Testing, testing!

A Comparative Study of Usability Evaluation Methods for Mobile Systems

Andreas Bager & Lars Michael Hansen

Preface

This thesis was handed in on the 6th of June 2007 and is the result of the work conducted at the 10th semester of Informatics at Aalborg University.

With the work documented in this thesis we hope to contribute to the research in the field of Mobile HCI, by outlining how and why usability of mobile systems has been evaluated in the beginning of the new millennium. And by pointing out reasons why contextual evaluation methods should be considered when evaluating the usability of mobile systems. Furthermore by exemplifying how novel contextual methods can be devised we hope to inspire researchers to continue to focus on this area and entice practitioners to consider the potential of contextual evaluations when designing and developing mobile devices.

A thorough summary can be found on the back of the thesis and the video recordings from the usability evaluations can be found on the enclosed DVD.

We extend our appreciations for commendable counseling to Jesper Kjeldskov for his ability to exercise wise judgment and offer keen insight in the field of mobile HCI.

During the process a range of individuals offered their assistance and participated as usability expert and test subject. We would therefore like to thank the following for their willingly and open-minded participation: Jakob, Esben, Thomas, Claus, Julie, Troels, Camilla, Helle, Lars, Mads, Klaus, Peter, Søren, Christian, Kenneth, Nils, Mads, Anders, Caroline, Anne, Kamilla, Dennis, Katrine, Allan, Nanna, Niels, Signe, Peter, Kim and Jacob.

Andreas Bager
abager@cs.aau.dk

Lars Michael Hansen
larsmiha@cs.aau.dk

Content

Introduction

In the world of today, there seems to be a growing demand for new and advanced mobile systems that allow a higher degree of freedom and flexibility, in the professional as well as the personal sphere. The workforce is supposed to be flexible and the demand for mobile systems that allow us to work anytime anywhere is in high demand. Likewise, mobile devices that support our everyday life activities such as communicating with friends and family have become a necessity. Maybe the greatest indication of this trend is the widespread popularity that the mobile phone has gained during the last decade.

Nowadays mobile phones, smart phones and PDAs are commonly used as a substitute or extension of the desktop systems that we continuously use in everyday life. In fact as the computing power in the mobile devices increase the difference between desktop systems and mobile system are slowly diminishing. Even so there are still some differences worth considering, such as the mobility of the system, which allows you to take the mobile device anywhere you fancy and use it anytime you see fit, as opposed to the workstation that sits on your desk. It seems that the rise of the mobile technology industry is nourished by the pursuit of freedom and flexibility of modern life.

As the market for mobile devices is growing, so is the interest in producing usable systems that provide actual value to people. It is no longer enough to be pioneers in the field of an emerging technology and as Lindholm, Keinonen and Kiljander state in the book *Mobile Usability: How Nokia changed the face of the mobile phone*:

"It is not necessary to demand that the everyday needs of nursery school teachers and ironworkers be considered in the early development phases of low-power radio frequency transmissions. However it is vital that user needs meet technologies at some point in the development cycle, or they will die. The technologies, that is". (Lindholm, Kiljand and Keinonen 2003, p. 2)

Therefore, as the new technologies become part of everyday life, the need for ensuring good human computer interaction (HCI) is growing. But how do the professionals, that make a living of developing these systems, ensure that they meet the right demands in the right way when new types of technologies and systems emerge continuously? An important influence on the interaction between user and system is the usability of the system. Typically the usability of a system is ensured by conducting usability evaluations in order to identify and correct vital usability problems during development. However many of the techniques and methods that we apply in this process are founded in experiences from the desktop paradigm (Lindholm, Kiljand and Keinonen 2003, p. 3), hence the application of these methods in combination with mobile systems must pose a challenge to usability engineers.

It is this problematic aspect of mobile HCI that we have taken interest in, and thus it will be the focus of this thesis. However, before we set out on our exploration in the field of mobile usability, we will establish a general understanding of the two primary concepts in this thesis: *mobility* and *usability*.

Mobility

We start off by narrowing the notion of mobility from the broadest possible definition down to the mere characteristics of mobile systems. And as we see it, mobile systems can be defined in different ways. One way of defining them is by their technical attributes (Gorlenko and Merrick 2003) and another way is to define them by the way in which they are used (Kristoffersen and Ljungberg 1998).

By defining systems in terms of their technical attributes, such as weight, size and battery life, we are able to define some systems as suitable for mobile work and others as primarily desktop tools. Gorlenko and Merrick use this technical approach to define a spectrum of device mobility, which can be used to assess the mobility of a given device. For instance a laptop would be assessed as slightly more mobile than a desktop computer, because it is easier to unplug and bring along. However the construction of the device requires it to be placed on a flat surface before proper use can take place, and as such it is not as mobile as for instance a mobile phone, which can be operated by one hand while moving and thus is considered a fully mobile system. This approach to define mobile systems is limited to the device itself, and thus we look to Kristoffersen and Ljungberg, in order to expand the definition.

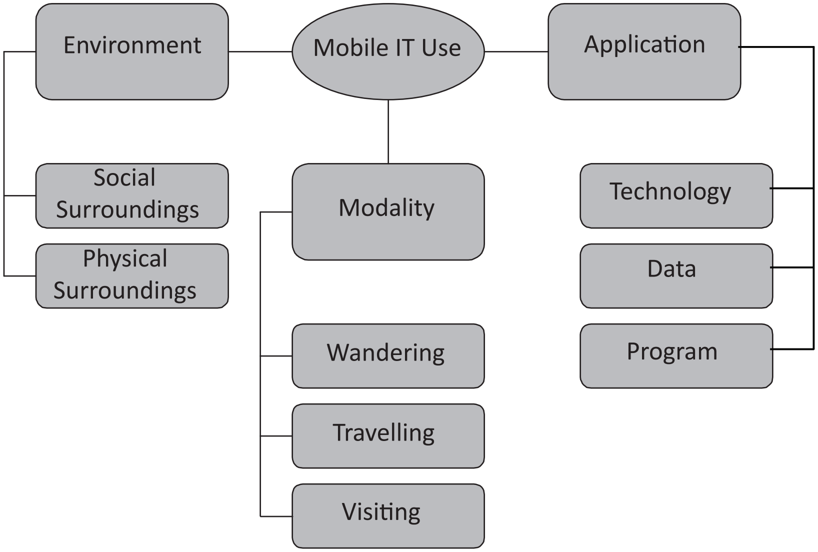

Figure 1 Model of the factors of influence on use of mobile IT

Kristoffersen and Ljungberg acknowledge that the technical device and its attributes is in large part responsible for the mobility of the system, but they introduce other parameters of influence and combine these in a model that serves to illustrate the aspects that encompass and define mobile use (See Figure 1 (Kristoffersen and Ljungberg 1998)).

This model illustrates how the human use of mobile systems is influenced by factors, such as the social and physical context of use, the modality of use and the hardware and software that constitute the system.

Kristoffersen and Ljungberg use the term *modality* to describe the patterns of mobility that the user is subject to in a given use situation. For instance, the mobile

use situation *travelling* should be perceived as a situation in which the user is located in some form of vehicle, which affects the use. The mobile use situation *wandering* covers situations where a user as a part of his work routines is required to move around in the office. And the mobile use situation *visiting* covers periods of activity outside the normal use situation, for instance when an office worker ventures into the field for a brief period. The different mobile use situations influence the usage as well as the environment in which the system is used.
The term *environment* can be divided into the *social* and *physical* environment, which both plays an important role in the use of mobile systems, since they can impose limitations on the use. Take for instance a mobile phone, which in many cases is a well functioning tool, but if we were to use it for communicating while driving a motorcycle, we would almost certainly put ourselves and the surrounding public at risk. Likewise the use of mobile phones is considered rude in many social settings, for instance at funerals or similar solemn occasions, where social standards and norms regulate the use of such systems.

The perception of mobile use as depicted in the model, encompass the application of the system as an important factor, but furthermore introduces modality and environment as factors of influence. Thus if future mobile systems are to meet the requirements of the users in the right way, these factors must be considered when designing and evaluating the usability of such systems. After all the usability of a system can only be evaluated with at least a minor understanding of the use. However, in order to understand the problems related to evaluating the usability of mobile systems, we first need to define usability and the purpose of evaluating it.

Usability

Usability evaluations are made either in order to identify problems and improve the system or in order to measure and assess the overall usability of the end system, i.e. formative and summative evaluations. We will focus on formative evaluations in this thesis because our primary interest is the challenges evaluators meet when conducting usability evaluations of mobile systems with the intention of identifying usability problems in order to inform the redesign of the system..

As we see it usability is a complex concept with many interpretations; however we believe that it should be seen as an important attribute of a mobile system, when used in a certain context. This interpretation complies with the definition given in the documentation of the ISO standard, ISO 9241-11:

"Usability: 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." (Bevan 2001)

A key aspect of this definition is the reference to context, which reflects the factors that characterizes mobile use presented in the previous section. If the use of mobile systems is affected by contextual matters such as, environment and modality, then the usability of the systems must also depend on these aspects. Thus we perceive the use and usability of a system as intertwined and thus they should be evaluated in the light of the context.

The definition of usability provided by the ISO standard closely resembles the one proposed by Jakob Nielsen (Nielsen 1994), where usability is defined through the terms: easy to use, efficient to use, easy to remember, having few errors and subjectively pleasing to use. However this definition does not encompass the contextual influences that have no direct effect on the interaction, which for instance means that social context, is not an aspect that influences the usability of a system, unless it affects the aforementioned subjective perception of the use situation. We believe that this definition of usability is too narrow, if we consider the fact that the quality of use of a mobile system is highly dependent on the context of use. Another view on usability is presented by Nigel Bevan, who criticizes Nielsen's definition of focusing on user interaction with the system as a closed system, instead of taking a holistic perspective. Bevan proposes that usability should be defined as quality of use (Bevan 1995). This broader definition allows us to view social acceptability and utility as aspects of usability, which to us makes sense, because it means that a system with high usability is a system that provides quality of use in a given context. This link between usability and quality of use is not included in Nielsen's definition, which means that according to his definition a system with high usability is not necessarily a system that provides quality of use.

In order to evaluate usability as we have defined it, we need to consider the environment and the mobile use situation, and in some way include this knowledge in our evaluation method. However it seems that the existing usability evaluation methods might not be suitable for this purpose. Take for instance the well known usability test conducted in a usability laboratory. By applying this method to evaluate a mobile system we wonder is we can ensure that the environment and the mobile use situation that affects the use in the test setup are in compliance with the actual context of use? Likewise it seems unlikely that experts conducting a heuristic inspection are able to take these factors into consideration. And how can we trust that the problems they identify will pose actual usability problems when used in the proper context of use? These questions are not easily answered, and even though a lot of effort has gone into research in this area, researchers still disagree whether or not it is worth the hassle to venture into the field with new and alternative methods.

Thus in this thesis we intend to examine *how* and *where* mobile systems should be evaluated. In order to do so we have defined three research questions.

Research Questions

The first step in examining how and where usability of mobile systems should be evaluated is to examine which usability evaluation methods are being used in the field of mobile HCI and how researchers deal with the challenges of evaluating mobile systems. This leads to the first research question:

> Research question #1: *Which methods are used in the field of HCI to evaluate the usability of mobile systems?*

Secondly we want to gain insight into how the methods, that we have established as the currently most used methods, perform when evaluating a mobile system and which trade-offs and practical obstacles there can be identified when they are applied. This leads to the second research question:

> Research question #2: *What characterizes the application and outcome of evaluating the usability of a mobile system with the current methods and how do the characteristics of the contextual and non-contextual methods differ?*

Finally we want to apply the knowledge that we gain from answering the previous questions to answer how and where usability of mobile systems should be evaluated, which leads to the third research question:

> Research question #3: *How can we utilize the first-hand knowledge that we have gained and devise a method that overcomes the challenges imposed by the nature of mobile systems?*

By answering these three questions we are able to contribute to the research in new methods for evaluating the usability of mobile systems. We are also able to offer an understanding of current practices and an insight into the problems researchers and practitioners face when evaluating the usability of mobile systems.

Research Methodology

In this chapter we account for the methodological procedure that we apply in order to approach the research questions presented in the previous chapter, along with the practical procedure that will be applied. This chapter will also present an overview of the content of the following chapters.

Research Question #1

In order to answer the first research question, we need to gain insight into the methods currently used for evaluating the usability of mobile systems. To learn how other researchers evaluate the usability of mobile systems, we might argue that direct observations of actual evaluations would be the best way to understand why mobile systems are evaluated the way they are. But since it would be difficult to describe what the typical approaches are when merely observing the available evaluation sessions and without the possibility of observing evaluations that already took place, we choose a descriptive approach that does not involve direct observation, namely a survey.

The survey will in this case take on the form of a literature review, from which we intend to collect descriptive data of the current practices in the research community. Thus we intend to do a survey on published research papers describing usability evaluation of mobile systems, which will allow a large sample size. Furthermore the knowledge that can be derived from several years of research will represent the tendencies of the period. The purpose of the literature review is to describe and understand past and present research by examining and classifying relevant research papers and hereby achieving an overview of practices and tendencies. The purpose of the literature review in this case would be to establish the proper knowledge base needed to answer the first research question.

Research Question #2

To answer the second research question we will need to obtain in-depth knowledge that allows us to compare and assess the methods that are currently applied when evaluating the usability of mobile systems. To obtain this kind of data we will conduct usability evaluations with the various methods that were described during the literature review, on an existing mobile system

We wish to evaluate the methods on their own terms and with respect to existing guidelines and recommendations, in order to apply the methods on the same terms as they would in a normal use situation. A case study therefore seems appropriate, since it allows us to gain detailed insight into the application of each evaluation method, when applied to evaluate a mobile system. In order to characterize the results and application of each method when applied to the same system we will adopt a holistic perspective and examine the results qualitatively.

However, as we conduct the evaluations and handle the analysis personally the method closely resembles applied research. By taking active part in the process instead of observing from a distance we are able to gain first-hand knowledge of the practical aspects of the evaluations, which we believe is necessary in order to answer the research question. By applying this mixed approach we will be able to characterize the methods and elaborate on their differences.

Research Question #3

By drawing on the experiences and results that we have obtained while answering the first two research questions, we will approach the third research question by constructing a new method that should attend to the problems related to usability evaluation of mobile systems. The purpose of constructing a new method is to test whether or not a more suitable method can be constructed and whether or not this should be the way to evaluate the usability of mobile systems.

Thus our methodological construction will be based on our own first-hand experiences as well as prior research, which implies that the research method in this part of the thesis is applied research. One of the characteristics of applied research, according to Wynekoop and Conger, is that we utilize prior knowledge, intuition, deduction and induction to analyze a specific research question (Wynekoop og Conger 1990). And when using applied research we should be aware that one of the major disadvantages of the method is that it often leads to results with poor generalizability. Thus we chose to apply this research method acknowledging the risk of losing the ability to apply our conclusions to mobile systems in general.

Methological Overview

The methods that we apply in order to be able to answer the research questions constitute are depicted in Table 1.

Research question	Research method	Procedure	Research objective
Which methods are used in the field of HCI to evaluate the usability of mobile systems?	Survey Research	Literature review	Understand and describe the methods that are currently applied in order to evaluate the usability of mobile systems
What characterizes the use and outcome of evaluating the usability of a mobile system with the current methods and how do the characteristics of the contextual and non-contextual methods differ?	Case Study / Applied Research	Conduct evaluation samples and compare them	Understand and evaluate the methods that are currently applied
How can we utilize the first-hand knowledge that we have gained and devise a method that overcomes the challenges imposed by the nature of mobile systems?	Applied Research	Construct a method and conduct a sample evaluation	Engineer and evaluate a novel method that deals with the challenges involved when evaluating mobile systems

Table 1 The research methods that we apply in order to answer the three research questions in the thesis.

The table illustrates which research methods we will apply to answer the research questions, and furthermore includes the procedure and objective of our research as described above.

In the process of answering the second and third research question we intent to apply secondary research methods in order to gather empirical data on the performance and application of different evaluation techniques. For instance, we will be conducting laboratory and field experiments in order to evaluate the mobile system with different usability evaluation methods.

Readers Guide

In extension to the above-mentioned account for our methodological procedure a short overview of how the thesis will be structured is appropriate. The structure of the thesis is illustrated in Figure 2.

A chapter will be dedicated to examining and answering each research question where the first chapter documents the literature review and answers the first research question. The following chapter documents the application of the different usability evaluation methods and their comparisons in order to answer the second research question. The following chapter documents how a novel usability evaluation method can be put together and evaluated in order to answer the third question. In the following discussion we will present recommendations on how to evaluate mobile systems under specific circumstances and elaborate on the potential value of conducting usability evaluations in context. In the following chapter we will return to the research questions and present the conclusions that we are able to draw. Finally we will account for the limitations of our work and propose suggestions for further work.

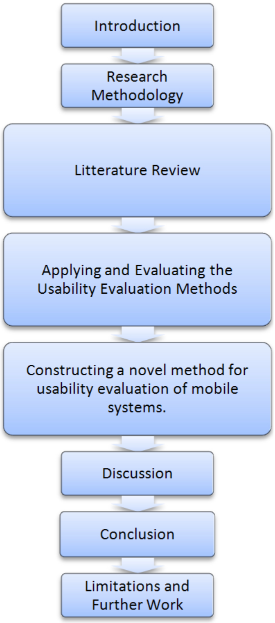
Introduction
Research Methodology
Litterature Review
Applying and Evaluating the Usability Evaluation Methods
Constructing a novel method for usability evaluation of mobile systems.
Discussion
Conclusion
Limitations and Further Work

Figure 2 The structure of this thesis.

Literature Review

This chapter will document the method that we applied in order to conduct the literature review and our findings will be presented.

The purpose of this review is to gain insight into how and why mobile systems are evaluated the way they are - by which we seek to shed some light upon the current practices of HCI researchers. By performing this survey we are able to answer the first research question that we set up in the previous chapter.

> Research question #1: *Which methods are used in the field of HCI to evaluate the usability of mobile systems?*

Furthermore we wish to understand how these researchers handle the challenges that the mobility of the systems impose on the evaluation process.

Method

The primary objective of this review is to enable us to characterize the usability evaluation methods that are currently applied in order to evaluate the usability of mobile systems. Thus we set out to identify and select a number of outlets that are relevant to the field of mobile HCI. This process resulted in the following list of outlets, limited to the papers published in the period 2000 to 2006. Most outlets are annual but some are bi-annual (marked *):

- Conference on Human Computer interaction with Mobile Devices and Services, Mobile HCI[1]
- Conference on Human Factors in Computing Systems, CHI
- Conference on Mobile and Ubiquitous Multimedia, MUM[2]
- Symposium on User Interface Software and Technology, UIST
- International Journal of Human-Computer Studies, IJHCS
- Symposium on Designing Interactive Systems, DIS*
- Transactions on Computer-Human Interaction, TOCHI
- Proceedings of the working conference on Advanced visual interfaces, AVI*[3]
- Conference on Computer Supported Cooperative Work, CSCW*
- Journal of Usability Studies, JUS[4]

This resulted in a massive set of papers (2360 in total), which were to be processed in a systematic way in order to subtract the set of papers that dealt with usability evaluations of mobile systems.

In order to select the candidate papers for further reading we went through three phases where papers were selected or dismissed depending on whether or not they fulfilled certain selection criteria. The phases are pictured in Figure 3.

[1] No conference was held in 2000

[2] Started in 2002 (but the proceedings from this year is not available)

[3] Proceedings from 2002 unavailable

[4] Start year: 2005

During the first phase we reviewed the abstracts of all the papers in each outlet and selected those that dealt with either development of a mobile system (with the potential of a following usability evaluation) or usability evaluation of a mobile system. Hereby the scope was limited to only 432 papers.

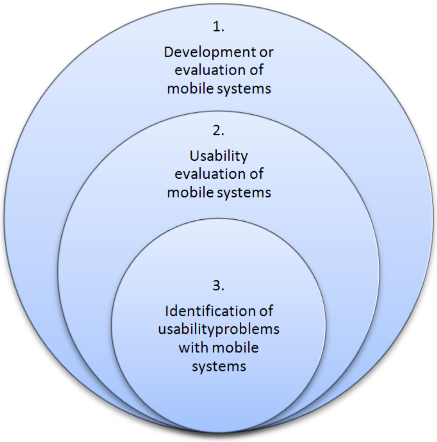
1.
Development or evaluation of mobile systems
2.
Usability evaluation of mobile systems
3.
Identification of usabilityproblems with mobile systems

Figure 3 The selection process of the literature review.

Next we searched through the resulting set of selected papers from the first phase, looking for those that did in fact concern usability evaluation of mobile systems. Due to our experiences from the first phase we knew that usability evaluation of computer systems is a broad concept often used indiscriminately, so we deliberately tried to identify and discard the papers that solely focused on proof of concept. In order to do so we reviewed the entire paper, which led to a set of papers that would be sufficient if we were interested in both formative and summative usability evaluations.

However as we have stated in the introduction, our focus is on formative usability evaluation, and thus we needed to identify and discard those papers that only concerned summative evaluations. Thus in the third and last phase we read all papers, in order to pick out those who dealt with the actual identification of usability problems. These papers were then processed again in order to determine where and how the identification takes place.

Preliminary Results

Through this selection process we ended up with the following distribution of papers from the different outlets (see Table 2).

Outlet	Number of papers in outlet	Development or evaluation of mobile systems	Usability evaluation of mobile systems	Identification of usability problems with mobile systems
Mobile HCI	302[1]	210	75	35
CHI	598	71	31	11
MUM	77[2]	47	12	7
UIST	204	29	9	1
IJHCS	497	21	5	1
TOCHI	113	17	6	5
DIS	143	15	6	3
AVI	220[3]	15	1	1
CSCW	190	5	3	3
JUS	16[4]	2	2	2
Total	2360	432	150	69

Table 2 Distribution of papers on the different search criteria.

