had identical tasks descriptions, except from the physical aspect in task 4. The final result of the two evaluations is undoubtedly influenced by the evaluator effect, as analysis of the data require interpretation from the evaluators

7.3. Joint Problem List

Through our data analysis, we also experienced an evaluator effect when assembling the joint problem list. Several problems found in both evaluations were described in different ways or in different detail, but by discussing problems, and thereby reaching consensus, this effect was diminished.

58% of the problems identified in both evaluations are unique, which could indicate that it is important to conduct evaluations of both types if a broad and varied measurement of the usability of a mobile system is to be obtained. The 58% could also be the result of the evaluator effect, but it is interesting that the more severe a problem is, the more likely it is identified in both evaluations, as seen in section 6.1.2.

7.4. ISO

Concerning the ISO usability assessment, there are differences in the result of the aspects *efficiency* and *effectiveness*, which results in a better overall usability rating of the system in the laboratory evaluation. According to ISO [1998], this is not surprising as the context of use influence the usability of a system. This confirms that more realistic context settings in an evaluation provide more valid information about the overall usability of a system.

The NASA-TLX shows no significant difference in the workload between the two usability evaluations. The

ratings from the post-questionnaire contradict this as the mental demands and the frustration level are higher for participants in the laboratory evaluation. It can be argued that the reason for the difference can be found in how the ratings are obtained. In the NASA-TLX tests the participants were not able to directly see how their answers would influence the result, but in the post-questionnaire, the ratings were conspicuous.

7.5. Reading Disabilities

It became apparent during the evaluations that some of the participants had reading disabilities. This meant that in some sessions the test monitor had to read the tasks out loud. Furthermore, it became apparent when the participants were answering the different questionnaires, which also had to be read out loud and explained. One could argue that when reading and explaining the tasks and questionnaires, the test monitor might influence their perception of the tasks and questionnaires.

When the intended target group of a system include people with reading disabilities¹ it goes without saying that it has to be taken into consideration in the design and implementation of the system. As a result, a usability evaluation should be designed accordingly.

One of the reasons that we did not encounter the problem prior to the evaluation was that we conducted our pilot evaluations with internal participants, as prescribed by Rubin [1994]. This suggests that, when insufficient knowledge about the target users are available, the pilot test should be conducted using the target users, in order to obtain domain specific knowledge about these persons.

8. Conclusion

Based on the number and the nature of the identified problems, both in terms of severity and themes, and the results of the ISO, we present the following conclusions.

Both evaluations utilized identical data collection techniques, which provide a foundation for comparing the outcome of the evaluations. Data collection is difficult when conducting field evaluations. However, using a mini-camera to capture the screen of the mobile system was of great value to us. It provided excellent recordings for later analysis and added only a little hassle for the participants. The conclusion is that it is possible to diminish the problems of complicated data collection that small screens of mobile phones and field evaluations inflict.

Furthermore categorizing the usability problems in themes revealed that 43.5% of the total usability problems identified were related to *feedback* and *information*, thereby giving the designers an important clue to where the main part of the problems was.

58% of the problems identified were unique, but the more severe a problem was, the more likely it was to be identified in both evaluations. Overall the field-based evaluation was more successful in uncovering usability problems, as it identifies significantly more problems, and were the only type of evaluation, which identified problems related to *cognitive load* and *interaction style*. This implies that evaluations conducted in field settings can reveal problems not otherwise identified in laboratory evaluations.

The ISO usability assessment furthermore exhibits the importance of the context, as the field evaluation get a lower overall usability rating, which emphasize that the context of use has influence on the usability of a system as stated in ISO 9241-11 [1998]. The conclusion is that it is worthwhile conducting user-based usability evaluations in the field, as it identifies more usability problems and problems relating to themes not found in the laboratory evaluation.

The findings in this paper are subject to limitations concerning the realistic settings in the field-based evaluation. It can be argued that vital aspects of a realistic environment did not exist, such as weather conditions and transportation issues. This is not a constraint for the problems identified, but evaluating in a more realistic use context might have altered the number and theme of identified problems.

Based on the results of our experiment we find that it would be interesting to conduct a similar experiment, were both the laboratory and the field settings are more realistic. This would give us the possibility to discover if the results change, and to triangulate the results gained, with the results of Kjeldskov et al.[2004], who perform a similar experiment.

References

- Baillie, L. (2003). *Future Telecommunication: Exploring actual use*, In Proceedings of IFIP TC13 International Conference on Human-Computer Interaction – INTERACT '03, Zurich, Switzerland, IOS Press.
- Dix, A., Finlay, J., Abowd, G. & Beale, R. (1998). *Human-Computer Interaction*, Prentice Hall Europe, Second Edition.
- Esbjörnsson M., Juhlin O. & Östergren M. (2003). *Motorcyclists Using Hocman – Field Trials on Mobile Interaction*, In Proceedings of the 5th International Mobile HCI 2003 conference, Udine, Italy, Springer-Verlag, LNCS.
- Frøkjær, E., Hertzum, M. & Hornbæk, K. (2000). *Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated?* In Proceedings of the ACM CHI 2000

¹ A nationwide study from 1991 showed, that 12% of the adult Danish population have a reading disability
[<http://www.dvo.dk/ordblind/index.htm>]