Out of the 432 candidate papers concerning the development or evaluation of mobile systems 150 papers include usability evaluation and of these only 69 papers included descriptions of formative usability evaluations of mobile systems. Out of the ten different outlets the *Conference on Human Computer interaction with Mobile Devices and Services* (Mobile HCI) stands out as being the primary contributor supplying half of the papers in each phase of the selection process. Despite some of the other outlets are biannual it is quite significant, yet understandable considering the title of the conference, that no other outlet deals with the identification of usability problems with mobile systems in such an extent.

When the resulting papers were reread in order to achieve an understanding of the purpose and method used in each evaluation session we were able to identify the most commonly applied usability evaluation methods. We learned that at the current state, researchers tend to use four types of evaluation methods; expert evaluations, laboratory evaluations, field evaluations or longitudinal evaluations (see Table 3).

Evaluation Method	Papers	Total
Expert Evaluation	(1)(2)(3)(4)(5)(6)(7)(8)	**8**
Laboratory Evaluation	(9)(10)(11)(12)(13)(14)(15)(16)(17)(18)(19)(20)(21)(22)(23)(24)(25)(26)	**18**
Field Evaluation	(2)(15)(16)(22)(24)(27)(28)(29)(30)(31)(32)(33)(34)(35)(36)(37)(38)(39)(40)(41)(42)(43)(44)	**23**
Longitudinal Evaluation	(45)(46)(47)(48)(49)(50)(51)(52)(53)(54)(55)(56)	**12**
No explicit method mentioned	(57)(58)(59)(60)(61)(62)(63)(64)(65)(66)(67)(68)(69)	**13**

Table 3 Types of usability evaluation methods identified in the literature review.

The papers that did not contain descriptions of the applied evaluation method will not be processed any further, since we can't say with certainty how and why the usability evaluation has been conducted. In the following we will present our findings regarding how and why these four evaluation methods are applied by HCI researchers.

Currently Used Evaluation Methods

Below we document the four evaluation methods that are applied in the research papers that involve formative usability evaluations. The purpose of this section is to describe the way in which the evaluation methods are applied and customized in order to better suit the purpose of evaluating the usability of mobile systems.

Expert Evaluation

An expert evaluation is typically regarded as an inexpensive way to perform usability evaluations. Several of the articles in this study employ either heuristic inspections or cognitive walkthroughs in order to identify usability problems early on in the development phase (1) (2) (3) (4). Expert evaluations tend to be described as inexpensive and rapid, but despite this they seem to be the least preferred evaluation method in the reviewed articles, only applied in 8 of the 69 research studies. This might be because none of the existing expert evaluation methods take the context into account.

This lack of contextual involvement in the methods is a theme that the researchers approaches in (5) where a set of heuristics tailored especially for use in evaluations of mobile systems is constructed. They conclude that mobile heuristics although better suited for evaluating mobile systems should not be considered an alternative to user studies but synergic. In a similar study (6) researchers concluded that a set of adapted heuristics are well suited for the mobile nature of the systems that are evaluated. In (7) the researchers set out to explore the potential of improving the cognitive walkthrough method by supplying the experts with video of users using the system in the proper context. In this way the experts gain valuable insight in the context of use and the researchers conclude that this approach is just as good as conducting cognitive walkthroughs in the context of use and only slightly more expensive than a standard cognitive walkthrough.

The challenge of introducing context in an expert evaluation is approached from a different angle in (8), where the researchers compare three types of expert evaluations; heuristic inspection, heuristic walkthrough and contextual walkthrough. The researchers come to the conclusion that it is possible to introduce contextual detail and hereby bridge the ‘realism gap’, by introducing scenario based tasks in the heuristic walkthrough, thus providing a semi-realistic use case to the expert evaluators. Expert evaluations should thus be taken into consideration when evaluating the usability of mobile systems, especially if rapid and inexpensive evaluations are the primary concern.

Laboratory Evaluation

Usability evaluations performed in a usability laboratory as prescribed in (Rubin 1994) is a de facto standard in the HCI community. 18 of the reviewed papers involved usability evaluations conducted in a laboratory. Out of these a majority follows standard procedures and applies pre test questionnaires to determine demographics, tasks to control the use of the system and a controlled environment wherein the evaluation is conducted (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21). Generally data is collected by the use of think aloud protocols and video feeds capturing the users interaction with the system. In most cases, the researchers tend to utilize the method without any regard to the mobile nature and diverse context of use of the system they evaluate.

However some of the research documented in the papers questions the methods ability to properly identify problems related to the mobile context that the systems are intended to be used in. Thus some researchers try to simulate the context of use in the laboratory, by simulating realistic use settings, i.e. a hospital ward (22), a shopping mall (23) or a military command post (24). In this way the researchers attempts to recreate a somewhat realistic environment, without losing the advantages of the laboratory. Thus it is still possible to retain a high level of control over the environment and take advantage of the specialized data collection technology of the usability laboratory.

In (22), the researchers come to the conclusion that this approach to usability evaluation with ecological validity is in many ways superior to usability evaluations conducted in the field because of the previously mentioned advantages and the fact that an evaluation in the actual hospital ward is a resource demanding procedure. It was however not possible to impose the same sense of criticality in the laboratory that the users experienced in their daily work at the hospital. While these results might indicate that it is possible to recreate ecological valid environments in the laboratory, different studies reach the opposite conclusion. In (15) the researchers come to the conclusion that the problems identified in the field and the laboratory differ significantly, and that the users' attitude towards the system also depends on the situation in which it is used.

Another approach to adapting the usability evaluations conducted in the laboratory to the mobile nature of the systems is to recreate the challenges of moving physically while using the system (25) (26). In this way the users operate the systems while being mobile, hereby enabling the researchers to observe how the mobility of the user affects the evaluation of the system. The mobility of the users was realized by letting them perform tasks while using different kinds of fitness equipment or walking on predetermined paths, laid out on the laboratory floor. It was concluded in (26) that the mobility of the users did not contribute to the identification of additional usability problems, but did in fact result in less identified problems than an evaluation conducted while the users was seated at a table. Whether or not laboratory based usability evaluations are capable of identifying all usability problems seem unlikely, but with a few modifications it might be a suitable method, although it probably should be used in correlation with other methods.

Field Evaluation

As an alternative to conducting usability evaluations in the laboratory, many researchers have experimented with field evaluations. In fact field studies are represented more times in this review than any other method (23 papers involve field evaluations). The reason as to why the field evaluation has received so much attention in the field of mobile HCI is directly related to the mobile nature of the systems. The typical argument for evaluating systems outside the laboratory is that the context of use of the system is not replicable in the laboratory (16) (27), be it the social, the physical or the temporal context of use. Thus many researchers argue that it is not appropriate to evaluate a mobile system in static surroundings, where the user interaction is limited by the setup of the laboratory.

Some researchers argue that it is a necessity to take the evaluation into the field in order to identify the correct usability problems in the system – problems that cannot be identified in the laboratory because of the rigid environment (15) (28) (29) (30) (31). Ecological validity thus seems to be of tremendous importance when evaluating mobile systems. Some researchers even claim that they are willing to sacrifice some of the control that the laboratory can provide in order to achieve ecological valid evaluations (32) (33).

In the research papers where this method is described, it is evident that in most cases the evaluators tend to apply well known techniques also used in the typical laboratory based evaluation. In general most of the papers describe evaluations where the users are to use the system in a natural environment, while being observed by a researcher, in some cases the users are asked to think aloud while using the system, thus providing the researchers with valuable information regarding the interaction with the system in the natural context. In most cases the think-aloud protocol is used in combination with a series of tasks, guiding the use of the system (16) (15) (34) (35) (32) (36) (2) (28) (37). Other researchers discard the think aloud protocol and rely solely on post test, semi structured interviews in order to avoid affecting the users while they are interacting with the system (38) (39) (33) (40) (41) (42). Likewise the users are allowed to explore the system without being restricted or guided by assigned tasks or imposed time limits (27) (33) (24) (43). Hereby the researchers hope to retain a higher level of ecological validity.

Four of the papers in this review contain descriptions of how system logs can provide valuable data sources, by informing the researchers about the users' interaction with the system (41) (33) (27) (44). This approach to obtaining data regarding the user interaction is non intrusive, but somewhat restricted by the fact that the interaction logs cannot collect the users' subjective input. Thus in most cases the logs are used in combination with video recordings and post test interviews in order to gain insight in the users' subjective thoughts and comments. Potentially system logs can be used to recreate or simulate the user interaction instead of collecting this type of data with custom cameras fixed on the devices or hand held camera operated by the researchers.

Generally the field evaluations are described as a useful way to gain insight in the usability problems that emerge when the system is used in the proper context, but some researchers claim that the same information could be obtained through laboratory evaluations as well (16) (22). In both cases the researchers conducted comparable laboratory and field situated evaluations and concluded that there is little to gain from evaluating in the field and that evaluating in the field was more resource demanding.

However other researchers reach different conclusions (15). In this paper the researchers conclude that while the field evaluation is slightly more resource demanding than the corresponding laboratory based evaluation, it is also capable of revealing many significant problems that were not identified in the laboratory, thus the researchers argue that the cost-benefit is higher than that of the evaluation conducted in the laboratory. Whether or not evaluating in the field is better than in the laboratory is hard to say, but it is obvious that it is a popular method in the field of mobile HCI.

Longitudinal Evaluation

The fourth approach to evaluating the usability of a mobile system is longitudinal evaluations or field trials as they are often referred to. In this review we found 12 papers describing the use of this technique as a mean to identifying usability problems with a mobile system. The common denominator in these evaluations is the long period of use, varying from 3 days (45) to 9 months (46), which allows the users to get acquainted with the system and use it on their own premises. When and where the users choose to use the system is left to themselves to decide in order to avoid unrealistic usage. The fact that the users, due to the time span of the evaluation, are able to get to know the system is also an attempt to identify the important problems that occur in everyday use instead of problems that results from poor knowledge of the system. Thus the researchers hope that the problems they identify are problems that bother new users as well as more experienced users.

In the longitudinal evaluation the researchers do not supply the users with tasks as is typical for the field and laboratory based evaluations described earlier. The reason as to why the researchers do not use tasks to guide the use of the system is because the researchers try to preserve the realism of the use situation (47) (48) (49). Actually the researchers tend to avoid any pollution of the ecological validity by letting the users use the systems without any interference, which also is the reason why almost none of the evaluations are conducted with the use of observers.

In some cases the researchers rely on data collected by the users themselves through diaries (50) (51) or system logs that can provide insight in the use of the system (52) (46) (53). In almost all cases the researchers use post test interviews (52) (45) (53) (50) (54) or focus groups (46) (55) (51) (56) as a source of information. Hereby the researchers are able to gather user input and identify usability problems that should be dealt with prior to the release of the system.

Whether to use focus groups or interviews is an unaddressed issue, but both are regarded as suitable techniques for the purpose. In one case the researchers use video containing use situations as a source of inspiration in a focus group session, letting the participants relate to the usage shown on screen and thereby jump starting and structuring the discussion. It seems that the biggest challenge when conducting a longitudinal evaluation is the data collection process. The problem seems to be the gathering of dense and high quality data without disturbing the user and thereby minimizing the realism in the evaluation setup.

To sum up, evaluations of mobile systems are typically conducted using one of four methods: expert evaluation, laboratory evaluation, field evaluation or longitudinal evaluation. These methods are not rigid and can be varied in many ways using different techniques in different phases of the evaluation. Thus they should not be seen as homogeneous and static, rather they constitute a framework that can be used to construct an evaluation method that fits the purpose.

Evaluation Objectives – Three Variables

In the research literature it seems that there are three objectives that are dominant when constructing or choosing evaluation methods and techniques: *realism*, *control* and *resources*. These objectives can be seen as parameters or variables that the evaluators can adjust by modifying or customizing the methods and techniques used.

Realism

Realism, or ecological validity as some researchers tend to refer to it, can be achieved in an evaluation by letting the users roam free in a natural use habitat with as little disturbances from the evaluators as possible. Often realism is perceived as a variable directly dependant of the spatial context or location in which the evaluation is conducted, but that might be too simple a conception. Realism can also be seen as an expression of the social and temporal context in which the user is supposed to interact with the system. For instance the users might identify one set of usability problems if they are placed in a crowded concert hall during a philharmonic concert and then asked to evaluate a novel menu structure on a mobile phone. A totally different set of problems might be the result if the concert hall was empty or if the evaluation was conducted in a usability laboratory or while using public transportation. Thus realism should be seen as a result of the spatial, social and temporal context and their respective coherence with the intended use situation.

In the reviewed papers the researchers often try to increase the realism of an evaluation by reducing their own influence on the users, for instance by using self reporting (51) (54) or post interviews (27) (38) (39) instead of direct observations captured on video and think-aloud protocols, but not all researchers take such radical steps to ensure realism. For instance some researchers try to introduce a degree of realism in a cognitive walkthrough by giving the experts access to video recordings of actual use situations involving real users. Hereby the experts gained valuable insight into both the context of use and the typical user profile, which helped them structure their own review of the system so it resembled actual use cases (7).

Control

Control or ecological consistency as it sometimes is referred to, can be seen as the amount of control the researchers have over a given evaluation setting, with respect to where, when and how the user is supposed to use the system. The typical laboratory evaluation is a good example of a setup that provides the evaluators with a great amount of control. For instance the evaluators can control whether or not the user should be disturbed during the evaluation. Likewise the evaluators are able to structure the use of the system by providing the user with tasks. In this way the evaluators can utilize the setup of the laboratory evaluation to gain a high level of consistency in the evaluation.

The laboratory is as such a suitable environment for evaluations when the primary objective is a high level of control. However, when moving the evaluation into the field, it can be harder to achieve a high level of control. This is due to the fact that such evaluations often are conducted in a public space, where the evaluators have little or no control over the surroundings. For instance it might be difficult to avoid interaction with other people when the user moves around in a public place like a park or a subway station. Likewise it is impossible for the evaluator to control environmental factors such as weather and lightning conditions. Thus it is often perceived as hard to achieve control and thereby consistency in the field environment. However many researchers tend to emphasize the need for realistic evaluation contexts when evaluating mobile devices, and if needed they will be willing to sacrifice some of the control and consistency that can be achieve in the laboratory (32) (33) (15) (47).

Resources

The third variable resources, is often perceived as an expression of the combination of physical resources and man-hours spend on the evaluation. This definition should however be extended to encompass the competences required of the evaluators, since competences are regarded as a valuable resource in most modern software companies. If the primary objective of an evaluation is to identify usability problems within the limits of a low budget and a short period of time, most evaluators probably would turn to expert evaluation methods, such as the heuristic inspection or a cognitive walkthrough. However these techniques in their original form, has proven to be unsuitable for identification of usability problems related to the context, which is why some researchers are working on improving the methods, in order to adapt them to evaluation of mobile systems (8) (7) (5).

While expert evaluation is one way of minimizing the resources needed in order to conduct an evaluation, an alternative is lowering the cost of more expensive methods, such as the laboratory or field evaluation. These evaluation are commonly perceived as more expensive than expert evaluations, but never the less, they are often used when evaluating mobile systems (53 out of 69 evaluations in the reviewed papers are conducted in a lab or in the field, while only 8 are conducted as expert evaluations). Many researchers are currently experimenting with new methods and techniques that are supposed to lower the cost of evaluating software. Such methods and techniques are sometimes referred to as “discount” or “quick and dirty”. Examples of such experimental techniques are the use of system logs instead of expensive and time consuming video recorded observations as a way to gain insight in user interaction with a mobile system (39) and the use of novices as moderators instead of experts in laboratory based usability tests (21).

Interdependency of the Variables

Realism, *control* and *resources* should not be seen as isolated variables, rather they should be perceived as interconnected and dependant on each other. For instance it is likely that the evaluators might compromise the realism in a field evaluation by imposing a set of control mechanisms such as restrictions on the context of use as well as predefined tasks restricting the use of the system. Likewise it is likely that the evaluators control over the evaluation is compromised when conducting unsupervised field evaluations without predefined tasks. Thus the evaluators will have to decide on which aspects they wish to focus. This interdependency of the variables should not be seen as a linear function, where actions taken in order to affect one variable, has a proportional negative effect on the other variables. Instead we use an elasticity metaphor when we consider the two variables realism and control. This metaphor was found suitable because of the fact that not all attempts to maximize one variable has a negative effect on the other, thus it can be conceived as an elastic band which can be stretched without breaking. The adoption of this metaphor of course also implies that the band can be stretched too far and break – a break represents impossible and utopian evaluation setups. The third variable, resources, is likewise interconnected with the aforementioned; since actions taken to lower the amount of resources needed in some cases can influence the control or realism of the evaluation. Thus it is up to the evaluators to assess and apply the appropriate methods and techniques to adjust the variables and achieve their objectives.

Mapping the Variables in a Matrix

Based on these three variables we propose a matrix which can be used as a categorization tool for different methods. By mapping the state of each variable in different methods we are able to create a categorization system that might provide evaluators with an overview of the available “tools” in their “toolbox”. This allows us to compare the realism and control provided by different methods as well as their respective demand for resources. The high-medium-low scale used in this matrix can be modified if needed, but in order to keep the representation simple and easily understood we choose to stick to this three point scale. The methods mapped in this matrix, represents the typical methods applied in the reviewed papers in order to identify usability problems in mobile systems. Thus it should not be perceived as an exact and static representation of the state of the three variables in any method. The four methods described earlier might be applied in a variety of ways, which of course would result in different mappings.

	Expert-based	Laboratory-based	Field-based	Longitudinal
Realism	Low	Low	Medium	High
Control	High	High	Medium	Low
Resources	Low	High	High	Medium

Table 4 Evaluation methods emphasis on realism, control and resources

The matrix (Table 4) is an attempt to illustrate how different evaluation methods emphasize different variables depending on their objectives. For instance the expert evaluation is regarded as a method with a low emphasis on realism, much like the laboratory evaluation. This is due to the fact that, in their standard form, little or nothing is done to introduce ecological validity and natural use settings. Instead the focus is on controlling the way the system is used or inspected in order to ensure consistency and scientific validity. The field evaluation and the longitudinal evaluation both are examples of evaluation methods that emphasize realism, at the expense of lower control. It is typically attempted to introduce ecological valid use settings that allow for realistic usage. In the field evaluation though, the use of the system is often controlled to some extent, since the purpose is to test certain parts of a system in a relative short period of time compared to the longitudinal evaluations. Thus the field evaluation is not nearly as focused towards realism as the longitudinal evaluations. The amount of resources spent on the individual evaluation methods vary a lot but in general it seems that the highly controlled laboratory and field evaluations are costly procedures. This is due to the extensive analysis of collected data and the cost of establishing a suitable use setting, be it in the laboratory or the field. The reason as to why the resource consumption of the longitudinal evaluation is mapped to medium is because a lot of the hassle in this method is left up to the participating users. Another reason as to why the longitudinal evaluations are cheaper to conduct is because of the quantity and quality of the collected data. The overview provided by this matrix depicts the methods as they are generally applied which might not correspond to specific implementations. This however should not be seen as a limitation, since the matrix should be perceived as a way of expressing the different foci of different methods in a high level abstraction.

Answering the Research Question

Based on the knowledge that we have gained through the literature review we are now able to answer the research question we initially set out to answer. Based on the literature review we have reached the conclusion that researchers currently are applying four types of usability evaluation methods:

- Expert evaluations
- Laboratory evaluation
- Field evaluations
- Longitudinal evaluations

These methods are implemented and utilized in a variety of ways, as an attempt to cope with the challenges that arise when evaluating mobile systems. We have documented how researchers attempt to achieve different objectives with different methods, and accounted for the three variable objectives: *realism*, *control* and *resources* that influence the implementation of the methods that are described in the reviewed literature.

Applying the Methods

In this chapter we will apply representative methods of four currently most used methods that we documented in the previous chapter. This is done in order to gain first-hand knowledge about the characteristics of the individual methods and to document the distinctive features of the contextual methods as well as the non-contextual methods. By applying the methods and describing their characteristics we can approach an answer for the second research question that we proposed in the introduction:

> Research question #2: *What characterizes the application and outcome of evaluating the usability of a mobile system with the current methods and how do the characteristics of the contextual and non-contextual methods differ?*

Initially we will describe the procedure of our evaluation process and document the application of the individual methods.

Procedure

Before describing how we applied the different methods we will briefly account for the objective of the four evaluations, the procedure used to categorize and rate the severity of the identified problems, the system that we selected as our case and on which we applied the usability evaluation methods and the selection of users and experts who participated in the evaluation sessions.

Evaluation Objective

The evaluations were all conducted with one primary objective, namely the identification of as many usability problems as possible with the individual methods. In order to be able to compare the problems identified with the individual methods, we will refrain from generalizations when describing the identified usability problems. Instead we will emphasize detailed descriptions of every single problem and try to keep them at an atomic level. As Hartson et. al. states in Criteria for evaluating usability evaluation methods:

"Comparison requires complete, unambiguous usability problem descriptions that facilitate distinguishing different types of usability problems" (Hartson, Andre and Williges 2003)

Thus by applying a high level of detail in the problem descriptions we will be able to compare the findings of the evaluation methods.

Problem Severity Taxonomy

In order for us to be able to rate the severities of the identified usability problems in a uniform way, we chose to use a simple three phase taxonomy inspired by Molich (Molich, 2000). The three steps in this taxonomy is based on a number of factors, such as the amount of time spent on the problem, the time it takes the user to learn to operate the system containing the problem and the inconvenience that the user feel and expresses. All of these aspects where expressed into simple statements that express the three types of severity and converted into the following taxonomy:

- Critical problems: Problems that hinder further use of the system and are experienced as being very irritating, thereby minimizing the user satisfaction.
- Severe problems: Problems that delay the user several minutes and irritates the user, but do not hinder the use of the system. These problems often lead to users misunderstanding the actions and status of the systems.