- Conference on Human Factors in Computing Systems, ACM Press – preprint version.
- Gray, W. D. & Salzman, M. C. (1998). *Damaged merchandise? A review of experiments that compare usability evaluation methods*, Human-Computer Interaction, 13(3), 203-261.
- Hart, S. G., & Staveland, L. E. (1988). *Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research*. In P. A. Hancock & N. Meshkati (Eds.), Human mental workload, Elsevier Science Publishers, B.V.
- Hertzum, M. and Jacobsen, N.E. (2001). *The Evaluator Effect: A Chilling Fact about Usability Evaluation Methods*. International Journal of Human-Computer Interaction, 13(4), Lawrence Erlbaum Associates, Inc.
- ISO (1998). The international Organization for Standardization (1998), Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability (ISO 9241-11).
- Johnson P. (1998). *Usability and Mobility; Interactions on the move*, In Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices, Glasgow, Scotland, GIST Technical Report G98-1.
- Karat, C., Campbell, R. & Fiegel, T. (1992). *Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation*, In Proceedings of the SIGCHI conference on Human factors in computing systems 1992, Monterey, California, United States, ACM Digital Library.
- Kjeldskov J. & Stage J. (2004). *New Techniques for Usability Evaluation of Mobile Systems*, Accepted for publications in International Journal of Human-Computer Studies, 60, Elsevier (forthcoming 2004).
- Kjeldskov, J., Skov, M.B., Als, B.S. & Høegh, R.T. (2004). *Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field*. Accepted for publication in Proceedings of the 6th International Mobile HCI 2004 conference, Glasgow, Scotland. Lecture Notes in Computer Science, Berlin, Springer-Verlag
- Molich, R. (2000). *Brugervenlige EDB-Systemer*, Ingeniøren|Bøger, 2. edition, 2000.
- Molich, R., Ede, M.R., Kaasgaard, K. & Karyukin, B. (2004). *Comparative usability evaluation*, In Behaviour & Information Technology Journal, January-February 2004, vol. 23, no. 1.
- Newman, W.H. & Lamming, M.G. (1995). *Interactive System Design*, Addison-Wesley Pub Co, 1st edition, 1995.
- Nielsen, C.M., Overgaard, M., Pedersen, M.B. & Stenild, S. (2004). *The Development of a Mobile System for Communicating and Collaborating – An Object-Oriented HCI Approach*, Department of Computer Science, Aalborg University, 2004.
- Norman, D. (1990). *The Design of Everyday Things*, Doubleday and Company, 2002 Edition.
- Nyssönen, Roto & Kaikkonen (2002). *Mini-Camera for Usability Tests and Demonstration*, Presented in Demo Sessions at the 4th International Symposium on Human Computer Interaction with Mobile Devices, 2002, Nokia Research Center.
- Pedell, S., Graham C., Kjeldskov J. & Davies, J. (2003). *Mobile Evaluation: What the Data and the Metadata Told Us*.
- Pirhonen, A., Brewster, S. & Holguin, C. (2002). *Gestural and Audio Metaphors as a Means of Control for Mobile Devices*, In Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, 2002, Minneapolis, Minnesota, USA, ACM Digital Library.
- Rubin, Jeffrey (1994). *Handbook of usability Testing – how to plan, design, and conduct effective tests*, John Wiley & sons, Inc.

Providing Feedback to Designers: Are Usability Reports Any Good?

Christian Monrad Nielsen, Michael Overgaard, Michael Bach Pedersen & Sigge Stenild
Department of Computer Science, Aalborg University, Denmark
{monrad, mio, mbp, stardust}@cs.auc.dk

Abstract

This paper examines how to provide feedback to designers based on a usability evaluation through a case study experiment. The experiment relies principally on quantitative data collection through semi-structured interviews with two software developers from a small Danish software company. From the experiment a number of lessons learned, in relation to the content and structure of usability reports, are presented, which can help usability evaluators provide better feedback through usability reports. The findings indicate the importance of detailed descriptions of problems and that logs of both video and system interaction are used by the developers, when trying to understand the nature of the usability problems.

1. Introduction

In this paper, we explore usability reports as a mean for providing feedback to the designers on usability problems found in a user-based usability evaluation. The relation between these two central topics, which is depicted in Figure 1, needs to be defined.

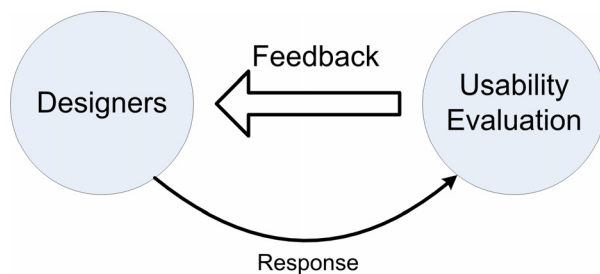


Figure 1: The process in focus in this paper.

Feedback, in this case, is defined as the return of information about the results of a process. It can also be viewed as an evaluative response based on a number of activities. The process and activities mentioned in this paper is the act of performing usability tests and analyzing the data from these.

Designers are a range of different individuals in a software development organization that are involved in the design of an information system. It covers a diverse group of people such as engineers, programmers, and graphical designers, which is anyone in a development

group, who, through their work and decision making, affects the usability of an information system.

Research in this area is important, because if the usability of a system should be improved, after a usability evaluation, it is important that the people, who are going to fix the problems, acknowledge and understand them. Usability and the related activities can eventually help designers to make better decisions; and thereby allowing them to do their jobs more effectively [Radle & Young, 2001].

1.1. Related Work

Much of the research that has been published in the area of usability feedback focuses on the organizational and interpersonal aspects of introducing usability evaluation in an existing organization. Some describe how usability engineers are best adopted and introduced into existing groups, consisting mainly of engineers and software developers [Mayhew, 1999a; 1999b]. While others attend to how organizational focus, on all levels of the organisation, can be directed towards usability [Ehrlich et al., 1994] [Radle & Young, 2001].

Nielsen [1994] focuses on quantitative measures of the physical characteristics of usability laboratories and calls for research concerning which placement usability groups should have in organizations. The schism is whether usability specialists should be centralized or distributed. If the organizational setup is based on usability specialists being a part of the development team, then there is little need for formal reports, because results are taken directly into the development process [Bærentsen & Slavensky, 1999].

Rohn [1994] portray a usability engineering group inside SunSoft, which provide support and performs usability evaluations *across* the organization. Several authors describe the use of specialized usability groups/departments employing usability professionals [Salzman & Rivers, 1994] [Lund, 1994] [Blatt et al., 1994] [Palmiter et al., 1994] [Fowler et al., 1994] [Zirkler & Ballman, 1994] [Muller & Czerwinski, 1999]. In a study of six different companies in the US and Denmark, Borgholm & Madsen [1999] find that there is a tendency toward separating designers and usability specialists in distinct organizational units. Despite differences between Scandinavia and the US

and regardless of organizational placement, usability specialists agree that the acceptance and credibility of their work cannot yet be taken for granted in many companies. An alternative, to being either centralized or distributed, is third-party vendors providing services to other companies [Dolan & Dumas, 1999].

A review of papers presenting usability evaluations of mobile systems showed that in all of the 58 papers examined, the designers and the evaluators were the same individuals [Nielsen et al. 2004a]. This can be related to why the problems with providing feedback to the designers have not been studied in much detail, because in research experiments it is often the designers themselves, who perform the usability evaluation.

When designers and evaluators are the same persons, it can have both advantages and disadvantages concerning the outcome of a usability evaluation. The advantage is that the evaluators are familiar with the application domain as well as the functionality and design of the system [Hartson et al., 2004]. On the other hand, the lack of independence between the designer and the evaluator might result in a less objective evaluation since the designer is biased towards the system [Bachrach & Newcomer, 2002].

1.2. Challenges in Relation to Designers

Receiving feedback that describes problems in a system, which the designers have personal involvement in, can be a discouraging task. When usability issues in a design are pointed out as being problematic, the designers will sometimes make an effort to defend the design, described as ‘design defensiveness’ [Spencer, 2000]. These problems are often caused by the lack of basic understanding about what usability really is [Mayhew, 1999a:414]. This prompts for the need to investigate the communicative mechanisms at play, when providing feedback to designers.

1.3. Scope

As it can be seen from the above presentation of research, focus is mostly on the governing organizational conditions. None, or at least very few, research papers focus on what mechanisms are in play when providing feedback to the designers of a system. The type of feedback examined in this paper is *usability reports*.

In addition to this, we want to investigate whether the designers interpret problems found in laboratory and field settings differently. This is relevant, since evaluation in field settings propose a number of practical problems [Johnson, 1998] [Kjeldskov & Stage, 2004] [Baillie, 2003]. Lately Kjeldskov et al. [2004] have also questioned whether evaluation in the field is worth the added effort compared to simulating a field-like experience in the laboratory.

The above discussion translates into a more general research question, which is: *How can effective feedback from a usability evaluation be provided to the designers of a system in a usability report?* The experiment we perform, which is described in section 3, focus principally on qualitative data.

In section 2 we elaborate on usability reports and chapter 3 describes details on the experimental design. Section 4 presents the result. Finally, section 5 discusses the findings and section 6 is the conclusion.

2. Usability Reports

The experiment described in this paper revolves around usability reports. In Dumas & Redish [1993], Rubin [1994] and Molich [2000] usability reports are suggested as a mean for communicating the results of a usability evaluation.

A study has shown that test reports are very common and standardized documents [Borgholm & Madsen, 1999]. Muller & Czerwinski [1999] also describe the use of reports within Microsoft to share findings and usability engineers’ recommendations, by making them available on the company intranet.

2.1. Structure and Content of Usability Reports

Sy [1994] and Redish et al. [2002] are some of the few, who presents specific advice on the structure and content of a usability report. The advices presented in Sy [1994] are:

- Include the goals of the test.
- Problems should be ordered according to how critical they are.
- Use bulleted lists, tables, and graphical presentation for quick retrieval of information.

Redish et al. [2002] and Perfetti [2003] mention that the report should:

- Not be too long and present a manageable number of problems.
- Include an executive summary.
- Include severity classifications.
- Include the number of users, who experienced the problem.
- Include positive findings.

The reports used in our experiment adapt many of the advices above.

2.2. The Reports Used in the Experiment

In this section we briefly describe the structure and content of the usability reports. A thorough description and comparison of the two usability evaluations can be found in Nielsen et al. [2004b]. Before the writing of each usability report, an agreement on the structure and content was made. The reports had the following structure:

1. **Summary**
2. **Method**
 - a. Purpose
 - b. Procedure
 - c. Test participants
 - d. Test procedure
 - e. Location & equipment
 - f. Identification & categorization of problems
3. **Results**
 - a. Workload (NASA-TLX)
 - b. Time used
 - c. Problem overview
 - d. Detailed description of problems
4. **Conclusion**
5. **Appendix**
 - a. Tasks
 - b. Interview guide
 - c. Questionnaires
 - d. Video log-files
 - e. System log-files
 - f. Task solutions

Apart from minor adjustments, this structure is based on Rubin's [1994:288-293] description on how to structure a usability report. The enumeration in this structure will be used as a reference later in the paper, where the developers' opinion on what parts of the report they found the most important is presented.