- Cosmetic problems: Problems that only delay the user for a short amount of time, but as the user gets acquainted with the problem he/she learns to work around it. These problems have very low impact on both the use and the user satisfaction.

By applying this taxonomy we are able to assess the severity of the identified problems in a uniform way, supported by the thorough descriptions as well as reviews of collected data material.

Selecting the Case – EQO Mobile®

The system that we chose to evaluate was the java application EQO Mobile® (Beta version 0.94.3) (Figure 4), which is a multi protocol instant messenger client for mobile phones. The system allows the user to use IM services such as MSN Messenger®, Google Talk®, ICQ® and Jabber® without installing any additional software. Furthermore it contains functionality that allows the user to use the popular VOiP technology Skype® through any java compatible GSM mobile phone. The system enables the user to stay in contact with IM contacts when being away from the computer by making the IM services mobile and allowing the user to be available at all times. There could be problems associated to this transformation from well known stationary IM clients used on desktop computers to a single mobile client supporting several protocols. Therefore, in order to realize the number of intended evaluations we limit the scope of the evaluation by focusing on the implementation of the MSN Messenger protocol in EQO Mobile®. The target audience for EQO Mobile® is IM Service users, that at the same time are experienced users of mobile phones and GPRS based services.

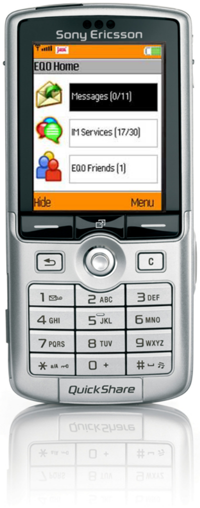

Figure 4 EQO Mobile on a SE K750i

This particular system was chosen because of the fact that it is currently under development, and even though it is quite stable it does contains usability problems. The system can be characterized as a consumer product intended for use in a wide range of contexts, or simply a multi contextual consumer product. Thus EQO Mobile® can be applied in a variety of environments and under various forms of mobility (modalities in Figure 1). By applying the four different methods on this particular system we will thus be able to evaluate the methods' ability to cope with the rich contextual variations. Besides, EQO Mobile® extends a well known IM service to the mobile phone, opening up for new use contexts and represents a new technology that is yet to gain popularity in the mobile consumer market.

Participating Users and Experts

The participants in the laboratory evaluation, the field evaluation and the longitudinal field evaluation, were chosen from master students at Aalborg University, all in the age group 20 - 26 years. Out of the 16 users involved in the three empirical evaluations (laboratory, field and longitudinal) 5 were female and 11 were male. All participants were regular MSN Messenger users and familiar with the mobile phone that was used during the evaluation (SE K750i). All participants were equipped with a SE K750i set up for GPRS and with EQO Mobile® preinstalled, for use in the evaluation. In the heuristic evaluation we used four HCI master students, currently finishing their master thesis, as experts. All had previous experience in conducting heuristic inspections and were well acquainted with Jakob Nielsen's heuristics that we used in the inspection.

The number of participating users and experts in the individual evaluations varied from 4 to 6. This variation in the number of participants is a deliberate choice, since we wanted to evaluate the evaluation methods on their own terms we have chosen to follow the guidelines that exist on the matter. Thus we have 4 experts participating in the heuristic inspection, which is well within the 3-5 expert recommendation of Jakob Nielsen (Nielsen n.d.). Likewise the Laboratory evaluation is conducted with 6 participants, due to the 4-8 participant recommendation in the guidelines set out by Jeffrey Rubin (Rubin 1994, pp. 128). This number of participants is repeated in the field evaluation, since the field evaluation essentially is an attempt to recreate the laboratory evaluation in a natural environment. For the longitudinal evaluation there exist no exact guidelines on the appropriate number of participants. Thus we have chosen to apply four users, which is due to the fact that we perceive the longitudinal method as an approach that provides rich data over an extended period of time with few participants.

The Applied Methods

As previously described we chose to apply four evaluation methods that represent our findings in the literature review. Thus we applied two non-contextual methods (a heuristic inspection and a laboratory evaluation) and two contextual methods (a field evaluation and a longitudinal evaluation). The distinction between contextual and non-contextual methods is defined by the context in which the evaluation takes place. If the evaluation is conducted in the real context of use, the methods will henceforth be defined as contextual methods and the methods that are not conducted in the real context of use will be defined as non-contextual methods.

The heuristic inspection was chosen as a representative of the expert evaluation methods. The reason why we decided to go with the heuristic inspection was that it was the most commonly applied method in the reviewed papers. Likewise the laboratory evaluation, the field evaluation the longitudinal evaluation were deliberately applied in a way that resembles the most common application of the methods in the reviewed papers.

In order to conduct these four evaluations in a reasonable amount of time we chose to start with the longitudinal evaluation. During the extended period of time the longitudinal evaluation required the users to use the system in we were able to conduct both the laboratory evaluation and the heuristic inspection before we were able to finish the longitudinal. Finally based on some of our experiences from the longitudinal evaluation we were able to conduct the field evaluation. This chronological order is not represented in the following, which is structured by the order in which the methods where described in the previous chapter.

Heuristic Inspection

The first evaluation that was conducted was a heuristic inspection conducted with HCI experts that inspected the system guided by the set of heuristics developed by Jakob Nielsen (Nielsen, Usability Engineering 1994). The purpose of this method is to identify usability problems in a cost-efficient way without involving end-users. It can be conducted almost anywhere, depending on the nature of the system. Besides it is easy to plan and execute due to a simplistic setup and an easy to grasp procedure, where the expert inspects the system using a given set of heuristics and notes when the system violates a given heuristic. The method is commonly used and well documented in the literature.

The evaluation took place at Aalborg University in our office and no particular modifications were made to the office since the physical use situation was not intended to resemble the context of use or include elements of mobility. The expert was situated at a desk in the office and a participating researcher sat next to him equipped with a laptop. The role of the researcher was to observe and document the problems that the experts identified, which allowed the expert to focus on the system and its relation to the heuristics. The reason as to why we chose to include the researcher is that we wanted detailed problems descriptions that would be comparable to those we would derive from the other evaluations. Thus the researcher was able to ensure that the descriptions were stringent and easy to interpret.

The four experts were all HCI master students with prior experience in carrying out heuristic inspections and each expert inspected the system one at a time. The procedure of each inspection was that before the expert was introduced to the system he was reminded of the ten heuristics (see Appendix A) and they were discussed in order to reach a common understanding of their meaning and application. After the review of the heuristics the experts were introduced to the system that they were to inspect and asked to use the system for a short while so that we could make sure that everything worked as planned and that the experts understood the basic principles of the system.

When the experts had been properly introduced to the heuristics and the system, the inspection was commenced. During the inspection it was left up to the expert how to use the system. Every time the expert noted a violation the researcher asked the expert to elaborate on the violation in order to obtain a detailed description of the problem. The researchers function was limited to this descriptive function, and thus they refrained from communicating with the experts in all other situations.

The compilation of a list containing all the problems identified by the four experts was initially done by each researcher individually. Each researcher assessed the severity of the listed problems and afterwards a merged list of usability problems and their assessments was agreed upon by discussing each problem and until an agreement was reached. The merged list therefore ended up containing a description of the problem, information about who encountered the problem, the heuristics that had been violated and the severity of the problem.

Laboratory Evaluation

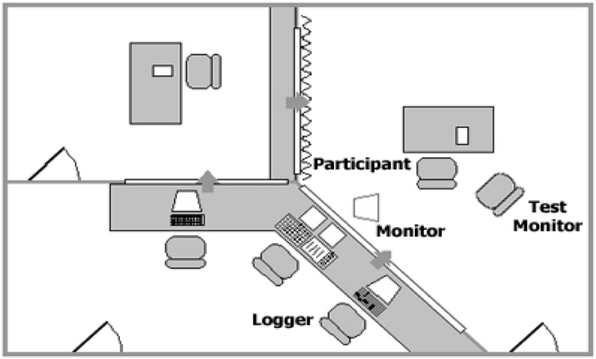

Figure 5 The laboratory setup.

The next method was a usability test executed in a usability laboratory with potential end users as described by Jeffrey Rubin (Rubin 1994). The purpose of this method is to identify usability problems, in a controlled environment so that problems or incidents can be reproduced. The evaluation was conducted using a think aloud technique based on K. A. Ericsson and H. A. Simon's work (Ericsson and Simon 1980, Ericsson and Simon 1984), which along with video and audio recordings is a standard procedure for data collection in a laboratory evaluation. The think aloud technique allowed us to preserve the volatile thoughts that the users expressed during the evaluation in a non-volatile way by preserving problems with audio and video. The usability problems were then subsequently derived from a thorough analysis of the audio and video data. Typically usability laboratories are equipped with high definition cameras and microphones in order to capture this kind of rich data.

We conducted the laboratory evaluation in the usability laboratory at Aalborg University, which is a fully equipped usability laboratory allowing control of the environment and capturing of high definition audio and video. The test setup can be seen in Figure 5, where one researcher took on the role as test monitor alongside with the participant and the other researcher operated the cameras and took on the role as a logger, taking notes in the adjacent monitoring room behind the one-way mirrors. The evaluation was conducted with the participation of six master students attending Aalborg University. The six participating users were to evaluate the system one at a time and prior to the evaluation they were all given a short introduction to the test system along with a tour of the laboratory. This was done in order to demystify the users and prepare them for the evaluation.

The evaluation in itself was conducted in two phases. In the first phase the users was presented with a series of tasks that they were asked to solve using the system. The tasks were designed to resemble small scenarios that contained common use situations, such as creating a MSN profile, adding and removing contacts, opening and closing chat sessions and setting the user status. During this phase, the participating users were asked to think aloud, but since this technique is hard to master, we took great care to instruct each participant before the actual evaluation began. During these instructions the participants were also informed about the purpose of the evaluation and the confidentiality issues regarding the recorded video and audio material. This information was standardized in written form to ensure that all test persons were equally well informed (see Appendix B). During the evaluation the test monitor encouraged the participants to think aloud if they forgot to do so. The second phase of the evaluation consisted of a short semi-structured interview based on the problems the user had experienced in the previous phase. This interview allowed us to clarify uncertainties regarding the problems and the users' overall opinion on the system.

Subsequently the combined video and audio data was analyzed individually by the two researchers which resulted in two lists of severity assessed usability problems. These two lists were then discussed till a merged list was agreed upon. The merged list ended up containing all the problems identified by the researchers along with the names of the users who experienced the problem in order to be able to count occurrences of each problem and lastly the observations that led to the identification of the problem.

Field Evaluation

The field evaluation was designed to resemble the typical field evaluation described in the reviewed papers. That includes users using the system to accomplish a set of predefined tasks in a predefined context while the researchers observe and record the session on a digital video camera. It was the intention that the ecological validity provided by the natural context of use would add a realistic touch to the evaluation, which in turn would enable us to uncover usability issues related to the use setting.

The process of choosing a proper context for the evaluation was aided by the knowledge gained from the longitudinal field evaluation. We found that the users most frequently used the system while using public transportation. Thus we chose to conduct the field evaluation in a city bus in Aalborg. Since the context was to be authentic we decided to conduct the evaluation in a bus in service during the late afternoon and evening hours. In this period the busses on route 12 was not overly crowded which allowed us to occupy the space needed to conduct the evaluation.

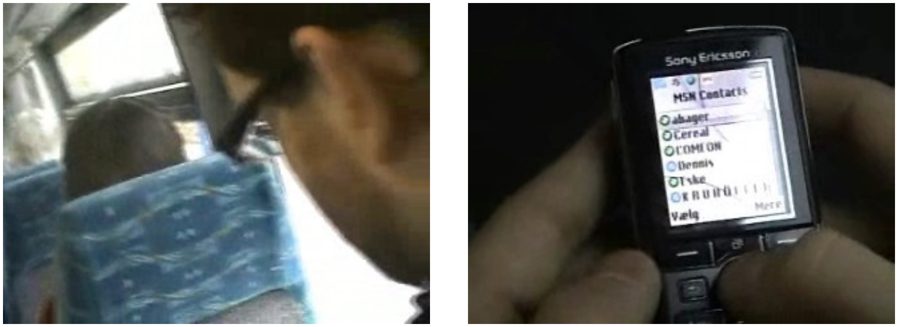

Figure 6 The field evaluation setup.

In order to collect high quality data for later review and analysis we chose to apply a hand held video camera. This data collection technique is typically applied in conjunction with fixed cameras on the device itself. In this evaluation however we found that it would hinder the participants natural use of the system because of the confined space in which it was to be used. In order to deal with this issue, we used the hand held camera to capture the surroundings as well as the interaction with the system (see Figure 6). The camera used was a digital video camera capable of providing high quality images of the display and buttons on the mobile phone. The camera was operated by hand in order to deal with the movements of the device caused by the movements of the participant and the vibrations of the bus. The build-in microphone of the camera was used to capture audio, which was necessary to capture the thoughts uttered by the users.

The evaluation procedure was structured in three phases. The first phase was an introduction phase, where the participants were introduced to the purpose and the procedure of the evaluation and the think aloud technique that we intended for them to use during the evaluation. This phase took place in a small office on campus. The reason as to why this phase did not take place in the bus as the second phase was that the system had proven to be somewhat unstable and we needed some buffer time in order to be able to deal with problems before entering the bus. In the first phase we asked the user to create a MSN profile and log on to the system, since the successful completion of this task indicated that the system was stable and hopefully would remain so for the rest of the evaluation. Furthermore during this process the participants were able to familiarize themselves with the application and the phone that was used in the evaluation.

In the second phase, the participants took the bus from the campus area to the bus terminal in down-town Aalborg; a trip lasting approximately 30 minutes. During the trip the participant was placed at a window seat with a researcher, undertaking the role as test monitor, occupying the seat next to him/her. The camera operator was seated in the seat right behind the participant and the test monitor. From this position it was possible to capture the interaction with the system as well as the surroundings, although not at the same time. During the second phase the participants were instructed to complete four tasks that they were given by the test monitor. The tasks were designed to structure the use of the system in a way so all the relevant functionalities were evaluated. In this way we were able to ensure that the participants used all the functions related to the implementation of the MSN Messenger part of EQO Mobile®. This includes adding and removing contacts, setting the user status, chatting with contacts and logging on and off MSN Messenger (See appendix C).

As in the laboratory evaluation, after the tasks were completed the participants were asked to express their overall impression of the system and comment on the problems that they experienced during the evaluation. The final phase was the analysis phase where the evaluators individually reviewed the recordings and identified usability problems for each of the participants. Each problem was noted with comments on the actions of the users and their verbal comments. The result of this procedure was once again two lists containing the severity rated usability problems. These lists were then merged in a combined effort by the two researchers.

Longitudinal Evaluation

The longitudinal evaluation was deliberately conducted in a way that resembled the common longitudinal evaluations described in the literature review. The evaluation was conducted over a two week period, without any direct observation by the researchers. Instead we adopted the self reporting techniques described in some of the reviewed papers, using only interviews and diaries as data sources. Therefore this method will be referred to as the *diary evaluation* in the remainder of the thesis. The primary objective of this evaluation method was to identify usability problems that occur in different contexts over a longer period of use. Thus we had to give up the total control of the context that we might be able to achieve in a laboratory. This loss of control was the price we had to pay in order to gain a broader insight in problems related to the use of the system in diverse contexts.

The overall setting or the context of use in the diary evaluation was not determined by the researchers, but instead it was left up to the participants. Hereby the usage was not restricted to a specific place or time, but instead the users where free to determine when and where the system should be used. This freedom to choose the context of use was a deliberate attempt to achieve a higher level of realism, which in turn we hoped would result in a diverse pattern of use. Diverse usage of the system was one of the primary objectives of the evaluation method, since we believed that it in return would lead to a broader range of problems related to the use in different situations.

The data collection technique in the diary evaluation was self reporting through diaries combined with post and mid evaluation interviews. Prior to the evaluation the participants were each equipped with a diary where the first page contained a short introduction to the data collection process they were to take part in (see Appendix D). This introduction contained information on what kind of data we were attempting to collect and how we expected them to collect it. Thus we hoped that the diaries in the end would contain detailed records of the experiences of each of the participants.

The evaluation was structured in three phases: A pre evaluation briefing, a two week evaluation phase and an analysis phase. At the pre evaluation meeting, the participants were introduced to the purpose of the evaluation, the system and the use of the diary. Furthermore we cleared any technical problems related to setting up a GPRS connection and installing EQO Mobile®. In the second phase the participants were to use the system over a two week period. During these two weeks the participants were free to use the system in whatever way they liked. The decision to conduct the evaluation without tasks to structure the use of the system was a deliberate attempt to maintain ecological validity.

During the second phase there was no direct contact between participants and researchers. The only contact during the evaluation was two scheduled interviews; the first interview after 4 days and the second interview after 8 days. These interviews where intended to be used for gathering user experiences, for later use in the post evaluation interview and was conducted through EQO Mobile®. Likewise it was the intention that the participants could be stimulated to further exploration of the system through these interviews. At the end of the two week evaluation period, the diaries were collected and during the next couple of days the participants were interviewed for the final time. This interview was the last chance to explore and analyze the problems that the participants had identified and noted in the diaries or reported during the two mid evaluation interviews. As a result of the last interview we were able to compile a list of usability problems experienced by each of the participants with detailed descriptions of the problems that they encountered during the evaluation. In the third and final phase the lists were compared and discussed and in the end we were able to merge them into one list containing all the problems that were identified by one or more of the participants. These problems were then assigned a severity rating in the same way as it was done in previous evaluations.

Merging the List of Usability Problems

The four resulting lists of identified and severity assessed usability problems from each evaluation method were finally merged into one combined list of problems containing the detailed descriptions and the severity ratings along with the number of experts or users that identified or encountered the problem. This was again done by discussing each problem based on the description and comparing similar problems to each other in order to refine the list of usability problems when possible, and to avoid duplicate problems. This way the assessments of severity was also double checked.

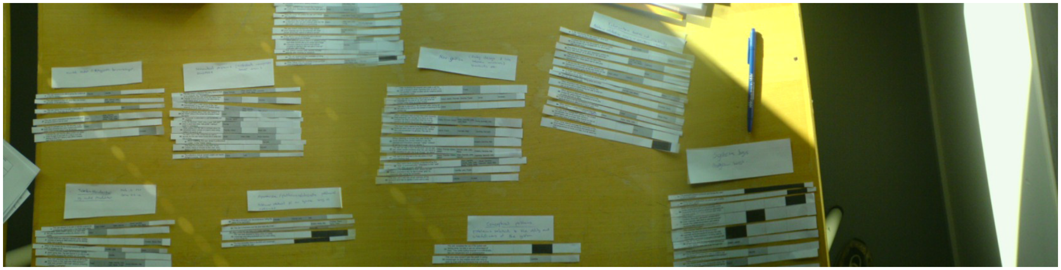

Figure 7 Categorization of usability problems in themes.

As a part of the process of merging the lists the problems were clustered thematically in order to structure the final list of problems (see Figure 7). This should help us compare the different methods and determine what type of problems each method identified. The themes were conceived in the process of clustering the problems.

Each problem was discussed and a theme was agreed upon, based on the experiences gained from the four evaluations considering symptom, cause and effect of each usability problem. The following 9 themes were identified and agreed upon:

1. **Standards**: Problems related to the users knowledge of the standards used in this type of technology and the adherence of standards defined by the phone and the metaphors in use.

2. **Mental Model**: Problems related to disagreements between the users perception of the system and the actual system; in most cases due to previous experience with MSN Messenger on PC's.

3. **Semantics**: Problems related to misinterpretation of menu items or icons. Poorly designed icons and bad phrasing of labels are typical for this type of problems.

4. **Ergonomics**: Problems related to the physical use of the system. Often the problems in this theme are related to the technical attributes of the mobile device or a result of the physical use of the device.

5. **Feedback**: Problems related to feedback, such as the lack of feedback on user actions and poorly phrased feedback. Missing feedback that results in confusion is often a source of frustration.

6. **Navigation**: Problems related to poorly designed interface navigation and dialog design/task flow. Problems like these result in poor user control and possibly confusion, when interacting with the system.

7. **Information**: Problems related to the lack of information or poor visualization and availability of information. Information problems could be missing status indicators, letting the users guess the status rather than showing it, results in confusion.

8. **Utility**: Problems related to the utility and usefulness of the system. Weak concepts that provide little or no value to the end user could be defined as utility problems.

9. **Program Bugs**: Problems directly caused by bugs in the code that hinders the user in interacting with the system. Problem of this type often occur randomly and can be hard to reproduce.

This list of themes resembles an abstraction over the problems that were identified in the four evaluations, and thus should not be perceived as a general list that can be applied without modification in other studies. The final list of usability problems clustered by themes can be seen in Appendix E. In the following the findings of the four usability evaluations of EQO Mobile® will be documented and analyzed.

Results

Initially the general tendencies of the evaluations are depicted to form an overview of the results, including a presentation of the total amount of usability problems and the distribution of problems over severity and themes. Additionally we will sum up and present the temporal resources spent on each of the evaluations and account for the most demanding aspects of each method.