A number of characteristics associated with usability reports are that they are often very extensive, take a long time to produce, and involve a heavy workload for the author [Borgholm & Madsen, 1999]. With this being the case, it is paramount that the feedback designers receive from such reports are useful, otherwise producing the report would be a waste of resources.

3. Experimental Design

Below we describe the specific approach on how we examined the change in the developers' opinion on what the major challenges and advantages were, as they read and reviewed the two reports.

3.1. Developer Opinions

Table 1 shows the 5 steps of the experiment concerning the two developers. Their initial understanding of usability, usability evaluation, and expectations to the

usability reports were uncovered. Following this, they were interviewed about their initial opinion on strengths and weaknesses in the system (step 2). The task of describing and explaining strengths and weaknesses in the system was repeated after each of the reports had been read. Each time strengths and weaknesses had been identified, the developer was asked to rank them relative to each other.

Step	Designer A	Designer B
#1	Outline the process for the developers, without revealing details.	
#2	Semi-structured interview on initial opinions on advantages and disadvantages.	Semi-structured interview on initial opinions on advantages and disadvantages.
#3	Receive and read the laboratory usability report. Semi-structured interview based on step #2. Interview is conducted by one of the writers of the laboratory usability report.	Receive and read the field usability report. Semi-structured interview based on step #2. Interview is conducted by one of the writers of the field usability report.
#4	Receive and read the field usability report. Semi-structured interview based on step #3. The designer is asked to comment on the usefulness of the reports and the individual parts. Interview is conducted by one of the writers of the field usability report.	Receive and read the laboratory usability report. Semi-structured interview based on step #3. The designer is asked to comment on the usefulness of the reports and the individual parts. Interview is conducted by one of the writers of the laboratory usability report.
#5	Group discussion where the designers are presented with each others advantages and disadvantages. The two developers are asked to agree on a joint list of advantages and disadvantages.	

Table 1: The five steps in the experiment.

The laboratory and field reports were presented to the developers in opposite order (step 3 + 4), to see if the order of the two reports would influence how they perceived them. After reading the two reports and generating the lists of strengths and weaknesses, the final lists were compiled by the two developers and were written on a white-board without the ratings. They were then asked to discuss and finally agree on a rating for all of the items of the two lists (step 5).

The reason for having two different interviewers perform the interview (step 3 + 4) was to ensure that the one conducting the interview always had in depth knowledge about the evaluation and report in question.

Despite the ratings being important, it also served the purpose of forcing the developers to discuss and reflect on each item. This can ultimately help them, and us, to gain a better understanding of each item and the underlying reasons for why exactly these are the ones mentioned.

3.2. Semi-structured Interview

The approach used during the interview was a semi-structured interview, also known as a qualitative research interview [Kvale, 1997]. The interview guide used by the interviewer had the following overall structure:

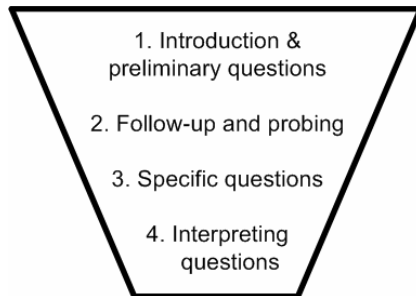


Figure 2: The funnel approach used during the interviews.

Figure 2 illustrates that the interviewer starts out with the most general questions in order to gain some initial knowledge concerning the developer. He then moves on to ask follow-up questions, which leads on to specific questions on specific topics. Based on the answers from the developer, the interviewer finishes off by asking questions, which indirectly aim at interpreting the statements made by the developer.

3.3. Analyzing the Interviews

In order for us to perform an analysis of the interviews, they were recorded on Mini-Disc. The analyses of the interviews were done in two steps. First the recordings were transcribed using opinion condensation. Then the resulting transcriptions were analyzed to uncover statements that can be associated with a number of themes, which relates to our research question.

3.3.1. Opinion Condensation

Three interviews with each of the developers and one joint interview were conducted. The length of the joint interview was 74 minutes, while the three interviews with each of the developers lasted 72 minutes and 101 minutes respectively.

To analyze the interviews, we used opinion condensation as described by Kvale [1997:186-206]. This was done two days after the interviews. Through this kind of transcription, opinions expressed by the interviewees were transformed into shorter and more precise formulations. The intention of the condensation is to be as precise as possible, which means that we use the same words as the interviewee did in the situation. Longer pieces of speech are condensed into a single or few sentences. The advantage of opinion condensation is that it can help present a relatively large amount of empirical data in an easy-to-read fashion, while both

preserving and clarifying important issues. Opinion condensation can never be considered equal to ‘traditional’ transcription of the interview, which has significantly higher level of detail and involves less processing of the original text (audio recordings). For example, opinion condensation implies that all non-verbal communication is lost during transcription.

3.4. Case Description

For the case to be as realistic as possible, we conducted the experiment in cooperation with developers from a real company, who are currently working on the development of a mobile system.

3.4.1. The Developers

During the experiment, we cooperated closely with two of the developers in a software company. These two developers were responsible for the design of the user interface in the system we evaluated. Developer A was 31 years old and had an education as data-engineer and developer B was 32 years old and was educated as datamatician¹. Both have been a part of the company since it started in 2000.

Apparently the company has recognized the benefits of focusing on usability. The following quote is taken from their web-site: *“To us the technical part of the solution is not sufficient in order to create a good solution, user friendliness is just as important.”* How, or to what degree, this affects the daily work of the company is unknown, but in relation to the challenges described in the introduction, it indicates that the company has recognized the competitive edges of usability.

3.4.2. The System

The system that the developers are working on is used for registering use of time, materials, mileage, and equipment and providing online access to the inventory, while working in the field. The system runs on a regular mobile phone with barcode scanner attached. The system relies heavily on the use of barcodes for performing registrations and interactions with the system. According to the company, the target user group is e.g. servicing engineers, home-helpers, carriers, crafts- and workmen.

3.4.3. The User-based Usability Evaluation

Two separate usability reports were written. One describing an evaluation performed in a state-of-the-art usability laboratory at Aalborg University and the other in field-settings at Vitus Bering CEU (technical high school) in Horsens. The evaluation method used was think-aloud and each evaluation involved seven differ-

¹ In Danish: “Datamatiker”.

ent users from Vitus Bering CEU. The tasks, which the users were asked to solve during the test, were devised in cooperation with a technical teacher from the same school.

3.4.4. Data Analysis and Problem Descriptions

The laboratory and field evaluations were each conducted by teams of two persons. The entire process of analyzing the data and writing the reports were done by the same team that conducted the test. The two teams were not allowed to discuss any results or findings before the entire process was completed. In advance, it was agreed, which type of severity ratings were to be used.

Table 2 shows the number of usability problems documented, described and rated according to severity in the reports. Severity ratings were based on three ratings proposed in Molich [2000].

	Field	Laboratory
Critical	15	14
Severe	16	14
Cosmetic	17	6
Total	48	34

Table 2: Number of usability problems found in both evaluations according to severity.

As illustrated in Table 2, the number of critical and severe problems found is almost identical in the two evaluations. A more thorough comparison of the results can be found in [Nielsen et al., 2004b].

4. Results

This section presents the key findings from the experiment. Sections in the usability reports are referred to by e.g. '(3a)', which relates to the section on workload in Table 2. References to steps, in the experiment, are referred to by '(step x)'. Quotes from the interview are in *italics*.

4.1. The Concept of Usability

Both developers described themselves as having comprehensive experience in designing user interfaces. They have both gained their experience through their jobs and not through their educational background. They find that usability and user interface design are important and necessary parts of their job, but that they cannot spend much time analyzing and considering different ideas when implementing parts of the user interface. Developer A explicitly says that they are software developers and engineers and that this is their main strength. They have never worked on a project, where usability evaluation was part of the process.

Generally, both developers were able, quite specifically, to formulate what they understand as usability.

Developer A finds that 'intuitive' is the word that describes it best, but he also mentions 'easy' and 'straightforward' to use, without having to read several manuals. Developer B defines usability as the specific screens. The design should target the user and the information presented should be relevant. Additionally, the user interface should be easily understood and nice to look at. They both describe the specific usability requirements, in relation to their system as; the interaction should involve a low number of scans and limited data entries.

4.2. Developers' View on System Advantages and Disadvantages

As it can be seen in the description of the experiment, both developers formulated three lists each and one joint list about usability advantages and disadvantages in the system. This section presents the findings directly related to these lists.

4.2.1. Developer A

Developer A had some difficulties in naming five advantages and disadvantages, especially in the beginning. He never succeeded in mentioning more than three advantages in the system. He was also somewhat reluctant in prioritizing the items in the lists. The only real change in advantages mentioned, following the initial list (step 2), was that he added that the system was highly adaptable. This advantage was rated as number 3 in the second list (step 3) and number 2 in the third list (step 4). In all the lists, developer A found the most important advantage to be that the system was online, and thereby has the ability to present real and accurate data to the user.

We have identified a number of changes in the three lists, focusing on disadvantages in the system from developer A. None of the three advantages he initially mentions are found in the succeeding two lists (step 3 + 4). The first three problems initially described were very general and abstract, but when presented with the very specific problems in the usability reports, he changed his focus significantly to match many of the problems found.

Basically the problems described in the two final lists (step 3 + 4) are the same, only one new problem is added in the last list, and two other problems were combined into one. The most important problem in the final list (step 4) is: 'Social/human resistance towards the introduction of the system'. Despite not being a specific usability problem, developer A mentions it as number 5, after having read the field-usability report, and after reading both reports, he rated it as the most important problem.

4.2.2. Developer B

We found that both ratings and the subjects mentioned in the positive list changed very little with developer B. A few changes did occur, where a subject changed one position in the list. The biggest change, in the advantages mentioned, was that the system was simple and required limited interaction. In step 2 he abandoned the view that it required limited interaction. In step 3 this subject appeared as the second most important advantage.

The fact that the barcode scanner cannot be used with Nokia mobile phones has proven to be a problem for the developers in relation to the users' attitude towards the system. This can be one of the reasons why this point was on the initial list, but disappeared as soon as he was presented with the system related usability problems, described in the first report.