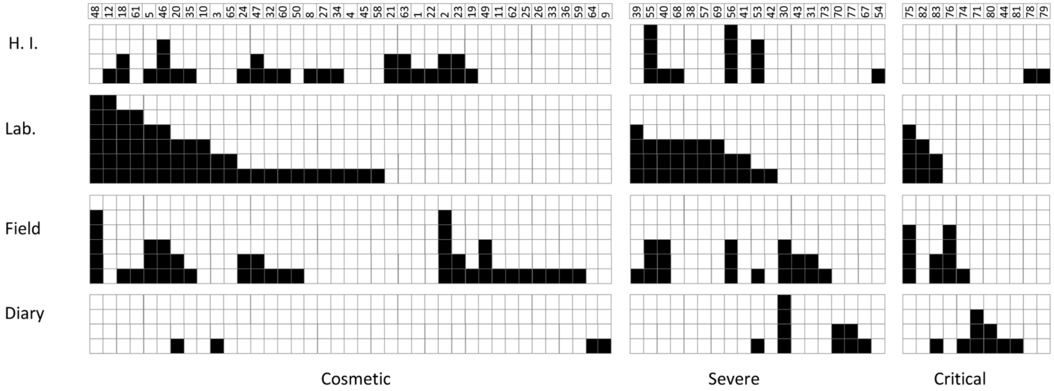

Figure 8 Distribution of usability problems based on severity.

The overview is followed by a thorough analysis of the individual methods identifying advantages and disadvantages of each evaluation method. By doing so we are able to document the trade-offs and outcomes of the individual methods and evaluate their applicability in the mobile paradigm. Finally we will try to depict the characteristics of the methods and compare them based on their individual strengths and weaknesses and analyze the different characteristics of contextual and non-contextual methods.

Overview

The four evaluations revealed 69 different usability problems with EQO Mobile®. As shown in Table 5 the heuristic inspection identified 2 critical, 6 severe and 20 cosmetic usability problems. The laboratory evaluation identified 3 critical, 11 severe and 22 cosmetic usability problems. The field evaluation identified 4 critical, 9 severe and 23 cosmetic usability problems and the diary evaluation identified 6 critical, 5 severe and 4 cosmetic usability problems.

	Inspection	Laboratory	Field	Diary
Critical	2	3	4	6
Severe	6	11	9	5
Cosmetic	20	22	23	4
Total	28	36	36	15

Table 5 Identified usability problems including severity assessments.

As expected, did the evaluations not identify the same amount of problems and the distributions on severities were not identical either. The laboratory evaluation and field evaluation did however identify the same amount of problems with a similar distribution of severities, meaning that they perform quite similar and possibly have characteristics in common, which require further analysis to uncover. The diary evaluation stands out as the only method identifying more critical and severe problems than cosmetic problems, but is at the same time the method which identifies most critical problems. We will look closer in to these observations in the following analysis.

Some of the problems were identified by multiple methods while other was uniquely identified by only one of the methods. This is illustrated in Figure 8 where the distribution of problems identified by each method is depicted along with the number of participants experiencing the problem for each severity category. Each column represents an identified usability problem and each black square represents a participant who encountered the problem. Reviewing Figure 8, we see that only two problems were identified with all four methods, problem number 20 being a cosmetic problem and problem number 53 being a severe problem. The table also shows that 30 of the 69 problems were identified by two or more methods and that the majority of the problems (47) were identified by at least two participants, although not necessarily two participant in the same evaluation.

As previously mentioned, the problems were thematically clustered in order to be able to abstract from detailed problem descriptions and visualize the general outlines of the types of problems each method revealed. The distribution of identified problems in each method on the different themes can be seen in Figure 9.

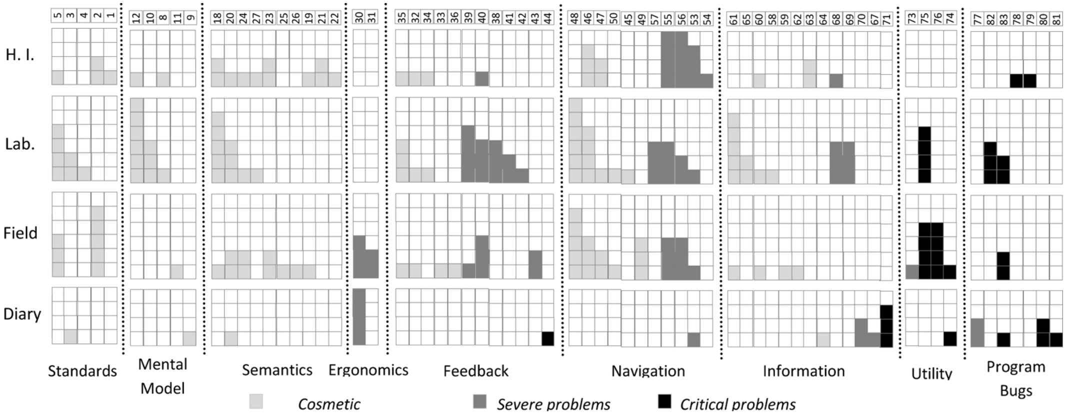

Figure 9 Distribution of problems based on themes.

Reviewing the figure we see that the most common themes are *semantics*, *feedback*, *navigation* and *information*. These themes are associated to the interface of the system, since they typically represent either bad phrasing of labels and menu items, lacking feedback to the user's actions, inconsistent navigational paths or inaccessible information.

We also see that all problems related to *standards*, *mental model* and *semantics* are exclusively cosmetic problems, and are thus experienced as being less severe than many of the problems related to the other themes.

Furthermore, all critical problems are related to the themes *utility* or *program bugs* except two problems identified in the diary evaluation; one being a problem related to lacking *feedback* and the other being a problem related to inaccessible *information*. And a final observation is that no problems related to *ergonomics* were identified in the heuristic inspection or the laboratory evaluation.

The resources used on each method are depicted in Table 6, which shows how many man hours the researchers spend on each evaluation.

	Inspection	Laboratory	Field	Diary
Planning	6 h	10 h	10 h	14 h
Execution	3 h	12 h	9 h	10 h
Processing	6 h	31 h	22 h	2 h
Time spend	15 h	53 h	41 h	26 h
Duration	1 day	3 days	3 days	14 days

Table 6 Time spend on each method, not including the participants. The test duration represents the time span of the execution of the method.

Four experts participated in the *heuristic inspection* and we spend a total of 15 man hours planning, executing and processing the evaluation and the data derived from it. The four heuristic inspections were executed the same day in our office at Aalborg University. Compared to the other methods it was easy to plan and execute because of the rule-of-thumb-based nature of the evaluation procedure and the existing guidelines in how to conduct such an evaluation (Nielsen 2003).

Six participants were recruited from the student body at Aalborg University for participation in the *laboratory evaluation*. The evaluation required 53 hours of work, distributed on planning and executing of the evaluation and processing of the derived set of audio and video data. The evaluations were executed over a three day period in the usability laboratory at Aalborg University. As mentioned earlier this method is well documented, which made it easy to plan and execute (Rubin 1994).

Six participants were also included in the *field evaluation* and we spend 41 man hours on planning, executing and processing the evaluation and the data derived from it. In planning and executing the evaluation we drew on the experiences from the literature review, and structured the method in much the same way as we would if we were to conduct a laboratory evaluation. The evaluations were conducted over a three day period and the processing of the audio and video data was the main contributor to the total amount of man hours spent on the evaluation. The reason why we were able to conduct the field evaluation faster than the laboratory evaluation was due to the fact the participants in the field completed their task faster which led to a smaller amount of video data.

Four participants were included in the *diary evaluation* on which we spend 26 man hours on planning, executing and processing. The processing of the derived results and the compilation of a final list of usability problems was easier in this case, due to the fact that the data from the final interviews that concluded each evaluation session was a list of usability problems that the participant encountered. Therefore the data from this method was by far the fastest to process as Table 6 illustrates. The time used on execution was primarily used for instructing and interviewing the participants, since there was no supervision or direct observation of the participants' usage of the system.

Analysis

In order to compare the methods, we will first analyze the individual trade-offs of each method. This will allow us to evaluate and illustrate the actual value of applying each method and compare them on a valid basis. Therefore, in order to characterize the methods the following section will document the advantages and disadvantages of each method, and evaluate the application of each method in the mobile paradigm. When accounting for the advantages we will focus on the individual methods' ability to identify problems related to the nine themes. The disadvantages will encompass the limitations of the methods with respect to the problems that they cannot identify and the practical disadvantages and obstacles that we have experienced during our application of the methods.

Heuristic Inspection

The heuristic inspection was conducted with minimal resource consumption as the primary objective. Thus it stands out in this aspect when compared to the other methods, with a time consumption of only 15 man hours during a one day period. The equipment needed for the evaluation was likewise very limited and thus the evaluation proved to be easy to conduct in a short timeframe. As a result of the low costs and the short period of time invested in the inspection we initially expected that it would mostly reveal lesser problems. This presumption was proven to be correct, since most of the usability problem that we were able to identify using this method were rated as cosmetic problems (20 out of a total of 28). This tendency is often documented in the research literature and thus came as no big surprise.

Advantages

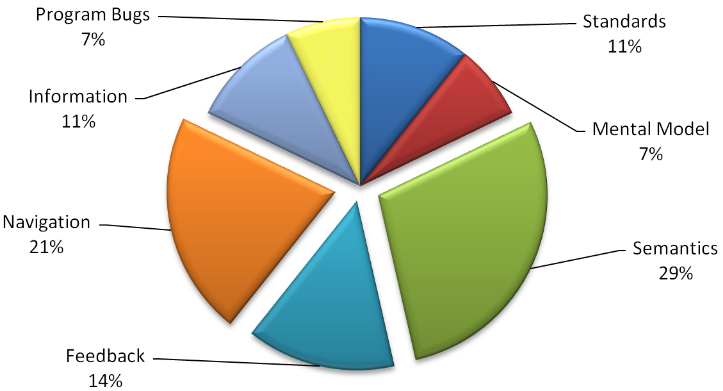

Figure 10 The distribution of problems over themes identified in the heuristic inspection.

If we take a closer look at the problems identified during the inspection, as depicted in Figure 10, we can see that the most common theme is *semantics*, which constitutes 29% of the identified problems. Poorly phrased menu items and abstract labels are the primary cause of this type of problems. An example of a problem related to semantics is the "ok" label assigned to the left soft button that appeared when the experts logged on to MSN Messenger. Almost all the experts that used this functionality thought that they had cancelled the login process and tried to reconnect though they were already connecting. The severity of this particular problem was rated as cosmetic, since it did not delay the users, but merely confused them. This problem illustrates how this type of problem often results in misunderstandings. In most cases however the true meaning of the labels were revealed through trial and error and the experts did not seem to be bothered by these problems in the long run. Although these problems might seem insignificant, it would be advisable that they were addressed in order to flatten the learning curve of the program.

The second most common theme is *navigation,* which constitutes 21% of the identified problems. Four of the problems related to this theme have been rated as severe. This is due to the fact that these problems result in frustration and confusion over a longer period of time. An example that illustrates this is the inconsistency of the back button, which in some cases returned the expert to his place of origin as expected and in other brought him to a new destination. Thus there is little or no logic in the way the back button works and it was very hard for the experts to figure out where the button would take them, which caused some frustration and delay.

The third most common theme is *feedback* which constitutes 14% of the identified problems. Three of the identified problems related to this theme were rated as cosmetic, while one was rated as severe. An example of the typical problem related to this theme, were the lack of feedback on the experts actions when they tried to interact with the system, while it was connecting to the EQO network. Instead of informing the experts as to why their attempts to interact with the system were futile, the system simply refused to co-operate which left the experts wondering what might be the cause. The problems related to feedback often caused the experts to second guess the actions of the system, which in turn led to some confusion.

Disadvantages

The drawback of the heuristic inspection is the inability to identify problems related to the utility of the system and ergonomics. The experts experienced no problems related to the physical interaction with the system in the static context of the office. Many of the participants in the field and diary evaluation reported that the limited input capabilities of the small keyboard and the normal rate of communication in a MSN Messenger chat, proved an obstacle that affects the usability of the system in a negative way. Since the inspection did not include long and tedious chat sessions, the experts did not seem to bother about the slow input rate and thus did not identify the issue as a hindrance to the usability of the system. The experts' lack of attention to utility aspects might be explained by their focus on the heuristics and their poor insight into normal use situations.

The Role of Heuristic Inspections in the Mobile Paradigm

It seems that the standard heuristic inspection under the circumstances presented in this report fares well when it comes to identifying problems related to the system interface and fails at identifying contextual problems such as those related to *ergonomics* and *utility*. Thus the standard heuristic inspection can be perceived as a method with focus on the structure and components of the interface, fit for identifying problems related to these areas.

The primary advantage of the inspection is the low resource requirements and time consumption, which makes it ideal for rapid evaluation during short development cycles. The heuristic inspection should be applied when the primary objective is rapid evaluation of new or redesigned interfaces. In order to improve and adapt the method to the mobile paradigm, steps could be taken to integrate the available knowledge of the context of use into the physical evaluation setup as well as in the applied heuristics.

Laboratory evaluation

The laboratory evaluation was conducted in a fully equipped usability lab over a period of three days, with six participants who in total identified 36 usability problems. Thus the relative amount of resources required to conduct this evaluation is somewhat higher than the resources spent on the heuristic inspection. Especially the man-hour spent on planning the evaluation in detail and scheduling times for the participants were a resource demanding task.

Advantages

The most common theme in the laboratory evaluation was *navigation* and the problems related to this theme constitute 25% of the problems that were identified using this method. The problems related to this theme are characterized by poorly designed structure and flow between dialogs and windows, like the issue related to the back button inconsistency, that we mentioned earlier. The participants in the laboratory evaluation sessions paid much attention to this aspect of the system, which is evident if we take the total number of identified problems belonging to this theme into consideration. Out of the 11 problems in total only 2 were not identified in the laboratory evaluation. This indicates that the laboratory evaluation as a method might have an advantage over the other methods when it comes to identification of *navigation* problems.

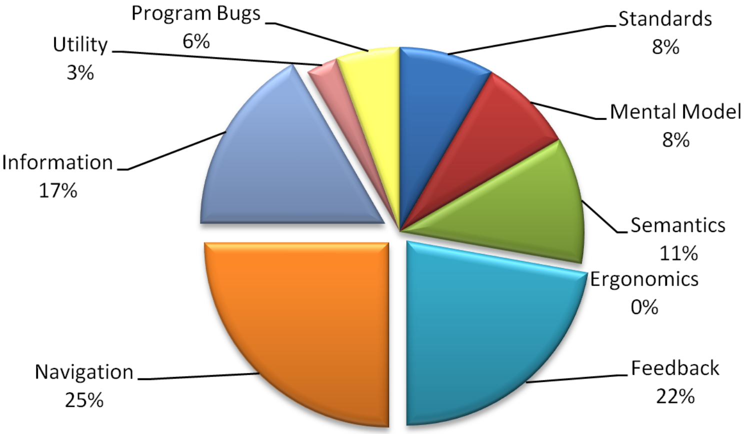

Figure 11 The distribution of problems over themes identified in the laboratory evaluation.

Likewise are 22% of the problems indentified in the laboratory related to *feedback*. The problems related to this theme are typically problems that are caused by the absence of feedback or poorly phrased feedback, which in both cases result in confusion and sometimes frustration. An example of this is the "data network error" that the users receive if EQO Mobile® fails to connect to the EQO server. This error message is phrased in technical language and offers no explanation of the nature of the problem, which renders the participants clueless and confused. It is interesting to note that this problem was not identified by the experts in the heuristic inspection, which could be ascribed to the participating experts' background in computer science. Another example that can serve to illustrate the problems caused by lack of feedback is when the users minimized EQO Mobile® and later returned and experienced that their active sessions had been closed and that they had lost connection to the EQO and MSN networks. In this case the technical challenge of keeping the connections alive is not the issue, rather the problem is that the users are never warned about this behavior and thus cannot take any precautions.

The third most common theme, that constitutes 17% of the identified problems in the laboratory, is *information*. The theme covers problems that relate to poor information access and availability. A very common problem related to this theme, is the poor access to information regarding the online status of the user. This problem typically left the participants wondering, but in most situations the participants easily managed to get by without the information, and therefore most of the problems of this type are rated cosmetic. In general it seems that the participants in the laboratory evaluation paid more attention to the availability of information than the participants involved in the other methods. Thus the laboratory evaluation might have a slight advantage over other methods when it comes to identifying this type of problems.

Disadvantages

The laboratory evaluation and the heuristic inspection have very similar drawbacks. Like the heuristic inspection the laboratory evaluation is at a disadvantage when it comes to identification of problems related to utility and ergonomics. Apparently, the participants focus on the interface itself and not as much on the overall use of the system. This might be caused by the laboratory setting that differentiates the situation from everyday use. Even though the evaluation sessions were followed by a short interview, where the participants could express their overall impression of the system, none were able to identify problems related to the physical use of the system and only half of them expressed that they had issues with the utility. The only concern of the latter group was the time it took to connect to both networks would hinder impulsive use of the system which in turn would reduce the utility.

The Role of Laboratory Evaluation in the Mobile Paradigm

Laboratory evaluation has been a de facto standard in the desktop paradigm, but when it is applied to evaluate mobile systems, it seems that it is somewhat limited by the poor resemblance with the natural use situation. Even if the laboratory setting would be modified to simulate an actual use context, we would be confined to specific context and thus the idea of evaluating the natural use of the system is only half way realized, since EQO Mobile® as previously described is multi contextual.

Despite this limitation the laboratory evaluation is still suited for identification of flaws in the interface, especially those related to the themes: *navigation*, *feedback* and *information*. The choice of whether or not to simulate context, should be individually assessed on a system basis, since some mobile systems might have more clearly defined context, such as mobile systems for use in hospitals. If this is the case, evaluations with simulated contexts in a customized laboratory setting would properly suffice, but since we have chosen a case with multiple potential contexts of use we would need several setups. It is our assumption that the procedure of identifying the proper contexts to simulate and then recreating these in the laboratory would render the method inefficient.

Field Evaluation

The field evaluation was carried out with focus on increasing the realism of the evaluation. It was an attempt to focus on realistic use of the system and recreation of a realistic evaluation environment. Thus the intention was not to retain full control of the use situation, but instead let the environment affect the participants and their use of the system in a natural way. This meant that unforeseen events were welcome in order to explore the usability problems this would uncover. As mentioned previously the method revealed the same amount of problems as the laboratory evaluation, but the distribution of severities and themes is somewhat different. The resources spent on this evaluation, is less than what was spent on the laboratory evaluation, which is a result of the less amount of time spent on each evaluation session. The method itself is quite similar to the laboratory evaluation, which led us to expect that we would also see some similarities in the problems that we would be able to identify. However because of the realistic use setting the method was likewise expected to help us identify problems related to the context of use. In order to examine these two anticipations and shed some light on the advantages and disadvantages of the field evaluation method, we need to conduct a closer analysis of the problems that we were able to identify.

Advantages

The field evaluation excels in much the same way as the laboratory and the heuristic inspection. Similar to the laboratory the most common theme was *navigation*. In total 22% of the identified problems was related to this theme (see Figure 12). The one problem that was uniquely indentified in the field evaluation was a problem related to the structure of the procedure required to delete a contact from the con-

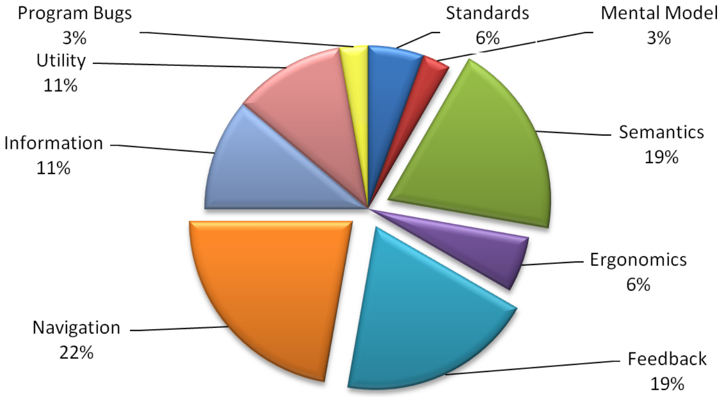

Figure 12 The distribution of problems over themes identified in the field evaluation.

tact list. The cause of the problem was that the participants thought that they had to choose a contact from the list in order to delete it, but instead they were only supposed to highlight the contact on the list and chose "remove contact" from the menu. The nature of this problem was in no way explicitly connected to the context of use and as such it does not provide us with evidence of contextual influence. The sheer amount of reoccurring problems in the field and laboratory however shows that the two methods perform quite similar when it comes to identifying problems related to *navigation*.

The second most common themes in the field evaluation were *feedback* and *semantics* which both constitutes 19% of the identified problems. Both of these themes are well covered by the laboratory evaluation as well, which underlines the similarity of the two methods. The five uniquely identified problems related to these two themes, were not related to the context of use and thus they could possibly occur in the non-contextual methods as well. For instance did the users complain about the menu items 'clear' in the chat window and 'hide' in the main menu as being difficult to interpret, since they found it hard to relate these verbs to their current situation. A similar problem related to missing feedback that occurred only in the field, was the fact that the users were returned to the EQO home menu instead of the contact list, when they successfully signed in to their MSN Messenger account. This action confused the users since they received no feedback indicating whether or not they were signed in.

All problems related to utility were identified during the field evaluation and two of those were unique for the field; One being the fact that the users did not see a point in using messenger on the mobile phone when they only were able to chat with the contacts that were online at the same time. And the other being the fact that the participants compared the use of the system to the SMS functionality of the mobile phone and therefore saw no advantage in using EQO Mobile® over something well known, trouble-free and inexpensive like the SMS service. The two problems related to ergonomics that was identified in total, were both identified in the field - both being severe; one problem being the fact that two of the participants became motion sick while using the system in the bus. The reason why this is noted as an independent problem is due to the rapid pace of the communication which forced the participants to keep a focus on the device in order to respond to new messages as quickly as they felt they would normally do at the PC. Another problem related to the pace of communication in real time chat, was the poor input capabilities of the mobile phone. Three of the participants complained that this issue severely crippled their communication. This problem is in fact very interesting because it illustrates that the input limitations of the mobile device might be okay for writing a SMS, but when it comes to real time chat like MSN Messenger it constitutes a problem. These problems are related to the larger context in which the system was used and illustrates that evaluations conducted in natural contexts of use might provide us with additional knowledge.