At first (step 2) a general problem, limited screen size, was rated as the most important problem. This problem was degraded to number 4 in the following two lists (step 3 + 4), where barcode descriptions, user training, and response time were the top three problems. In these two lists, a problem emerged stating that it was problematic that different mobile phones interpreted the user interface code differently. The screen size problem was rated lower, because he saw that the user actually did manage to perform some of the more complex tasks, without encountering problems related to this issue.

4.2.3. Joint List

The advantage mentioned in the joint list, 'the system is simple and uniform', is interesting, since many of the usability problems found contradict this advantage.

The five problems in the joint list reveal much on how they intend to solve the problems. It indicates that they might not entirely have recognized the realness of many of the usability problems described, despite their previous statements. It is interesting that they, through the lists, indirectly reveal that usability issues should be resolved through user education and system documentation. This contradicts a statement from one of the developers saying, when the system is sold through their partners, then literally no education and introduction is given to the system. Hence making a system that can be used without comprehensive education is important.

Despite the problem of 'Social/human resistance' towards the system only being mentioned in developer A's list, they quickly agree that this is the most significant problem.

4.2.4. Comparison

Many of the advantages mentioned by both developers are somewhat general and somewhat sales oriented, which probably reflects the every day focus of the developers. Combined with the fact that the usability reports focus little on positive findings, this can explain why only slight changes can be observed in the individual lists showing advantages.

Several changes occur in respect to the initial lists from both developers. This shows that the developers do in fact change their immediate perception of usability problems in the system, as a result of reading the first report. After reading the second report (step 4), developer B's list did not change at all, while developer A changes his mind concerning several issues.

4.3. Usefulness of the Reports

Both developers used the same approach when reading the reports. Basically the reports were read from the beginning to the end. Occasionally the appendices (5) were used to see the design of the tasks. The log-files (5d + 5e) were not read in their entirety, but were used to examine details concerning a problem, if they were uncertain why a problem had occurred. Developer B said: *'I used the log-files to gain further insight into what happened'*.

Both developer A and B mentioned that the overview of the usability problems (3c) and the elaborating descriptions (3d) were important in the future work on the system: *'I really like the problem list and it is something I can use concretely in my work'*. The log-files (5d) were good, because *'they describe what they (the test participants) did. It provided a better feel of what they did, why they could not figure it out, and what they did next'*. This shows that log-files are useful, for providing further insight when trying to understand some of the problems in detail.

Log-files can provide almost firsthand insight into what specific actions the user performed. Although they cannot be used directly to resolve the problems, they find them important to understand the conditions under which the tests have been conducted (2). This was mentioned by both developers as being very important in respect to how they rate the validity of the evaluation. On the contrary, developer B mentions that: *'The other assessments and similar are quite fun to read, but they are not very useful'*, referring to the summary (1) and the conclusion (4). It is important to note that executive summaries may still be important in a more general organizational context.

The developers found the NASA-TLX (3a) method interesting, but they experienced some problems in interpreting the table displaying the NASA-TLX results. Developer B found that the field report lacked a transcription of the debriefing conducted at the end of

each test. This was important, since: *'It would provide me with a better insight into the participants' attitude towards the system'*.

During the final interview (step 5), the developers brought up the issue of using video recordings. In relation to some of the problems encountered in: *'the first few minutes, when the user for the first time was presented with the mobile phone'*, it would have been beneficial if the video material had been available. This would have given him a chance to see the test participants' first reactions.

When asked, which of the two reports, they found to be the best, they both replied that it was the one they read as the first one. Developer B said that the laboratory-based report was the best because it included transcriptions of the debriefing following each test were available in the log (5d). He also mentions that: *'it reflects the reality I know best'* and that the field report appears more *'critical'*. The perception that the field report is the most critical may be due to it describing more problems. Developer A found the field based report to be the best, as he found that it was more detailed in its descriptions of the problems (3d).

4.4. Social and Organizational Aspects

Limited time is an overall issue throughout the interviews and in several occasions the two developers use this as an excuse for some of the existing usability problems. In the beginning, before having seen any of the reports (step 2), developer A says: *'We know that many of the things are there – many things that we would really like to correct if we had the time'*. Numerous times both developers mention that designing the user interface is an important and necessary part of their job, but that they cannot spend much time on analyzing and considering different ideas. They are simply too busy and therefore have not got the necessary time. Developer A expresses that this should be taken into account when evaluating the usability.

As developers, they often find themselves thinking in *'states'* and *'actions'*, but according to both developers, the reports can help them to gain further insight into how the users think, when they use the system.

4.5. Evaluation Setup

One of the issues frequently referred to during the interviews is that the users are very inexperienced, and if they were more experienced, the result of the evaluation would have been different. This is probably correct, but it does not imply that usability problems found by relatively inexperienced users do not exist. We see this more as a defensive reaction towards, a perhaps, overwhelming number of usability problems. This is supported by the developers accepting that many of the problems are relevant and should be fixed. When de-

veloper A is asked about his general opinion on the evaluations, he replies that: *'Many of the things mentioned have applicability in our further work'* and adds that he: *'can relate to the findings and use them positively'*.

Another point of critique presented by the developers is that the tasks are not realistic and that this might have affected the outcome of the tests. According to Molich [2004] this is a typical objection raised by developers. Still developer A mentions: *'I am impressed with how many strange errors the users manage to provoke, which we have never thought of ourselves'*.