Disadvantages

The limitation of the field evaluation was primarily the reduced quality of our recorded data. The hand held camera and build in microphone that we used often resulted in shaky images and reduced sound quality, and under more difficult circumstances it might have proven to be an issue, however in our case the quality of the resulting video recordings proved to be sufficient. Thus evaluation in realistic contexts can prove to be a practical challenge. The most serious limitation of this field evaluation is the fact that the evaluation is still confined to one context. It is possible to evaluate in several contexts, but it might prove impossible to evaluate multi contextual systems, like mobile phones, in every possible context of use. Thus ethnographical studies might be necessary in order to determine the context in which to evaluate, which means that extra resources would be necessary in both the planning and execution phase and thus potentially lower the cost efficiency of the method.

The Role of Field Evaluation in the Mobile Paradigm

The field evaluation can be used as a way to gain insight in specific contexts and the problems that might occur here. However it is a costly and time consuming affair to conduct evaluations in multiple contexts, which limits the scope of application of this particular method. However when evaluating systems with clearly defined contexts of use, the field evaluation could allow evaluators to identify contextually related problems. This might especially be useful if the context is unfamiliar to the evaluators and developers. In such a situation the field evaluation could be used to gain insight into the domain as well as to identify usability problems. In such a case evaluators should take great care in documenting the context and the way the system is used, in order to inform further development.

Diary Evaluation

The final evaluation was conducted with focus on increasing the ecological validity of the usability evaluation setting even further. In order to do so, the participants were instructed to use the system when and where they wanted to without any form of control of their actions or interference through data collection. In order to get as realistic use of the system as possible, and gain insight into the usability problems that occur at different stages of use, we let the evaluation run over two weeks. During this two week period the researchers were free to perform other tasks, and thus we were able to conduct both the laboratory evaluation and the heuristic inspection in this period. Despite being the method running over the largest span of time it required the least amount of man hours to process the data. During the two week evaluation period we found that the users used the system in a wide range of contexts varying from the tranquility and comfort of the couch to the hustle and bustle of public transportation. We found that not only the context of use, but also the purpose of use was subject to variations. It was for instance evident that the users used the system for different purposes such as: to stay connected in environments where no internet access where available through landlines, to conduct short and planned conversations with MSN Contacts as an alternative using SMS messages and as a way to conduct longer and unplanned conversations with whomever wanted to communicate at the current time. Thus the freedom of the setting allowed the user to explore the system in a wider range of social and spatial contexts of use. In order to determine the trade-offs of the method and be able to evaluate the advantages and disadvantages we need to take a closer look at the distribution of the identified problems and their corresponding themes.

Advantages

The diary evaluation identified 15 usability problems, and thus proved to be the method identifying the least amount of problems. Still it is worth noting that 9 out of the 15 problems were not identified with any other evaluation method, and that 4 of these unique problems were rated as critical problems. These four problems constitute a serious hindrance to the usability and success of EQO Mobile® and thus they should be identified and dealt with prior to releasing the system. For instance the problem related to *feedback*, which in fact was the only feedback related problem that was rated as critical, was the lack of feedback when a message could not be correctly delivered to the receiver. By not informing the users of erroneous transmissions, the communication could be severely crippled which would decrease the usability of the system.

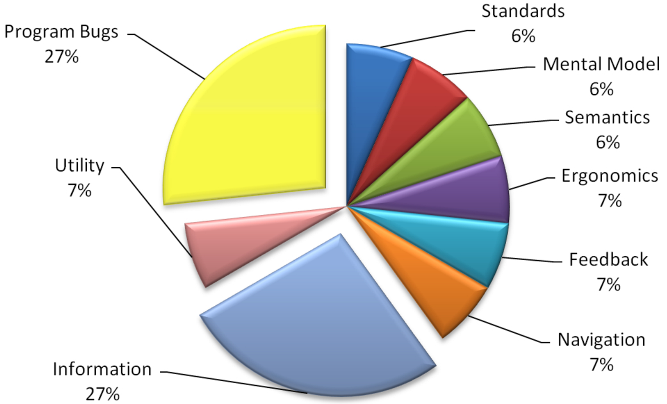

Figure 13 The distribution of problems over themes identified in the diary evaluation.

One of the two most common themes in the diary evaluation proved to be *program bugs* which constituted 27% of the total amount of identified problems in the method (see Figure 13). The same amount of problems were identified as related to the theme regarding availability and accessibility of *information*, but the program bugs were all but one rated as critical which makes this theme stand out. One of these critical program bugs was that the system would freeze randomly, forcing the participant to restart the phone. This program bug was likewise identified in the laboratory and the field evaluation. One critical bug that was not identified in any other evaluation was that the system sometimes influenced the functionality of the phone forcing the participants to restart the phone. For instance one participant experienced that his phone was unable to display images after using EQO Mobile®. Whether or not these program bugs could have been identified using other methods is not clear, but they seem to occur randomly, and thus, due to the longer evaluation period, it seems more likely that they will occur and thus be identified with this method.

The problems related to the availability and accessibility of *information* likewise constituted 27% of the total amount of identified problems in the diary evaluation. Four of the problems related to this theme were uniquely identified using this method out of which one was rated as a critical problem and two as severe problems. For instance when network coverage was lost, the user was returned to the main menu and any active session he might have at the current time would be lost. This meant that messages received just before the network coverage was lost, were not necessarily read by the participant. It was encountered by 3 of the 4 participants and was assessed to be a critical problem. Another *information* problem that none of the participants in the other evaluations encountered was that MSN contacts with similar names could not be told apart in the contact list because neither their email address nor their alias were visible. Likewise none of the participants in the other evaluations noticed that when they had conversations longer than what could fit in a single screen, a scroll bar would appear to the right covering underlying text, meaning some words could not be read.

Disadvantages

The diary evaluation was conducted over a two week period, which gave the users a chance to get acquainted with the system. Thus the users identified a different set of problems – in general they identified problems that continued to bother them, rather than trivial problems that they learned to deal with. For instance, it is remarkable that the diary evaluation revealed significantly less problems related to interface issues than any other evaluation method. Instead the users focused on the more serious problems, such as the four identified program bugs that proved to be a hindrance to their use of the system.

This focus is also evident if we look at the distribution of problems over severity, where the diary evaluation, unlike any of the other methods, identifies more critical problems than severe and more severe problems than cosmetic. This distribution supports the assumption that participating users focus on the problems that continue to bother them and ignore the less severe problems. This issue can be seen as a trade-off of the self reporting technique that we chose to use in the evaluation, since the diaries give us a poor insight into the problems that users quickly overcome and thus assign little or no relevance to. These problems would typically be cosmetic problems, that could be easily dealt with and thus improving the overall user satisfaction and flattening the learning curve of the system.

In retrospect, the decision not to control the users or their use of the system has influenced the results of the evaluation as well. It is our belief that the absence of tasks resulted in identification of fewer usability problems, simply because the users where able to get away with using only parts of the system and avoid other parts entirely. We therefore believe that the decision to pursue realism through freedom of use, as an alternative to structured usage might be futile, since it results in poor insight in the usability problems that might exist in the parts of the system that the participants neglects.

The Role of Diary Evaluation in the Mobile Paradigm

Diary evaluations should be considered when dealing with mobile systems that are intended to be used in multiple and possible unknown contexts. Under such circumstances the freedom to use the system where and when the users see fit would enable them to identify problems related to a variety of contexts. However, the data collection technique that we chose to apply seems somewhat inefficient, because it leaves it up to the users to identify the usability problems, which might cause the participants to focus on the major issues and forget the minor problems. If for instance video recordings of the use existed such a problem would be minimized, but then again such an action would certainly also affect the participants' use of the system. However this is a compromise that could be necessary in order to collect the data needed for later analysis.

Because of the extended period that the diary evaluation is conducted over, the utility of the method can be limited when applied to development projects where rapid iterations are the norm. Rather it must be used as a method that can provide the developers with insight into not only the usability problems that occur in a wide range of contexts, but also into the patterns of use in the various contexts themselves. The explorative nature of the diary evaluation holds some advantages that might make it more suitable for new concepts, such as mobile Messenger applications, but extending into other realms as well.

Comparing the Characteristics

As a result of our analysis, we are now able to present the characteristics of the four methods that we have applied. Furthermore in order to answer the research question that we set out to answer in this chapter we will focus on the characteristics of the contextual methods as opposed to the non-contextual methods.

Before we give a condensed description of the abilities of the methods we would like to sketch our experiences with the individual methods.

Table 7 serves to show that it is not only the problems that we were able to identify with the individual methods that are of interest, but rather all the aspects that practitioners and researchers alike should consider when applying the evaluation methods.

The heuristic inspection is characterized by being able to identify many of the cosmetic problems related to the system interface within a relative short amount of time. Thus the method should be applied to this specific purpose, and avoided when the evaluators suspect that the system may contain other types of problems. The prmary advantage of the inspection approach and the feature that makes it interesting

Evaluation Method	Practical disadvantages	Resource consumption	Type of problems that the method typically identifies	Problem types that the method seldom identifies
Heuristic inspection	Requires the participation of 4 HCI experts	15 man hours during a single day	Cosmetic problems related to interface issues	Problems related to the context of use and other aspects such as utility and ergonomics
Laboratory evaluation	The need for a laboratory or similar suitable location	53 man hours during a three day period	Problems related to interacting with the system; primarily interface issues	Problems related to the context of use and other aspects such as utility and ergonomics
Field evaluation	The context can complicate the data collection process	41 man hours during a three day period	Problems related to interface issues as well as problems related to the specific context	Problems related to unmet expectations of the users' to the functionality of the system
Diary evaluation	The time span of the evaluation and the fact that the users do not necessarily use the entire system	26 man hours during a two week period	Problems related to the context of use and problems that only occur during extended periods of use.	Problems related to interface issues in general

Table 7 Summary of our experiences with the methods.

is that it can be conducted quickly and under almost any circumstance. The only requirement that has to be met in order to conduct this evaluation is that the proper experts are available and that they are equipped with the system and the set of heuristics that will form the foundation of the evaluation.

The laboratory evaluation is a thorough method that will allow the evaluators to identify a wide range of problems, but primarily problems related to the interaction with the system without respect to the context in which the system is intended to be used in. The results derived from the laboratory evaluation will most likely be reproducible under similar circumstances, which provide the evaluators with knowledge on the frequency of which the problems will occur. This knowledge can be useful when the evaluators or developers shall decide which problems are to be corrected and in which order. In order for the method to be applied correctly a proper laboratory or similar suitable facility must be present.

The field evaluation can be seen as a variant of the laboratory evaluation, where the controlled environment is forfeited in favor of a realistic context of use. The similarity in the methodological procedure is reflected in the amount of problems identified with the method and the severity these problems. The methods differ in two important aspects. Firstly the field evaluation has an advantage over the laboratory evaluation when it comes to identifying problems related to *utility* and *ergonomics* and secondly the field evaluation is conducted in a shifting environment which affects reproducibility of the problems. The primary practical concern that the evaluators should be aware of when conducting a field evaluation is that it can be tricky to conduct a usability evaluation with cameras and crew in public places. On several occasions we had to leave the bus before we even started the evaluation due to lack of available seats.

The diary evaluation was unlike any of the other evaluations conducted as a longitudinal evaluation over a two week period, which might be perceived as a serious drawback in situations where fast results should be obtained. The evaluation yielded some interesting results and provided us with knowledge of several critical and severe problems that could not be identified by any of the other methods. Most of these problems were closely related to the extended period of time that the users spend with the system. Thus the longitudinal aspect of the evaluation allows us to reveal a special subset of problems. Likewise it was evident that the context influenced the usability of the system which allowed us to identify a couple of contextual dependent problems as well. However, despite the fact that the diary evaluation provided us with new and relevant knowledge it seems to lack the ability to identify the most simple interface problems. This is a drawback that we subscribe to the data collection technique and the freedom that allows the users to avoid parts of the system entirely.

Contextual vs. Non-Contextual

These characteristics of the different methods can be summed up by outlining how the contextual and non-contextual methods differ.

As we argue in the analysis of the contextual evaluations, the users are inclined to identify problems that relate to their own use of the system in the given context, whereas the users and experts in the laboratory evaluation and heuristic inspection are more focused on the system as an individual object with innate usability. This distinction is very important to make, since it can explain why the users in the laboratory do not identify problems that are directly related to the context provided by the laboratory itself. It is evident that the problems identified in the laboratory are strictly related to the users' interaction with the system instead of the interaction with the system in the context of the laboratory.

Even though this observation is based on only a small fraction of the problems that we identify, we find that the laboratory evaluation and the inspection provide an environment where the system and the objective of evaluating the innate usability of the system is the highest priority. Because of this heightened focus on identification of problems within the system it is clear that the laboratory and the heuristic inspection will excel in identification of problems related to the system interface.

Likewise contextual methods might have a slight disadvantage when it comes to identifying problems related to the interface. It is likely that the context might distract the users and influence them in such a way they focus more on the utility of the system in the given context, than on identifying problems related to the interface. Thus the results of contextual evaluations will contain a larger amount of contextually related problems and fewer interface problems. This hypothesis is supported by the characteristics and results provided by the applied methods.

In order to compensate for the weaknesses of the contextual and the non-contextual methods it will either be necessary to construct new methods that draw on advantages from both sides, or to use several methods in combination. In the following chapter we will try the first in an attempt to device a more suitable method for evaluating the usability of mobile systems.

Answering the Research Question

The above analysis and characteristics of the four applied methods allows us to answer the research question that we set out to answer in this chapter. It seems that the four methods can be applied in order to identify different problems and that they all are subject to certain limitations. In general we can conclude that the heuristic inspection and the laboratory evaluation are suitable for identification of interface problems and are limited when it comes to identifying contextually related problems. Likewise the field evaluation and the diary study have proven to be able to identify a number of problems closely related to the use of the system under realistic circumstances. This short description serves to illustrate the difference between contextual and non-contextual evaluations.

However there are individual practical differences as well that we need to account for. The heuristic inspection for instance is at an advantage when the need to perform fast and inexpensive evaluations are the top priority. Likewise the diary evaluation seems suitable if a contextual evaluation is required in the frame of a tight budget. The field and laboratory evaluations should be perceived as more complete, based on the fact that they identify a larger group of problems, but in turn they have higher resource consumption, which might make them less attractive. The field evaluation performs quite similar to the laboratory evaluation but reveals more contextual problems and fewer interface problems, which should be considered when choosing a method to apply.

A New Evaluation Method

In the following chapter we will describe how we constructed a new usability evaluation method and account for the objectives and techniques we applied to meet these. After having described and argued for our choices we will present and analyze the results that we were able to derive from applying the method and compare the new method to the previously applied methods.

This will serve to answer the third research question:

> Research question #3: *How can we utilize the first-hand knowledge that we have gained and devise a method that overcomes the challenges imposed by the nature of mobile systems?*

Objectives

As we documented in the previous chapter none of the methods stand out as the one best way to evaluate the usability of EQO Mobile. However the problems identified with the contextual methods indicate that there is additional knowledge to be gained from applying contextual methods and therefore we wished to construct a single contextual method that is able to provide us with this insight.

Thus the primary objective will be to construct a more suitable contextual method that will allow us to identify the problems that are closely related to the context of use and the problems that are only discovered during extended periods of use, while at the same time allowing us to identify problems related to the interface of the system. It is therefore our objective to devise an evaluation method that is better suited for evaluating mobile systems than the current methods.

We will base this new method on a foundation provided by the diary evaluation and seek to implement techniques that can help us identify the problems that were either neglected or simply not encountered during the diary evaluation.

Thus our objectives can be summarized:

- Strive for ecological validity by allowing the system to be used freely in all possible context and over an extended period of time

  While at the same time:

- Improve the data collection technique so volatile knowledge of user experiences is not lost
- Ensure that all parts of the system are subject to use, so that no potential problems are hidden from the users.

These objectives are inspired by the fact that we observed indications that the high level of realism led us to identify very interesting usability problems, but that the low level of control apparently allowed for shallow usage of the system. The lack of control over the way the participants used the system enabled them to ignore parts of the system and therefore problems related to those parts could not be identified. A solution to this problem could be to apply a technique from one of the more control oriented evaluation methods that encourage the participants to use all parts of the system. We therefore included the use of predefined tasks in the same way as we did in the laboratory and field evaluations. By doing so, we could ensure that the participants used all parts of the system.

Likewise we aimed at improving the data collection technique because the combination of diaries and interviews that we applied in the diary evaluation, allowed the users to forget or overlook usability problems. Hence it was impossible for us to uncover these problems. In order to remedy this problem we therefore wanted to apply techniques that allowed us to objectively collect volatile data about the usage of the system, so that we later could subject it to review and analysis. In order to pursue this objective we opted for the use of digital video cameras as a measure to retain the use of the system.

By applying predefined tasks and video recordings, we were forced to compromise the realism of the evaluation. However, we sought to minimize this unfortunate influence by applying the techniques in a way that allowed for natural and unsupervised usage as well as structured and video recorded usage. The methodological procedure and the practical approach to implementing the techniques are described in the following.

Procedure

The methodological procedure of this evaluation was inspired by cultural probing. Cultural probes are often applied as an ethnographic approach used in order to inform the design of new technology and software (Gaver, Dunne and Pacenti 1999), but in this case we intended to apply it with a different purpose. Instead of using probes to inform new designs, we intended to use the probing technique to evaluate an existing piece of software and thereby inform the redesign of the system. Thus we decided to define the new method as the **video probe evaluation**.

The users in the video probe evaluation were chosen from the same selection criteria as the users in the previous evaluations that involved real users, but this time we chose to conduct the evaluation with couples. The reason why we chose to use couples was that it would ease the data collection procedure, since one of the partners could operate the camera while the other could perform a task or describe an encountered problem. Secondly, we expected that couples would use the system to communicate on a daily basis and encourage and motivate each other to use the system, which would lead to further exploration and more realistic use over time.

We decided to equip each couple with digital video cameras, so that they could document the usability problems they encountered over the period of use. In the spirit of cultural probing, we packed the cameras in a small box containing all the material needed for the evaluation (See Figure 14). We anticipated that the novelty of the approach in itself would inspire and motivate the participants to use the system and collect valuable information regarding the usability of the system.

The only drawback of conducting the evaluation with couples was that it proved virtually impossible to find couples with matching phones, thus we decided that we instead would make sure that at least one of the partners were equipped with a SE K750i (the model used in the other evaluations). By letting the partners use their own phones we were aware that we potentially had polluted our evaluation setup. So in order to remedy this and avoid potential false positives we decided to double check and if possible reproduce all problems that occurred on the unauthorized phones on a reference SE K750i phone during the analysis of the problems the users encountered.

The evaluation kit consisted of the following materials (See Appendix E):

1. A digital camcorder and tapes (enough for 3 hours of continuously filming).
2. A description of the purpose of the evaluation with instructions on how to film each other when solving the tasks.
3. Numbered envelopes with the tasks and instructions.

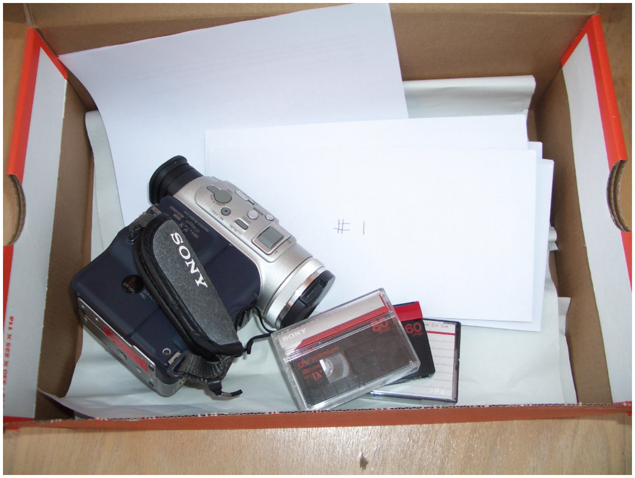

Figure 14 The video probe equipment kit

Prior to the evaluation we constructed a time table (see Figure 15) in order to get an overview of the evaluation and enabling us to prompt the participants to open the envelopes containing the tasks in due time. The time table also helped us schedule time for completion of the third task that required the participants to engage in conversation with a fictive contact controlled by us - the evaluators. This conversation was used as a secondary data source and allowed us to collect preliminary information about the participants' usage of the system and their experiences so far, and at the same time gave us a chance to encourage them to use the system.