5. Discussion

Before reading any of the usability reports developer B expressed that he had great expectations to the usability evaluation. However, he was somewhat worried that we might not have enough experience in using the system, whereas a potential user has a need for doing the things that the system can. He thinks the evaluation might have another outcome, if we did the evaluation one more time. We find this as an example of the developer being defensive [Spencer, 2000].

Developer A expresses the time pressure, they work with every day, should be taken into account when evaluating the usability of the system, which in relation to usability evaluation makes little sense. Usability problems exist in the system regardless of the time that has been available for development. The issue of designers being reluctant to add time in their schedule for HCI activities has been described before, for example in Radle & Young [2001] and Spencer [2000].

Whether to include and edited videotape as part of the feedback for the designers, to allow designers a first hand view of the problem, have been used at IBM [Fath et al., 1994]. Traditional reports are still used, but video clips can *'provide compelling evidence to developers who are reluctant to correct usability problems'* [ibid.]. A drawback associated with the use of video recordings is that it is very time-consuming task to edit such a tape [Borgholm & Madsen, 1999].

From a discussion about issues on confidentiality, originating from the fact that no formal contract existed, another idea emerged. Namely that the developers were allowed to write a few pages, which could then be put in the report, where they could explain the improvements, they will be implementing based on the reports. This approach can also be relevant in other situations, for example in larger companies, where a development group might not be satisfied that a report existed, within the organisation, describing usability problems in a piece of software they developed. By writing a few pages, describing how they used the results of the report to improve the usability of the

software, the developers may be less sceptical towards the existence of a usability report.

When suggesting solutions to the usability problems, they agree that some problems can be solved through minor modifications of the user interface. Despite having, at least on the surface, accepted the existence of usability problems, they rate more user education and making a user manual as being some of the most important strategies for solving the problems.

Very few positive findings were presented in the two reports. When the developers were asked, whether they would have liked the evaluation to focus more on positive findings, they replied that positive findings are always nice, but they cannot really use them for improving the system. Hence they do not find any reason for spending a lot of time and energy on finding positive aspects. Both Perfetti [2003] and Redish et al. [2002] support the idea of including positive findings in usability reports.

Alternative approaches for providing feedback have been proposed. In Redish et al. [2002] cooperation with the designers are taken a step further by suggesting that they are brought in and made a part of the planning and conducting of the evaluation, analysis of the data, and the communication of the results.

Spool [2004] and Redish et al. [2002] describe the KJ-method, which is used for group decision making. Notes taken by developers, who are observing the usability tests, form the basis of applying the KJ-method. Through a series of steps, the developers agree on a list of the most important problems. The advantages of this approach are that the results are immediately available, and the development team have themselves played an important role in defining the usability test results. This can potentially make them more receptive of the usability problems found.

Radle & Young [2001] also recognize the importance of interpersonal skills when addressing usability in relation to development teams. Sy [1994] presents additional advises on how communication of the evaluation results can be improved apart from a usability report. If possible, a meeting should be held to go through the findings with the appropriate people. During this meeting, it is important to refrain from any kind of confrontational attitude, and if possible, the meeting should be ended with a list of actions derived from a co-operative discussion. This is important in order to involve designers more actively in the resolution of usability problems.

6. Conclusion

Through the experiment described above we have learned a number of lessons, which are relevant in

providing effective feedback to designers through usability reports. This leads us to conclude that:

- A problem list providing overview of usability problems combined with detailed descriptions is important and essential for the designers when trying to understand a problem.
- Results of NASA-TLX, which are not explained by being put into context, are difficult for the designers to relate to.
- Log-files of user interaction, based on video recordings combined with system-logs, are used and considered important by the designers to understand specific details of the usability problems.
- Information on test setup, users, tasks, and test users' subjective opinions are important to the designers, but these are also the point of critique, when designers explain, why they find problems more or less real.
- General assessments and evaluations in usability reports have limited usefulness for the designers.

Contradictory to our expectations, the designers mentioned user training and writing a manual as a way to overcome some of the usability problems instead of making changes to the user interface.

In the interviews with the designers, we experienced situations similar to design defensiveness as described by Spencer [2000]. We have also experienced specific critique in relation to users and tasks, as described by Molich [2004]. The issues of making time for usability and HCI related work [Radle & Young, 2001] [Spencer, 2000] were also found during the experiments with the two developers.

In this type of experimental design, we rely heavily on qualitative data collected through interviews. All of the results we have found are subject to the proviso that the experiment only involved two developers, which implies that results cannot be considered general.

Based on the results of our experiment, we find that it would be interesting to perform similar feedback experiments with other ways of providing feedback to designers. Inspiration can be found in the area of interpersonal communication, for example, the ideas of the American psychologist and psychotherapist Carl Rogers, who have defined the psychological conditions necessary for open and fulfilling communication between individuals [Rogers & Freiberg, 1994] [Rogers, 1962]. In Rogers' opinion what is needed is the role of a facilitator.

Further research, exploring whether the whole process has resulted in actual changes that improve the overall

usability of the system, is relevant to conduct, in order to determine, whether the final system is considered more useful by the potential users. Which in the end is what usability evaluation is all about.

References

- Bachrach, C. & Newcomer, S. F. (2002). *Addressing Bias in Intervention Research*, in Journal of Adolescent Health, Volume 31, Number 4, 2002.
- Bærentsen, K. B. & Slavensky, H. (1999). *A Contribution to the Design Process*. Communications of the ACM, May 1999, Vol. 42, No. 5.
- Baillie, L. (2003). *Future Telecommunication: Exploring actual use*, INTERACT 2003.
- Blatt, L., Jacobsen, M. & Miller, S. (1994). *Designing and equipping a usability laboratory*. In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Borgholm, T. & Madsen, K. H. (1999). *Cooperative Usability Practices*. Communications of the ACM, May 1999, Vol. 42, No. 5.
- Dolan, W. R. & Dumas, J. S. (1999). *A Flexible Approach to Third-Party Usability*. Communications of the ACM, May 1999, Vol. 42, No. 5.
- Dumas, J. S. & Redish, J. C. (1993). *A practical guide to usability testing*, Norwood, NJ: Ablex Publishing.
- Ehrlich, K., Beth, M. B. & Pernice, K. (1994). *Getting the Whole Team into Usability Testing*, IEEE Interface – January 1994.
- Fath, J. L., Teresa, L. M. & Holzman, T. G. (1994). *A practical guide to using software usability labs: lessons learned at IBM*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Fowler, C., Stuart, J., Lo, T. & Tate, M. (1994). *Using the usability laboratory: BT’s experiences*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Hartson, H. R., Shivakumar, P. & Pérez-Quñones, M. A. (2004). *Usability Inspection of Digital Libraries: A Case Study*, accepted for publication in the Special Issue of Journal of Digital Libraries on Usability
- Johnson, P. (1998). *Usability and Mobility; Interactions on the move*, In Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices, Glasgow, Scotland, GIST Technical Report G98-1.
- Kjeldskov, J. & Stage, J. (2004). *New Techniques for Usability Evaluation of Mobile Systems*, accepted for publications in International Journal of Human-Computer Studies, Elsevier (forthcoming 2004).
- Kjeldskov, J., Skov, M. B., Als, B. S. & Høegh, R. T. (2004). *Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field*, accepted for publication in Proceedings of the 6th International Mobile HCI 2004 conference, Glasgow, Scotland.
- Kvale, S. (1997). *Interview – En Introduction til det Kvalitative Forskningsinterview*, Hans Reitzels Forlag, 1. udgave, 1997.
- Lund, A. M. (1994). *Ameritech’s usability laboratory: from prototype to final design*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Mayhew, D. J. (1999a). *The Usability Engineering Lifecycle*, Morgan Kaufmann Publishers Inc. San Francisco, California.
- Mayhew, D. J. (1999b). *Strategic Development of the Usability Engineering Function*, ACM Transactions – September /October 1999.
- Molich, R. (2000). *Brugervenlige edb-systemer*, Teknisk Forlag.
- Molich, R. (2004). E-mail correspondence with Rolf Molich 18/05/04.
- Muller, M. J. & Czerwinski, M. (1999). *Organizing Usability Work To Fit the Full Product Range*. Communications of the ACM, May 1999, Vol. 42, No. 5.
- Nielsen, J. (1994). *Usability Laboratories*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Nielsen, M. C., Overgaard, M., Pedersen, M. B. & Stenild, S. (2004a). *A Review of Literature on Usability Evaluation Methods for Mobile Systems*, Department of Computer Science, Aalborg University, 2004.
- Nielsen, M. C., Overgaard, M., Pedersen, M. B. & Stenild, S. (2004b). *Usability Evaluation of a mobile system: Comparison of a Laboratory and Field Evaluation*, Department of Computer Science, Aalborg University, 2004.
- Palmiter, S., Lynch, G., Lewis, S. & Stempski, M. (1994). *Breaking away from the conventional ‘usability lab’: the Customer-Centered Design Group at Tektronix, Inc.*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Perfetti, C. (2003). *Usability Testing Best Practices: An Interview with Rolf Molich*. Originally published: 07/24/2003, Found 10/05/04: http://www.uie.com/articles/molich_interview/.
- Radle, K. & Young, S. (2001). *Partnering Usability with Development: How Three Organizations Succeeded*, IEEE Software – January/February 2001.
- Redish, J., Bias, R. G., Bailey, R., Molich, R., Dumas, J. & Spool, J. M. (2002). *Usability in Practice: Formative Usability Evaluations – Evolution and Revolution*. Usability in Practice Session, CHI 2002.
- Rogers, C. R. & Freiberg, H. J. (1994). *Freedom to Learn*, (Third Edition), MacMillan Coll Div.
- Rogers, C. R. (1962). *The Interpersonal Relationship: The Core of Guidance*, Harvard Educational Review, Volume Thirty Two, Number 4, Fall 1962.
- Rohn, A. J. (1994). *The usability engineering laboratories at Sun Microsystems*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests*, New York, NY: John Wiley & Sons.
- Slazman, M. C. & Rivers, S. D. (1994). *Smoke and mirrors: setting the stage for a successful usability test*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’
- Spencer, R. (2000). *The Streamlined Cognitive Walkthrough Method, Working Around Social Constraints Encountered in a Software Development Company*, CHI Letter 2000 Volume 2 Issue 1.

Spool, J. M. (2004). *The KJ-Technique: A Group Process for Establishing Priorities*. Originally published: 05/11/2004. Found 10/05/04: https://www.uie.com/articles/kj_technique/.

Sy, D. (1994). *Bridging the Communication Gap in the Workplace With Usability Engineering*, ACM 1994.

Zirkler, D. & Ballman, D. R. (1994). *Usability testing in a competitive market: lessons learned*, In BIT Volume 13, Numbers 1 & 2, January –April 1994, Special Issue ‘Usability Laboratories’