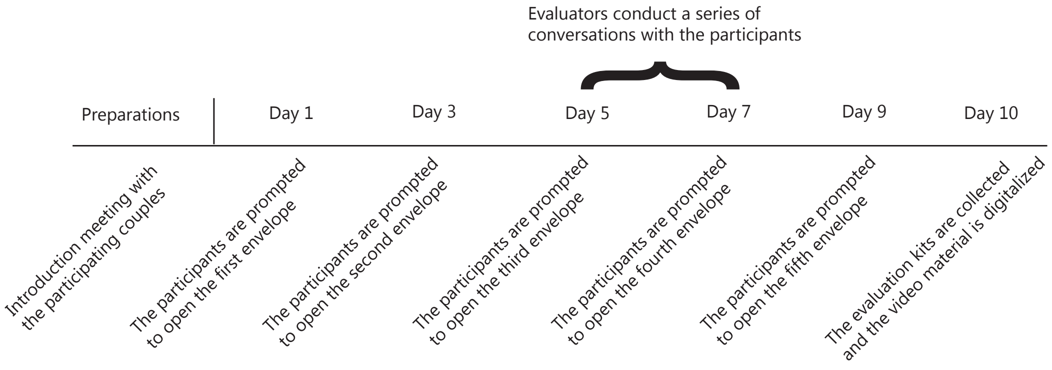
Evaluators conduct a series of conversations with the participants
Preparations
Day 1
Day 3
Day 5
Day 7
Day 9
Day 10
Introduction meeting with the participating couples
The participants are prompted to open the first envelope
The participants are prompted to open the second envelope
The participants are prompted to open the third envelope
The participants are prompted to open the fourth envelope
The participants are prompted to open the fifth envelope
The evaluation kits are collected and the video material is digitalized

Figure 15 Time table illustrating the evaluators' overview of the evaluation sessions

Before we began the evaluation, we visited the participating couples and made sure that they had EQO Mobile up and running and they understood the purpose of the evaluation and were able to use the video camera that we had provided. After the introductions were given, the evaluation was officially initiated by prompting the couples with a SMS asking them to open the first envelope containing the first task. The first task was designed to make the users setup a MSN Messenger profile on EQO Mobile so they were able to start their exploration of the system. Along with the task were instructions on the expectations we had to their use of the system during the next couple of days.

Likewise, the rest of the envelopes in the evaluation kit contained predefined tasks that the participants were to solve during the evaluation. These tasks resembled the tasks that were given to the participants in the laboratory and field evaluation with slight modifications. The reason why we chose to include tasks in the evaluation was due to the objective of ensuring that the users explored the same parts of the system as the ones explored in the laboratory and field evaluation. This way we were confident that the participants would not neglect or ignore essential parts of the system and thus reducing the potential outcome of the evaluation unnecessarily. The remaining tasks were to be solved one at a time during the ten day evaluation period. However, the participants were not given any instructions on when they were to solve the tasks. Instead we informed them that they would be prompted with an SMS approximately every two days with instructions. In this way we hoped that we could stimulate the curiosity of the participants and preserve their interest in the evaluation.

At the end of the evaluation we revisited the couples and collected the evaluation kits. The video material was digitalized over the next couple of days and both researchers were given a copy for further analysis. Like in the previous evaluations we analyzed the video material individually and produced a list of problems which later was merged through discussion with reviews of the video material as supporting data. Thus we ended up with an agreed upon set of usability problems that were rated by severity and clustered in themes.

Results

In order to determine whether or not the video probe evaluation is suitable for evaluating the usability of mobile systems, we will need to take a closer look at the results that we have been able to achieve with this novel approach. First we will try to establish a general overview of the amount and severity of the problems that were identified using this method. Furthermore we will look at the severity of the problems and the distribution of problem themes and compare the findings to the ones derived from the other methods. When a general overview has been established we will focus on the tradeoffs of the video probe evaluation and try to depict advantages and disadvantages, in order for us to be able to compare the method to the methods from the previous chapter.

Overview

The video probe evaluation identified a total of 37 usability problems distributed on 4 critical, 8 severe and 25 cosmetic usability problems (see Table 8).

	Inspection	Laboratory	Field	Diary	Video Probe
Critical	2	3	4	6	4
Severe	6	11	9	5	8
Cosmetic	20	22	23	4	25
Total	28	36	36	15	37

Table 8 Identified usability problems including severity assessments.

The distribution of problems identified with this method compared to the other methods are illustrated in Figure 16, which shows that many of the critical and severe problems identified in the video probe evaluation overlap with the problems identified using the other methods. Surprisingly we see that half of the cosmetic problems identified with this method are unique.

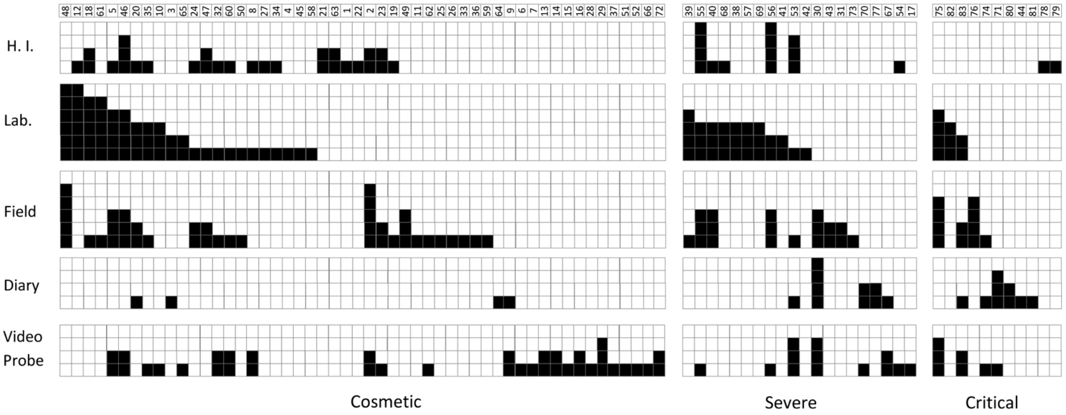

Figure 16 Distribution of usability problems based on severity.

If we take a closer look at how the usability problems are distributed over the nine themes, we see that the dominating themes in this method are *mental model*, *navigation* and *information* (see Figure 17).

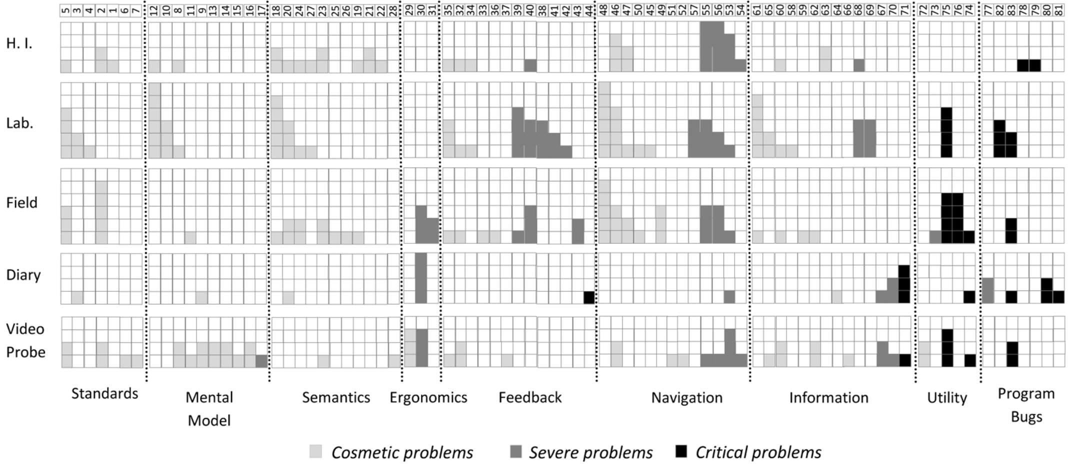

Figure 17 Distribution of usability problems based on themes.

Compared to the three non-longitudinal methods the video probe clearly identifies less problems related to *semantics* and *feedback* and remarkably more problems related to *mental model*. It performs quite similar when considering problems related to: *standards*, *navigation*, *information* and *program bugs*. If we consider the amount of problems related to *utility* it is outperformed by the field evaluation but it is still better at identifying problems related to this theme than the heuristic inspection, the laboratory evaluation and the diary evaluation.

Furthermore when comparing the video probe evaluation to the diary evaluation we see that the three *information* problems that were unique for the diary evaluation were also identified in the video probe evaluation (problems 67, 70 and 71). This might be because they are rare occurring problems that are more likely to be identified during longer periods of use, but we will return to this in the analysis.

	Inspection	Laboratory	Field	Diary	Video Probe
Planning	6 h	10 h	10 h	14 h	20 h
Execution	3 h	12 h	9 h	10 h	9 h
Processing	6 h	31 h	22 h	2 h	24 h
Time spend	15 h	53 h	41 h	26 h	53 h
Duration	1 day	3 days	3 days	14 days	10 days

Table 9 Time spend on different stages of the video probe evaluation and the previous evaluations.

If we look at the resources spend on the video probe evaluation (see Table 9) we can see that we have spend a total of 53 man hours on planning, executing and processing the evaluation which makes it just as resource demanding as the laboratory evaluation. The reason why this method took the longest time to plan is both because it is a novel method that was constructed more or less from scratch and that the procedure of finding suitable participants proved rather troublesome.

Analysis

Now that an overview has been established we look at the advantages and disadvantages of the video probe evaluation before we compare it to the previously applied evaluation methods.

Advantages

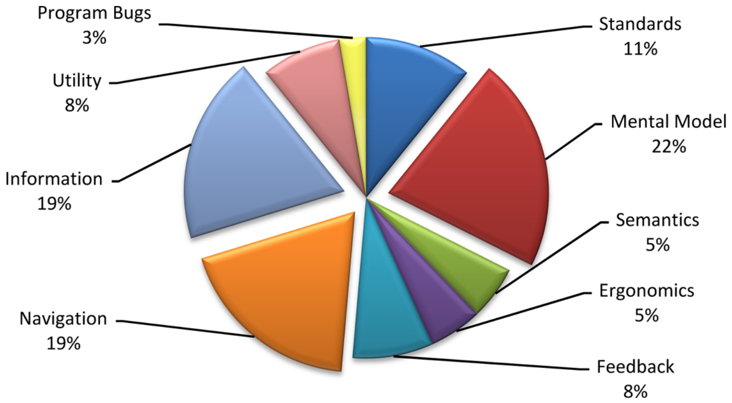

Figure 18 The distribution of problems over themes identified in the video probe evaluation.

The most common problem theme in the video probe evaluation was the theme *mental model*. The problems related to this theme constitute 22% (see Figure 18) of the total amount of usability problems identified during the evaluation, which correspond to 8 of 10 problems related to this theme identified by all methods (see Figure 17). The reason as to why we see this apparent advantage of the new method can be related to the self reporting technique that we applied. When the users recorded themselves, they took time to reflect on their own usage during the previous days as well as to solve the tasks that we provided. Thus the video material contained many reflections and perspectives on the system as an independent application, but also as a mobile variant of MSN Messenger. This is evident if we take a closer look at the problems related to the mental model theme, which all are related to expectations caused by prior experiences with MSN Messenger. For instance, several users expected to be able to change their display names, send offline messages and create new MSN accounts. All problems resulting from unfulfilled expectations caused discontent and confusion, but when the users realized that their perception was wrong, they tend to accept the shortcomings of the system and adapt their usage.

Another dominating theme is *navigation*, which constitutes 19% of the total amount of problems identified with the video probe evaluation. This theme is also one of the three most common themes in the heuristic inspection and the laboratory and field evaluations, and compared to these the video probe evaluation performs quite similar – especially with respect to severe problems. It is not the method identifying the most navigation related problems, but it identifies all but one of the severe problems related to this theme, and three cosmetic problems in addition. Compared to the diary evaluation this method is effective when it comes to identifying problems related to this theme, which most likely is due to the detailed data provided by the video recordings. This is supported by the fact that, many of the problems related to this theme were not explicitly noted by the participants, but were identified by the evaluators during the analysis of the video material. Likewise we believe that the problems encountered by the participants in the evaluation based on the video probes, could also occur during the diary evaluation, but since the participants did not perceive them as problems they were left out of the diaries. Whether or not the problems are verbally accounted for by the participants or identified later by evaluators, they should be noted as usability problems because of the confusion that they cause.

The problems related to the theme *information* likewise constitute 19% of the total amount of usability problems identified during the video probe evaluation. Out of the seven identified problems that relates to this theme only one problem was uniquely identified using this method. Just like the diary evaluation, the video probe evaluation allowed us to identify two severe problems and one critical, which were not identified by any of the non-longitudinal evaluations. These problems are interesting because they represent the type of problems that are most likely to occur in longitudinal evaluations (the diary and video probe evaluations in this case). One of these problems is the fact that each contact is represented by only eight characters in the contact list, which does not allow the users to tell contacts apart that share the same first eight letters in the display name. This problem occurred rarely and was only noted by the participants in the longitudinal studies, which support the claim that problems that only occur rarely are easier to identify over a longer period of time. Therefore the longitudinal evaluations allowed a rare problem such as this to occur several times, which made the users notice it and become irritated and frustrated enough to account for it during the camera sessions. Another longitudinal-specific problem was the poor representation of newly received messages, which forced the users to continuously keep eyes on the ongoing conversations in order to be aware of new messages. This problem is only identified in the longitudinal evaluations because the users in the non-longitudinal evaluations never had to put the phone down and wait for a contact to write. This problem shows how the realistic use situation in the longitudinal evaluations can provide us with additional information. The last longitudinal problem was a critical problem concerning chat sessions being shut down without warning or opportunity of saving when the phone lost network coverage. This is likewise a consequence of the freedom the users had to use the system as they liked.

Another advantage of this method that seems to be caused by the environment in which it is conducted is the method's ability to identify *utility* problems and *ergonomic* problems. These types of problems are almost exclusively identified when applying evaluation methods in context, except for the one *utility* problem that the laboratory evaluation allowed us to identify. The one problem that was also encountered in the laboratory was the problem with long connection times, which according to the participants, rendered the system useless for short and impulsive conversations.

Disadvantages

One of the drawbacks of the video probe evaluation is the poor insight into problems related to *semantics* and *feedback*. The five cosmetic problems related to these themes that we were able to identify with the video probe evaluation only constitute 5/24 of the total amount of problems that we have identified with all the methods, and thus it seems that this method is unfit for identifying these types of problems. The severity of this drawback becomes clear if we look at the feedback related problems identified by the other methods. Both the diary evaluation and the non-longitudinal evaluations identified severe and critical problems related to this theme, which we did not identify with the video probe evaluation. Therefore it seems that the video probe evaluation is insufficient when it comes to identifying this type of problems.

Another drawback is the seemingly poor ability to identify program bugs, however the nature of the problems belonging to this theme is volatile, and most of the problems occur at random which makes it a poor indicator of the methods abilities. However, as we have argued before, this method should have a slight advantage over the non-longitudinal evaluations in this area, since it runs over a longer time span and thus the chance of identifying such problems should be increased.

Besides the obvious drawbacks, related to problems that were hard to identify, the method has a potential drawback related to a technical issue. The quality of the video recordings were, in some cases, poor compared to those of the laboratory and field evaluation, and could have proven an obstacle to our analysis if it was not supported by high quality audio data. Therefore we find it very important to inform the participants in such an evaluation about the practical use of the camera and proper lightning conditions, while at the same time emphasize that they should comment on their actions and try to continuously think aloud or engage in dialog with their partners during the video recorded sessions. By ensuring a good video quality we might be able to retrieve even more information from the evaluation.

Comparing the Video Probe Evaluation to the other Methods

In general we have experienced that by using video cameras in the self-reporting data collection process of the longitudinal evaluation we are able to increase the amount of useful data that can be derived from the evaluation. The tasks that were given were devised to structure the use of the system so all parts of the system's functionality was used. When reviewing the recorded video material, it seems that this approach helps us identify many similar problems as in the laboratory and field evaluation where a similar set of tasks was used. It is however evident that the participants do not always verbally account for the problems they encounter, despite they sometimes were regular usability problems. In these cases the video recordings allowed us to identify these unspoken problems. Furthermore when solving the assigned tasks, the participants often took time to illustrate other problems that had bothered them, but were not directly related to the functionality that they evaluated at the time.

Answering the Research Question

Recalling the third research question, we have now documented how we have used our first-hand knowledge gained from the previous chapter to devise a contextual usability evaluation method that overcomes the challenges that mobile systems impose on usability evaluation methods.

The video probe evaluation combines some of the best characteristics of the contextual and non-contextual methods. By applying some of the structure and the data collection technique of the latter methods in a longitudinal method, we have been able to identify problems related to the use of the system in ecological valid contexts over a period of time, while at the same time maintaining the structure and insight provided by the tasks and video recordings.

However the method should not be perceived as a superior method for evaluating mobile systems, since it lacks some of the strengths of the other methods. But it is an example of how techniques could be combined in a single method such that the nature of the mobile system is embedded in the evaluation method. It is possible that a *one best way* can be constructed, but according to our results neither of the five evaluation methods that we have applied can be labeled as such.

Discussion

In this discussion we will process the data that we have collected differently than we did in the precious chapters. Initially we will try to identify the methods that identify the most significant problems in order to see, whether or not this new perspective can support the claims that we have made so far. Secondly we intend to rearrange our data in order to elaborate on the potential of using different methods in combination, which we hope will shed some new light on how the usability of a mobile system should be evaluated. Finally we intend to elaborate on the obstacles and threats that can prove a hindrance to wide spread use of contextual methods.

The Most Significant Problems

Identification and correction of the most significant usability problems are crucial for the success or failure of any given system when releasing it. Therefore it must be imperative to identify the problems that would improve the quality of use most significantly if they were corrected before the actual system is released.

Since our previous attempts at identifying a superior method for evaluating the usability of mobile systems has been futile – or ambiguous to say the least - we will try to reduce the complexity of our findings and point out the most significant problems with EQO Mobile that if corrected would add the most value to usability of the system. By doing so we assume that a method will stand out as the best one at identifying significant usability problems.

Although we have assessed the severity of all the problems identified with the five different methods and based much of our previous work on these severities, we acknowledge that the severity ratings we have assigned to each problem might not correspond directly to how the *quality of use* of the mobile system is affected by correcting the problem. For instance, what is the most important problem to correct if one problem is critical, yet rarely occurring, and the other is cosmetic and frequently occurring?

For now we will not elaborate any further on this question, but it serves as an example of what choices should be made when receiving a list of usability problems and what we as producers of such lists should consider when presenting them to developers. Obviously it comes down to identifying and ensuring the correction of the most significant problems with the system and the question is therefore how to know which problems are more significant than others and how they are to be identified.

In an attempt to point out the most significant problems we called upon two fellow HCI master students to act as experts. We asked them to rank the ten usability problems that would increase the usability of EQO Mobile the most if they were to be corrected. Furthermore the experts were instructed to consider the system and its problems from a holistic perspective. The experts were then presented with the complete list of usability problems and were asked to select the ten most significant problems and rank them from 1 to 10 (1 being most important). For instance if they believed that fixing problem number 33 would have the greatest impact on the overall usability of the system, they were to rank the problem as #1. Initially they did this individually and when finished they were asked to merge them into a final ranked list, which they both could agree upon represented the ten most significant usability problems with the system (see appendix F for the procedure and Figure 19 for the results).

When reviewing the individually produced lists of the ten most significant problems we saw that the experts were in complete agreement on 7 of the problems that made it to the final list. Each expert therefore ranked 3 problems that did not make it to the final list. Out of these problems one problem appeared on both of the experts lists, but did not make the final cut and was discarded during the merging.

Figure 19 depicts how the ten most significant problems were distributed over the five different evaluation methods that we applied. The figure can be seen as a reduced version of Figure 16 in the previous chapter. And as Figure 19 shows, none of the methods identified all of the significant problems.

The heuristic inspection only identified 1 of the 10 most significant problems - making it the least preferable method when it comes to identifying significant problems. The laboratory evaluation identified 3 of the problems including the most significant one, which indicates some limitations in its ability to identifying the most significant problems. The field evaluation identified 5 significant problems including the most significant one. It also identified the 3rd, 4th, 5th and 6th most significant ones making it rather solid despite the fact that it didn't identify all significant problems. The diary evaluation also identified 5 significant problems including the three most significant ones. Finally, the video probe evaluation also identified 5 significant problems including the most significant one as well.

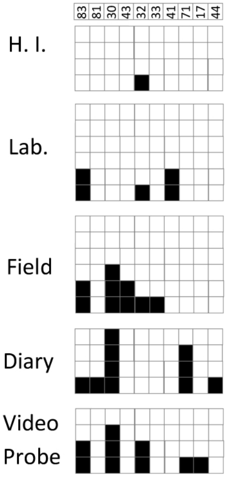

Figure 19 Top 10 of the most significant problems to correct (most significant to the left and least significant to the right).

Again we cannot sort out a single method that should be applied in order to identify all or most of the significant problems, but in order to find the most significant problems it seems like there is more to be gained from venturing in to the field with contextual methods than relying on non-contextual methods. The top 10 supports that there is additional knowledge to be gained by applying contextual methods compared to the non-contextual methods. This could very well be an indication that finding significant problems require methods with a high degree of realism.

Based on this and our prior assessments of the individual usability evaluation methods we see that no single method stand out as the one best way to evaluate the usability of mobile systems. Therefore in the next part of the discussion we will elaborate on the possibility of combining the methods with the intention of enabling ourselves to present recommendations on how to combine evaluation methods for practitioners and researchers alike.

Recommendations on Combining Methods

We have seen that all methods have different advantages and disadvantages. Thus in order to uncover as many usability problems as possible it is reasonable to consider combinations of methods instead of just combining techniques from different methods into a single method. We will do so based on our findings by comparing the possible combinations of methods when combined one and one.

In order to set the stage for comparing the different approaches we will point out the metrics that we will include in our comparison:

- Amount of identified problems
- Severity of the identified problems
- The efficiency of the combinations

The first metric is the methods combined ability to identify large quantities of problems, the second metric that we will include is the severity of the identified problems and the last metric is the efficiency of the various combinations.

Combinations

By comparing various combinations with respect to these metrics we will be able to recommend different combinations of evaluation methods. However each combination will be recommended with certain criteria in mind and thus the recommendations will only apply under the described circumstances. It is our intention that this discussion will lead to a better understanding of the various possibilities that exist when combining methods, and we will try to depict the various combinations in a way that clarifies their respective trade-offs.

Before we can start to compare and assess the potential of combining the methods, we will need to compute a new data set based on the problems lists of our previous evaluations. By calculating the union between the numbers of problems identified with two different methods we are able to represent the data in Figure 20, which also depicts the unified distributions of severities:

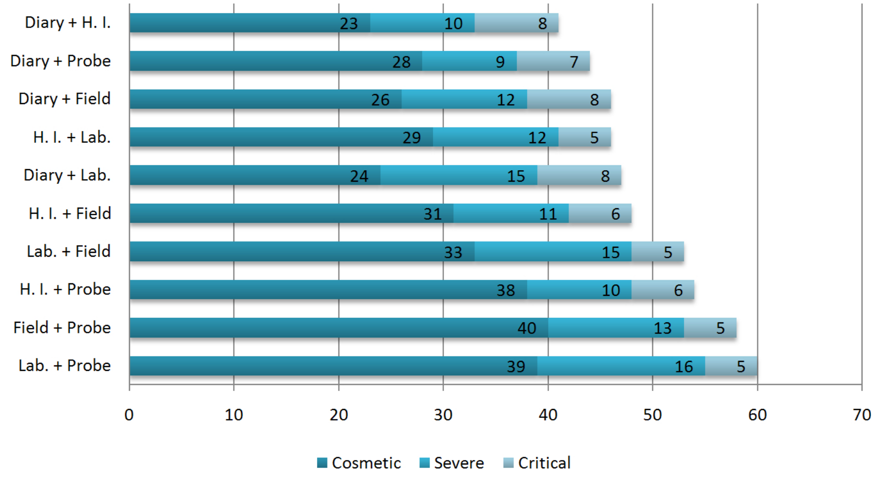

Figure 20 A representation of the total amount of problems identified (including severity) by each combination of methods.

Likewise, in order to represent the efficiency of the various combinations we need to consider both the amount of problems identified and the resource consumption in terms of the number of man hours spend on the evaluation. Thus by calculation the ratio between number of problems and hours spend on the combined evaluations we are able to construct a Figure 21 that expresses the efficiency of the combinations.

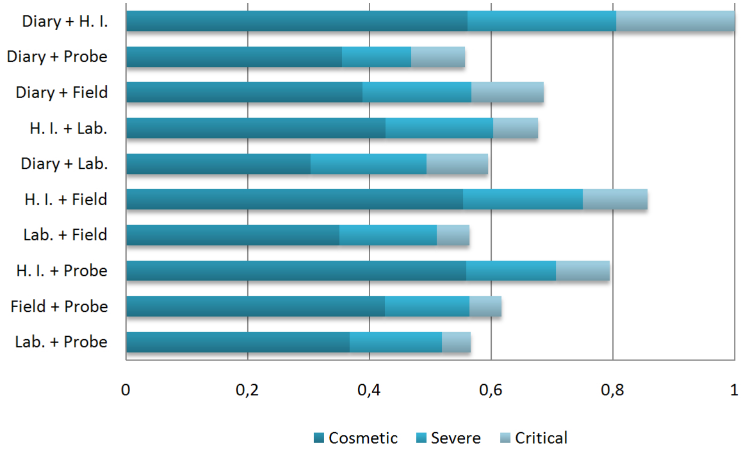

Figure 21 A representation of the number of problems identified in one man hour by each of the combintions.

By recalling the previous findings and reviewing Figure 20 and Figure 21 we have identified five combinations of methods that we would recommend under specific circumstances:

- Most efficient (Diary + H. I.)
- Shortest timeframe (H. I + Field)
- Largest number of problems (Lab. + Probe)
- Most critical problems (Diary + Lab)
- Best balance between efficiency and number of problems (H. I. + Probe)

These combinations and the circumstances under which they are recommendable will be documented in the following sections.

Most Efficient

The combination of methods that stands out when we regard the efficiency of the methods is the combination of **the diary evaluation and the heuristic inspection** (see Figure 21). This combination has very low resource requirement while it at the same time identifies problems quite well, and thus they have proven to be the most efficient combination. And since the efficiency is not only a representation of the amount of man hours put into the evaluation, but also a representation of the output in form of identified problems, we can conclude that this combination provides the most value for the money. A combination like this would be recommendable if keeping to the budget is top priority.

Shortest Timeframe

Even though the combination of diary evaluation and heuristic inspection would be an obvious choice if the evaluation budget is tight, it would be recommendable to consider the time span of the method as well as the resource consumption. If the situation requires for fast results acquired with minimal resources another approach would be recommendable, namely the combination of a **heuristic inspection and a field evaluation**. This combination is remarkably faster to conduct than the diary evaluation and the heuristic inspection because of the extended period of time (two weeks) needed to conduct the diary evaluation. The resulting findings from a combination of the heuristic inspection and the field evaluation is 48 usability problems identified using only 56 man hours, which makes this combination the second most efficient (see Figure 21).

Largest Number of Problems

In situations where budget and time frame is not a concern and the primary objective is to identify the largest number of problems possible, the obvious choice would be to go with the combination of a **laboratory evaluation and a video probe evaluation**. This combination is the top performer when only comparing the total amount of problems identified (see Figure 20) and as such it will be recommendable in situations where problem quantity is all that matters and the large resource consumption of this combination plays a minor role. Evaluators might argue that sometimes problems that are not identified and corrected could prove more expensive than the process of conducting such a thorough evaluation.

Most Critical Problems

A slight modification of the criteria's that the above recommendation is based on would result in a different combination. If the primary objective of the evaluators is to identify the most critical problems with no regard to resource consumption or time frame the preferable combination is harder to identify. If we take a look at Figure 20 three combinations stand out: Diary evaluation combined with heuristic inspection, diary evaluation combined with field evaluation and diary evaluation combined with laboratory evaluation. All of these combinations identify 8 critical usability problems, thus it is difficult to decide on what combination to recommend without also considering the severe problems identified by the combinations. By doing so it is evident that the **diary evaluation** should be combined with the **laboratory evaluation** in order to find the most critical and severe problems (23 in total).

Best Balance between Efficiency and Number of Problems

In order to recommend a combination of methods based on the combination's trade-off between efficiency and number of problems identified we have to rank the methods according to each metric. The combination of methods that stand out when considering these rankings is the combination of **the heuristic inspection and the video probe evaluation.** This combination of methods is both the one identifying the third most problems while at the same time being the third most efficient combination. Thus if we should recommend a method with the best balance between efficiency and number of problems identified this would be it.

These combinations should be seen as recommendations under different circumstances and therefore none are supposed to be regarded the universally best way to combine the methods. However it is remarkable that all recommended combinations include both a contextual and a non-contextual method. We will elaborate on this in the following.

Combining Contextual and Non-Contextual Methods

Based on the above discussion of the relevant combinations of methods we are left wondering, whether or not it is a coincident that all combinations consist of one contextual method and one non-contextual method? According to our understanding of the characteristics of the various methods we are convinced that the contextual methods have provided us with knowledge that we could not have obtained by applying only non-contextual methods and vice versa.

Thus our study indicates that the best way to approach usability evaluations of mobile systems would be to adopt a mixed approach. A combination of methods can be applied to explore different problem areas and provide the evaluators and developers with the knowledge needed for improving the usability of the system with respect to the context of use.

One advantage of combining the contextual and non-contextual methods that are not included in our recommendations is that contextual evaluations might provide the evaluators with a better understanding of the context which they can communicate to the rest of the development team. Thus by venturing into the field the evaluators can conduct usability evaluations with the objective of identifying usability problems, while at the same time being able to gain deeper understanding of how the users apply the system and behave in the context in general along with the contextual factors that influence them.

Usability engineers might argue that this information regarding the context should already be present when the product has reached a stage where usability testing is relevant. However by introducing new technology into the context, we might affect and alter the context itself. For instance by introducing a new technology, we might influence the internal relationships between coworkers, and thus alter the social context that we are designing for. Thus it will always be informative to observe the use of the system in context.

This serves to support the claim that when performing usability evaluations of mobile systems, evaluators can benefit from combining contextual and non-contextual methods. By combining methods the evaluators will be able to identify a broader range of usability problems and furthermore be able to gain valuable insight into the use of the given mobile system in the actual context of use. Depending on the available resources and objectives of the evaluation, decisions as to how methods should be combined might vary according to the recommendations proposed in the first part of this chapter.

Contextual Methods – Threats and Obstacles

In this last part of our discussion we will try to take one step back and elaborate on the potential threats and obstacles that can prevent researchers and practitioners from applying contextual methods in their work.

In order to describe the methodological distinction between contextual and non-contextual methods we turn to Richard O. Mason (Mason 1988), who describes the two primary attributes of knowledge producing activities such as the work presented in this thesis and the process of identifying usability problems in a mobile system in general. The attributes are: Tightness of control and richness of reality. These attributes are generally perceived to be in opposition to one another at the same level of knowledge. Therefore, researchers and practitioners must ultimately settle for a trade-off between these. The trade-off of these attributes can be used to illustrate the difference between contextual and non-contextual methods, where the latter often will be applied with focus on tightness of control and the first to ensure the richness of reality. This is in accordance to the observations on the objectives described as variables, that we presented in the literature review, where we used the term cotrol and realism. Thus when practitioners and researchers choose to apply non-contextual methods they also choose to utilize the tightness of control that they can achieve with such methods. The various methods that we have applied in this thesis can be placed on a continuum describing the trade-off between control and realism.

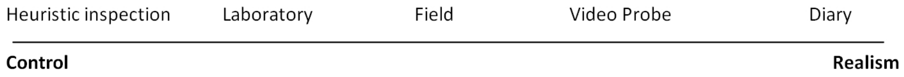

Figure 22 Continuum describing the relative focus on control and realism

The representation provided by the continuum in Figure 22 serves to illustrate the trade-off of each method when compared to the others. It should however not be perceived as an absolute definition of the trade-off of the individual methods. The locations of the methods in the figure express the relationship between the methods according to the *control-realism* continuum. These relationships should be seen relatively, for instance is the field more realistic than the laboratory and less controlled, but the figure is not expressing to what extent the methods differ.

The relevance of this discussion of methodological distinctions is that when researchers and practitioner opt for contextual methods they are forced to accept a reduced tightness of control in order to achieve the increased richness of reality or realism that the contextual methods are capable of providing. This might be an important obstacle for the adoption of contextual methods, since many researchers with backgrounds in the engineering disciplines will opt for the more controlled methodological approaches due to their positivistic heritage.

Another obstacle that we face when arguing that contextual methods can provide researchers with additional knowledge that they cannot gain by applying non-contextual methods, is the fact that researchers that do not adhere to our broad definition of usability, might question whether or not the additional knowledge that we have gained in our experiment is relevant when trying to identify usability problems. Thus a narrow definition of usability might leave the impression that the contextual methods are in fact less useful than we claim them to be. As the perceived potential of the contextual methods is reduced so would the incentive to apply the contextual methods.

It is our judgment that through our work as it is documented in this thesis, we have been able to illustrate the potential of contextual methods, not as a substitute for non-contextual, but as a different approach that can provide evaluators, and in the end, developers with different and valuable knowledge that will allow them to construct mobile systems that are more suitable for the context of use.

Conclusion

The purpose of this thesis has been to explore the practice and science of evaluating the usability of mobile systems. Thus we have worked towards an increased insight into how and why mobile systems are evaluated in the way they currently are and outline their strenghts and weaknesses. Subsequently we have sought to apply our knowledge in order to construct a method more suitable for the purpose. This has been based on a process structured by three research questions, the first of which sounded:

> Research question #1: *Which methods are used in the field of HCI to evaluate the usability of mobile systems?*

By performing a literature review we were able to shed some light on the research that has been conducted on evaluating the usability of mobile systems. Through our review we have identified four commonly used methodological approaches:

- Expert evaluations, such as **heuristic inspections** or contextual walk-throughs
- Usability evaluations conducted with real users in a **laboratory**
- Usability evaluations conducted with real users in the **field**
- **Longitudinal** evaluations conducted in the field

The expert evaluations are conducted by HCI experts or user interface specialists with guidance of design guidelines, heuristics, or scenarios. Typically these evaluations are conducted within a tight budget and within a very limited time frame.

The usability evaluations conducted in laboratories are typically performed in a way that resembles the typical evaluation of a standard desktop system, with the use of think aloud protocols and video recordings of the session used for later analysis.

The usability evaluations conducted in the field, is somewhat a derivative of the laboratory evaluations, and is usually conducted with the same techniques. Typically the only real difference between laboratory and field evaluations is the setting in which the evaluation takes place. The greatest difference between the two approaches is that in the field, the primary objective is often a pursuit of realism or ecological validity that is neglected in the laboratory evaluations.

The longitudinal approach resembles the field evaluations in some way, but typically the researchers tend to avoid using data collection techniques that influence the actual use situation. Instead they apply interviews, focus groups or diaries in order to collect data on the use of the system. The primary objective of the longitudinal evaluations are again a pursuit of ecological validity, that is sought realized through letting the users use the system for extended periods. Thus the longitudinal methods are often lengthy affairs that can run over long periods of time, but in turn provides extended insight in the users' reception and use of the system.

In general we see that the approaches that researchers apply in order to evaluate the usability of mobile systems are conducted with three important aspects in mind: **resources**, **control** and **realism**. The expert evaluation are typically chosen because of their low resource consumption. The laboratory evaluations are typically preferred by researchers who emphasize the benefits of the controlled environment, such as the increased reproducibility. The field and the longitudinal evaluations, are typically conducted with the intention of increasing the realism of the context of use.

With this new found knowledge we set out to conduct usability evaluations of mobile systems ourselves, in order to gain first-hand knowledge and experience with the methods that we found to be the ones typically applied by researchers. By doing so we were able to answer the second research question.

> Research question #2: *What characterizes the application and outcome of evaluating the usability of a mobile system with the current methods and how do the characteristics of the contextual and non-contextual methods differ?*

In order to provide conclusive characteristics of the four methodological approaches we summarize our experiences with each of the methods that we initially sought to explore.

The **heuristic inspection** is a method with focus on minimal resource consumption. Thus it came as no surprise that it was the cheapest method to apply. Besides being the only analytic approach to usability evaluation that we applied it was characterized by its ability to identify many cosmetic problems related to the interface of the system. As a stand alone method it is to be perceived as a suitable method for identifying non critical problems in user interfaces. It is not a recommendable choice in situations where the intention is to evaluate the entire system and identify the most possible problems, since it lacks the ability to identify many of the more severe problems, related to other themes than those covering the user interface. However, when combined with other methods it often stands out as a reasonable choice and therefore it is often a part of our recommendations.

The **laboratory evaluation** method is conducted with focus on control and consequently we found that it had poor resemblance of the realistic context of use. The results provided by our application of this method resembles the ones provided by the heuristic inspection, because it identified many problems related to the interface as well, but compared to the heuristic inspection the laboratory evaluation identified more problems overall with a larger group of severe and critical problems, thus making the method more complete on its own. We find that a laboratory evaluation might be useful for evaluating mobile systems, since it uncovers just as many problems as any other single method with approximately the same severity ratings.

The **field evaluation** was applied with focus on increasing the realism of the evaluation setting in order to attain ecological validity, without letting go of all the elements of control. This method was faster to conduct than the laboratory evaluation, but only because of the fact that the users' were able to conduct the tasks that we provided them faster than the users' in the laboratory. The method provided us with insight into the users' perception of the system as a whole and problems related to the overall utility of the functionality provided by the system. Likewise the field evaluation was able to identify several of the problems related to the interface that we also were able to identify by applying the laboratory evaluation and the heuristic inspection. Thus it seems that the field evaluation results in a better general understanding of the flaws in the system, but with lesser attention to interface related problems.

The **diary evaluation** was conducted as a longitudinal evaluation, with extensive focus on ecological validity; hence the use of the system under realistic circumstances was a top priority which forced us to give up some of the control that we had in the other evaluations. By avoiding obstructive data collection techniques that could taint the realism of the evaluation, we were forced to gather data in other ways. We chose apply diaries and to conduct interviews, which proved to be much faster to analyze than collected video material, which in turn meant that the diary method turned out to be the second least resource demanding of them all, despite the duration of two weeks. However the diary evaluation was also by far the method that identified the least amount of problems. The problems it identified were however characterized by being more severe and critical than those identified by any other method. The method identified a surprisingly large number of program bugs and some very serious problems related to the information theme that no other method had previously identified. Many of these identifications were closely related to the fact that the users used the system over an extended period compared to the other user in the other methods. On its own this method is unsuitable for general evaluations conducted with the intention of disclosing as many of the potential problems in a system as possible. However, because of its low demand of resources and the characteristics of the problems that it identifies, it is suitable for use in combination with other methods.

In conclusion, we found that there is no universally best method amongst the four methods we applied, when it comes to evaluating the usability of mobile systems. However we can conclude that all the applied methods can provide us with useful knowledge and that they all identified unique problems that no other method identified. Furthermore we can conclude that by applying contextual methods we were able to uncover a large group of problems that we were not able to uncover using the non-contextual methods. Thus we can conclude that there is something to gain from venturing into the field or in some other way conduct ecological valid evaluations.

In order to answer the third research question we decided to combine aspects of the contextual methods in order to construct a method that on its own was a more suitable method for evaluating the usability of a mobile system like EQO Mobile. This was an attempt to apply the knowledge that we had attained about the evaluation methods and served to answer the third research question:

> Research question #3: *How can we utilize the first-hand knowledge that we have gained and devise a method that overcomes the challenges imposed by the nature of mobile systems?*

Based on our findings and inspired by the literature review, we constructed a method that allowed the users to use the system over an extended period of time and in different contexts, while adding elements of control such as data collection with digital video cameras, in order to be able to capture and retain problems the users faced. Likewise we chose to use predefined tasks in order to structure the usage of the system in a way that would ensure that the users explored the entire system. These two techniques were adopted based on our experiences from the laboratory and field evaluation, and thus we suspected that the results of the new method would resemble that of the field and laboratory evaluation with regards to the amount of identified problems. Likewise we expected the problems that we would identify to be influenced by the context in which the users applied the system, and as such we expected to identify most of the problems that we uniquely identified with the diary evaluation. Based on the results derived from the evaluation, which we named the **video probe evaluation,** we can conclude that we were able to identify many but not all of the problems that were unique for the field and diary evaluation, and thus the method do not provide us with a single best way of identifying the problems that resembles the added value of contextual evaluations.

Based on our comparisons we decided to elaborate on the potential of applying methods in combination. Thus in our discussion we have proposed several recommendations based on different criteria's, which can assist or inspire practitioners and researchers alike in their decisions on how to evaluate a mobile system. Based on the various combinations that have been reviewed in the discussion and the insight that we have gained through this process, we are able to conclude that in many situations it will be advisable to combine contextual and non-contextual methods, and that evaluators by doing so can achieve a broader insight into the usability problems of a given mobile system.

Limitations and Further Work

In this final chapter we will account for the limitations of our work including thoughts on our own role in the process, the number of involved participants, the measurement of resources spend, the level of detail of the problem descriptions and how these elements affect the process of evaluating and comparing multiple evaluation methods. Furthermore we will suggest what could be done in order to overcome these limitations.

A limitation of our study that has influenced the validity of the findings is the so called evaluator effect. Because we decided to conduct the evaluations ourselves including the process of identifying and rating the severity of the identified problems, the study is limited by our bias. The knowledge we gained from applying the first evaluation method might have affected the outcome of the second and so forth such that we accumulated insight in the strengths and weaknesses of the system as we went along. Therefore chances are that we focused on the issues we knew where to find and neglected new aspects that we had not encounter before. In order to minimize this threat to the validity of the study, it would be necessary to let other researchers review our video recordings, diaries and notes.

A recommendation to further work based on the above mentioned limitation is, that in order to improve the overall validity and conclusive power of a comparative study like this, we would recommend that it was repeated with several groups of participant for each evaluation method. And each evaluation method should be applied independently with independent evaluators, in order to avoid that the researcher's bias contaminates the results. However it should be noted that in order to compare the results produced by different evaluators a commonly shared perception of the concept of usability would be necessary along with a strict set of guidelines for documenting the usability problems.

Another limitation of our work is the relatively small number of participants in the evaluations. This could be seen as limiting the validity of our results when comparing the evaluation methods. The number of participants in each group was chosen so that it was in compliance with the recommendations and guidelines of each method that we applied. However if we were to minimize this threat to the validity we would not increase the number of participants in each group, but instead conduct the evaluations several times with the same number of participants. We believe that by conducting further comparative studies of evaluation methods with groups of participants of the recommended size it is possible to depict the trade-offs of the methods as they would be applied in real life more precisely. Which potentially could provide practitioners with an extended knowledge of where and when to apply different methods.

Our decision to include the recommended number of participants in each evaluation was difficult to follow through since no guidelines exist for the appropriate number of participants in longitudinal evaluations, that we are aware of. Thus the decision to include four individuals and four couples in the two longitudinal evaluations was based on the impression that longitudinal evaluations would require few participants since they run over an extended period. Based on our study we believe that it will be advisable to conduct the video probe evaluation with the same amount or maybe even more participants than in the laboratory evaluation and field evaluation in order to explore its potential. However, more research would have to be conducted in order to determine a suitable amount of participants for such an evaluation.

Another limitation of our work is the resources spend on each evaluation method. Since we were familiar with some methods and others were new to us, we assume that the unknown methods took marginally longer to apply than the known ones did. A limitation of the comparability of the resource consumptions we see no way of overcoming without running trial evaluations of all methods beforehand. We therefore acknowledge that the consumption of the different methods is deeply rooted in previous experiences and knowledge of the method. In future comparative studies, guidelines should be established by researchers prior to the evaluations and the practical execution of evaluations should be outsourced to independent teams of evaluators.

A final limitation worth mentioning is the level of detail in which we described the usability problems. We decided to describe the identified problems as detailed as possible, as previously accounted for. Since the logic in the argument of looking at something in great detail is that it must consequently be easier to tell one thing apart from something else. Practically this meant that we described problems that were encountered in different parts of the system as separate problems even though they might be very similar. Thus we believe that a high level of detail in the problem descriptions does certainly not limit our ability to compare methods, but it might minimize the similarities between the compared methods. This duality of the descriptive level when looking for differences and similarities might be worth considering for other researchers when conducting comparative studies of usability evaluation methods.

Bibliography

References

Bevan, Nigel. “International Standards for HCI and Usability.” *International Journal of Human Computer Studies*, October 2001: 533 - 552.

Bevan, Nigel. “Usability is quality of use.” *Proceedings of the 6th International Conference on Human Computer Interaction.* Yokohama: Elsevier, 1995.

Gaver, Bill, Tony Dunne, and Elena Pacenti. “Design: Cultural Probes.” *interactions* (ACM Press, New York, NY), January/February 1999: 21-29.

Gorlenko, L, and R Merrick. “No Wires Attached: Usability in the connected mobile world.” *IBM Systems Journal*, 2003: 639-651.

Hartson, H. Rex, S. Terence Andre, and C. Robert Williges. “Criteria For Evaluating Usability Evaluation Methods.” *International Journal of Human-Computer Interaction*, 2003: 145-181.

Kristoffersen, Steinar, and Fredrik Ljungberg. “Representing modalities in Mobile Computing: A Model of IT use in Mobile Settings.” *Proceedings of Interactive applications of mobile computing* (Fraunhofer: Institute for Computer Graphics.), 1998.

Lindholm, Christian, Harri Kiljand, and Turkka Keinonen. *Mobile Usability: How Nokia changed the face of the mobile phone.* McGraw-Hill Professional, 2003.

Mason, Richard O. “Experimentation and knowledge – A pragmatic perspective, Knowledge: Creation, Diffusion, Utilization.” *Science Communication*, 1988: 3-24.

Nielsen, Jakob. “How to conduct a heuristic evaluation.” *useit.com: Jakob Nielsen on Usability and Web Design.* http://www.useit.com/papers/heuristic/heuristic_evaluation.html (accessed May 30, 2007).

Nielsen, Jakob. *Usability Engineering.* Academic Press Inc., 1994.

Rubin, Jeffrey. *Handbook of usability testing.* New York, NY: John Wiley & Sons, Inc., 1994.

Wynekoop, Judy L., og Sue A. Conger. »A review of computer aided software engineering research methods.« *Proceedings of the IFIP TC8 WG 8.2 Working Conference on The Information Systems Research Arena of The 90’s.* Copenhagen, 1990. 129-154.

Revieved Papers

1. Kjeldskov, Jesper and Paay, Jenny. “Just-for-us: A Context-Aware Mobile Information System Facilitating Sociality.” *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services.* Salzburg: ACM Press, New York, NY, 2005.

2. Cheverst, Keith, Nigel Davies, Keith Mitchell, Adrian Friday, and Christos Efstratiou. “Developing a Context-Aware Electronic Tourist Guide.” *SIGCHI Conference on Human Factors in Computing Systems.* Hague: ACM Press, New York, NY, 2000. 17-24.

3. Bach, Erik, et al. “Bubbles: Navigating Multimedia Content in Mobile Ad-hoc Networks.” *Proceedings on the 2nd International Conference on Mobile and Ubiquitous Multimedia.* Norrköping: ACM Press, New York, NY, 2003.

4. Huang, Elaine M., Michael Terry, Elizabeth Mynatt, Kent Lyons, and Alan Chen. “Distributing Event Information by Simulating Word-of-Mouth Exchanges.” *Mobile Human-Computer Interaction : 4th International Symposium, Mobile HCI 2002, Proceedings.* Pisa: Springer Berlin / Heidelberg, 2002.

5. Bertini, Enrico, Silvia Gabrielli, and Stephen Kimani. “Appropriating and assessing heuristics for mobile computing.” *Proceedings of the Working Conference on Advanced Visual interfaces.* Venezia: ACM Press, New York, NY, 2006.

6. Korhonen, Hannu, and Elina M.I. Koivisto. ”Playability heuristics for mobile games.” *Proceedings of the 8th Conference on Human-Computer interaction with Mobile Devices and Services.* Helsinki: ACM Press, New York, NY, 2006.

7. Gabrielli, Silvia, Valeria Mirabella, Stephen Kimani, and Tiziana Catarci. “Supporting cognitive walkthrough with video data: a mobile learning evaluation study.” *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services.* Salzburg: ACM Press, New York, NY, 2005.

8. Po, Shirlina, Steve Howard, Frank Vetere, and Mikael B. Skov. “Heuristic Evaluation and Mobile Usability: Bridging the Realism Gap.” *6th International Conference on Human Computer Interaction with Mobile Devices and Services.* Glasgow: Springer Berlin / Heidelberg, 2004.

9. Lamberts, Harald. “Case Study: A PDA Example of User Centered Design.” *Mobile Human-Computer Interaction : 4th International Symposium, Mobile HCI.* Pisa: Springer Berlin / Heidelberg, 2002.

10. Chincholle, Didier, Mikael Goldstein, Marcus Nyberg, and Mikael Eriksson. “Lost or Found? A Usability Evaluation of a Mobile Navigation and Location-Based Service.” *Mobile Human-Computer Interaction : 4th International Symposium, Mobile HCI.* Pisa: Springer Berlin / Heidelberg, 2002.

11. Koskela, Tiiu, and Inka Vilpola. “Usability of MobiVR Concept: Towards Large Virtual Touch Screen for Mobile Devices.” *Mobile Human-Computer Interaction – MobileHCI.* Glascow: Springer Berlin / Heidelberg, 2004.

12. Hyvärinen, Tuuli, Anne Kaikkonen, and Mika Hiltunen. ”Placing links in mobile banking application.” *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services.* Salzburg: ACM Press, New York, NY, 2005.

13. Hakala, Tero, Juha Lehikoinen, and Antti Aaltonen. “Spatial interactive visualization on small screen.” *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services.* Salzburg: ACM Press, New York, NY, 2005.

14. Büring, Thorsten, and Harald Reiterer. “ZuiScat: querying and visualizing information spaces on personal digital assistants.” *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services.* Salzburg: ACM Press, New York, NY, 2005.

15. Duh, Henry Been-Lirn, Gerald C. B. Tan, and Vivian Hsueh-hua Chen. “Usability evaluation for mobile device: a comparison of laboratory and field tests.” *Proceedings of the 8th Conference on Human-Computer interaction with Mobile Devices and Services.* Helsinki: ACM Press, New York, NY, 2006. 181 - 186.

16. Kaikkonen, Anne, Titti Kallio, Aki Kekäläinen, Anu Kankainen, and Mihael Cankar. ”Usability testing of mobile applications: A comparison between laboratory and field testing.” *Journal of Usability Studies*, 2005: 4-16.

17. Huang, Sheng-Cheng, I-Fan Chou, and Randolph G. Bias. “Empirical Evaluation of a Popular Cellular Phone’s Menu System: Theory Meets Practice.” *Journal of Usability Studies, Issue 2, Volume 1*, 2006: 91-108.

18. Ebling, Maria R., Bonnie E. John, and M. Satyanarayanan. “The importance of translucence in mobile computing systems.” *ACM Transactions on Computer-Human Interaction, Volume 9 , Issue 1*, 2002: 42-67.

19. Bederson, Benjamin B., Aaron Clamage, Mary P. Czerwinski, and George G. Robertson. “DateLens: A fisheye calendar interface for PDAs.” *ACM Transactions on Computer-Human Interaction, Volume 11 , Issue 1*, 2004: 90-119.

20. Chang, Angela, Sile O’Modhrain, Rob Jacob, Eric Gunther, and Hiroshi Ishii. “ComTouch: design of a vibrotactile communication device.” *Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques.* London: ACM Press, New York, NY, 2002. 312 - 320.

21. Kallio, Titti, and Aki Kekäläinen. ”Improving the Effectiveness of Mobile Application.” *MobileHCI 2004.* Glasgow: Springer-Verlag Berlin Heidelberg, 2004. 315–319.

22. Kjeldskov, Jesper, Michael B Skov, Benedikte S Als, and Rune T Høegh. "Is It Worth the Hassle? Exploring the Added." *MobileHCI 2004.* Springer / Heidelberg, 2004. 61-73.

23. Bohnenberger, Thorsten, Anthony Jameson, Antonio Krüger, and Andreas Butz. "Location-Aware Shopping Assistance: Evaluation of a Decision-Theoretic Approach." *Mobile Human-Computer Interaction : 4th International Symposium.* Pisa: Springer Berlin / Heidelberg, 2002. 155-169.

24. McGee, David R., Philip R. Cohen, R. Matthews Wesson, and Sheilah Horman. "Comparing paper and tangible, multimodal tools." *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves.* Minneapolis: ACM Press, New York, NY, 2002. 407-414.

25. Pirhonen, Antti, Stephen Brewster, and Christopher Holguin. "Gestural and audio metaphors as a means of control for mobile devices." *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves.* Minneapolis: ACM Press, New York, NY, 2002. 291-298.

26. Kjeldskov, Jesper, and Jan Stage. "New techniques for usability evaluation of mobile systems." *International Journal of Human-Computer Studies, Volume 60, Issues 5-6*, 2004: 599-620.

27. Bornträger, Christian, Keith Cheverst, Nigel Davies, Alan Dix, Adrian Friday, and Jochen Seitz. "Experiment with multi modal interfaces in a context-aware city guide." *Mobile HCI.* Springer-Verlag, 2003. 116-130.

28. Newcomb, Erica, Toni Pashley, and John Stasko. "Mobile computing in the retail arena." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Ft. Lauderdale, Florida, USA: ACM Press, New York, NY, 2003. 337-344.

29. Izadi, Shahram, et al. "Citywide: Supporting Interactive Digital Experiences Across Physical Space." *MobileHCI 2001.* Lille, France, 2001.

30. Holland, Simon, and David Morse. "Audio GPS: spatial audio in a minimal attention interface." *MobileHCI 2001.* Lille, France, 2001.

31. Cheverest, Keith, Keith Mitchell, and Nigel Davies. "Investigating Context-aware Information Push vs. Information Pull to Tourists." *Mobile HCI 2001.* Lille, France, 2001.

32. Fithian, Rachel, Giovanni Iachello, Jehan Moghazy, Zachary Pousman, and Stasko John. "The design and evaluation of a mobile location-aware handheld event planner." *Mobile HCI.* Springer-Verlag Berlin Heidelberg, 2003. 145-160.

33. Davies, Nigel, Keith Cheverst, Alan Dix, and Andre Hesse. "Understanding the role of image recognition in mobile tour guides." *Proceedings of the 7th international Conference on Human Computer interaction with Mobile Devices & Services.* Salzburg, Austria: ACM Press, New York, NY, 2005. 191-198.

34. Tähti, Marika, Ville-Mikko Rautio, and Leena Arhippainen. “Utilizing context-awareness in office type working life.” *Proceedings of the 3rd international Conference on Mobile and Ubiquitous Multimedia.* College Park, Maryland,: ACM Press, New York, NY, 2004. 79-84.

35. Aittola, Markus, Pekka Parhi, Maria Vieruaho, and Timo Ojala. “Comparison of mobile and fixed use of smartlibrary.” *MobileHCI.* Springer-Verlag Berlin Heidelberg, 2004. 383-387.

36. Teng, Chao-Ming (James), Chon-In Wu, Yi-Chao Chen, Hao-hua Chu, and Jane Yung-jen Hsu. “Design and evaluation of mProducer: a mobile authoring tool for personal experience computing.” *Proceedings of the 3rd international Conference on Mobile and Ubiquitous Multimedia.* College Park, Maryland: ACM Press, New York, NY, 2004. 141-148.

37. Tähti, Marika, Rautio, Ville-Mikko and Arhippainen, Leena.. “Utilizing context-awareness in office-type working life.” *Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia.* College Park, Maryland, 2004.

38. Esbjörnsson, Mattias, Oskar Juhlin, and Mattias Öst. “Motorcyclists Using Hocman - Field Trials on Mobile Interaction.” *Mobile HCI 2003.* Udine: Springer-Verlag Berlin Heidelberg, 2003. 32-44, 2003.

39. Demumieux, Rachel, and Patrick Losquin. “Gather Customer’s Real Usage on Mobile Phones.” *MobileHCI ‘05.* Salzburg, Austria.: ACM Press, NewYork, NY, 2005. 267-270.

40. Karlson, Amy K., George Robertson, Daniel C. Robbins, Mary Czerwinski, and Greg Smith. “FaThumb: a facet-based interface for mobile search.” *Proceedings of the SIGCHI conference on Human Factors in computing systems.* Montréal: ACM Press, New York, NY, 2006. 711-720.

41. Brown, Barry, Ian MacColl, Matthew Chalmers, Areti Galani, Cliff Randell, and Anthony Steed. “Lessons from the lighthouse: collaboration in a shared mixed reality system.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Ft. Lauderdale, Florida: ACM Press, New York, NY, 2003. 577-584.

42. Marsden, Gary, and Nicholas Tip. “Navigation control for mobile virtual environments.” *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services.* Salzburg, Austria: ACM Press, New York, NY, 2005.

43. Ojala, T, et al. “SmartRotuaari -- Context-aware mobile multimedia services.” *Proc. 2nd International Conference on Mobile and Ubiquitous Multimedia.* Norrköping, Sweden, 2003. 9--18.

44. Crabtree, Andy. “Design in the absence of practice: breaching experiments.” *Proceedings of the 2004 conference on Designing interactive systems: processes, practices, methods, and techniques.* Cambridge, MA: ACM Press, New York, NY.

45. Sawhney, Nitin, and Chris Schmandt. “Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments.” *ACM Transactions on Computer-Human*

Interaction, Vol. 7, No. 3 (ACM Transactions on Computer-Human Interaction, Vol. 7, No. 3), 2000: 353–383.

46. Luchini, Kathleen, Chris Quintana, and Elliot Soloway. "Design Guidelines for Learner-Centered Handheld Tools." *CHI 2004.* Vienna: ACM Press New York, NY, 2004. 135-142.

47. Abowd, Gregory D., and Elizabeth D. Mynatt. "Charting Past, Present, and Future Research in Ubiquitous Computing." *ACM Transactions on Computer-Human Interaction, Vol. 7, No. 1*, 2000: 29–58.

48. Koskela, Tiiu, Kaisa Väänänen-Vainio-Mattila, and Lauri Lehti. "Home Is Where Your Phone Is: Usability Evaluation of Mobile Phone UI for a Smart Home." *Mobile Human-Computer Interaction – MobileHCI 2004.* Springer Berlin / Heidelberg, 2004.

49. Stifelman, Lisa, Baron Arons, and Chris Schmandt. "The audio notebook: paper and pen interaction with structured speech." *Proceedings of the SIGCHI conference on Human factors in computing systems.* Seattle, Washington: ACM Press, New York, NY, 2001.

50. Sammon, Michael J, Lynne Shapiro Brotman, and Ed Peebl. "MACCS: Enabling Communications for Mobile Workers within Healthcare Environments." *Mobile HCI '06.* Helsink: ACM Press, New York, NY, 2006. 41-44.

51. Sefelin, Reinhard, Verena Seibert-Giller, and Manfred Tscheligi. "Xaudio: Results from a Field Trial Study on a Technology Enhancing Radio Listeners' User Experience." *MobileHCI 2004.* Springer-Verlag Berlin Heidelberg, 2004. 351–355.

52. Isaacs, Ellen, Alan Walendowski, and Dipti Ranganthan. "Hubbub: A sound-enhanced mobile instant messenger that supports awareness and opportunistic interactions." *CHI 2002.* Minneapolis: ACM Press, New York, NY, 2002. 179-186.

53. Milewski, Allen E, and Thomas M Smith. "Providing Presence Cues to Telephone Users." *CSCW '00.* Philadelphia: ACM Press, New York, NY, 2000. 89-96.

54. Jones, Matt, Jain Preeti, George Buchanan, and Gary Marsden. "Using a Mobile Device to Vary the Pace of Search." *Mobile HCI 2003.* Springer-Verlag Berlin Heidelberg, 2003. 390-394.

55. Fono, David, and Scott Counts. "Sandboxes: Supporting Social Play through Collaborative Multimedia Composition on Mobile Phones." Banff, Alberta, Canada: CSCW '06, 2006.

56. Bardram, Jakob, Thomas A. K. Kjær, and Christina Nielsen. "Supporting Local Mobility in Healthcare by Application Roaming Among Heterogeneous Devices." *Mobile HCI 2003.* Springer-Verlag Berlin Heidelberg, 2003. 161-176.

57. Newman, Mark W., et al. "Designing for serendipity: supporting end-user configuration of ubiquitous computing environments." *Proceedings of the conference on Designing interactive systems: processes, practices, methods, and techniques.* London: ACM Press, New York, NY, 2002.

58. Wright, Tim, et al. "Usability methods and mobile devices: an evaluation of MoFax." *Proceedings of the 4th international conference on Mobile and ubiquitous multimedia.* Christchurch: ACM Press, New York, NY, 2005.

59. Kuikkaniemi, Kai, Marko Turpeinen, Antti Salovaara, Timo Saari, and Janne Vuorenmaa. "Toolkit for user-created augmented reality games." *Proceedings of the 5th international conference on Mobile and ubiquitous multimedia.* Stanford: ACM Press, New York, NY, 2006.

60. Karlson, Amy K., Benjamin B. Bederson, and John SanGiovani. "AppLens and launchTile: two designs for one-handed thumb use on small devices." *Proceedings of the SIGCHI conference on Human factors in computing systems .* Portland: ACM Press, New York, NY, 2005.

61. Amir, Anat S. "O2 Active: Enhancing User Experience on Mobile Phones." *Mobile Human-Computer Interaction – MobileHCI 2004.* Glagow: Springer Berlin / Heidelberg, 2004.

62. Viljamaa, Timo-Pekka, Akseli Anttila, and Rob Van Der Haar. "Creation and application of mobile media design drivers." *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services.* Salzburg: ACM Press, New York, NY, 2005.

63. Black, Jason T., Lois W. Hawkes, Jean-Pierre Ketly, Isaac Johnson, and Marvin Lee. "A Paper Prototype Study of the Interface for a Children's Collaborative Handheld Learning Application." Glasgow: Springer Berlin / Heidelberg, 2004.

64. Hinckley, Ken, Jeff Pierce, Eric Horvitz, and Mike Sinclair. "Foreground and background interaction with sensor-enhanced mobile devices." *ACM Transactions on Computer-Human Interaction* (ACM Press New York, NY) 12, no. 1 (2005).

65. Hinckley, Ken, et al."Sensing techniques for mobile interaction." *Proceedings of the 13th annual ACM symposium on User interface software and technology.* San Diego: ACM Press, New York, NY, 2000.

66. Lam, Heidi, and Patrick Baudisch. "Summary thumbnails: readable overviews for small screen web browsers." *Proceedings of the SIGCHI conference on Human factors in computing systems.* Portland: ACM Press, New York, NY, 2005.

67. Bardram, Jakob E., and Thomas R. Hansen. "The AWARE architecture: supporting context-mediated social awareness in mobile cooperation." *Proceedings of the 2004 ACM conference on Computer supported cooperative work.* Chicago: ACM Press, New York, NY, 2004.

68. Nakanishi, Yasuto, Shouichi Kumazawa, and Takayuki Tsuj. "iCAMS2: Developing a Mobile Communication Tool Using Location Information and Schedule Information with J2ME." *Human-Computer Interaction with Mobile Devices and Services.* Udine: Springer Berlin / Heidelberg, 2003.

69. Krauss, Matthias, and Dennis Krannich. "ripcord: rapid interface prototyping for cordless devices." *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services.* Helsinki: ACM Press, New York, NY, 2006. 187-190.

Appendices

These appendices is a collection of the various guidelines, tasks, introductions etc. that we have used during our work. Most of it is in Danish, and purposely so, since all of the participating experts and users are from Denmark.

Appendix A - Heuristics

Visibility of system status

> The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

Match between system and the real world

> The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

User control and freedom

> Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Consistency and standards

> Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

Recognition rather than recall

Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Flexibility and efficiency of use

Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

Aesthetic and minimalist design

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

Help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Appendix B - Laboratory

Introskrivelse til usabilitytest i lab

Velkommen og tak fordi du vil deltage i dette forsøg. Forsøget går ud på at vi i den næste halve times tid skal teste et system der hedder EQO, der overfører noget af MSN Messengers funktionalitet til mobiltelefonen. Testen skal hjælpe os til at afsløre brugervenlighedsproblemer i produktet, som på nuværende tidspunkt er under udvikling. Det er således ikke et færdigt system du skal teste og der kan forekomme tekniske vanskeligheder under vejs. Testen foregår på den måde at du får lov at sidde med telefonen og bruge systemet til at løse en række opgaver som vi på forhånd har udarbejdet. Disse opgaver er udarbejdet således at de fører dig rundt i de væsentligste dele af systemet. Systemet der testes indeholder yderligere funktionalitet, men i dag fokusere vi på den del der har med MSN Messenger at gøre.

Opgaverne vil blive udleveret en af gange både mundtligt og på skrift. Du er på et hvert tidspunkt velkommen til at stille afklarende spørgsmål vedrørende opgaverne, men det er meningen at du selv skal forsøge at løse dem. Opgaverne er af varierende sværhedsgrad og det er ikke et mål i sig selv at udføre dem så hurtigt som muligt. Husk det ikke er dig vi tester, men systemet ☺

Under testen vil vi gerne anmode dig om at forsøge at tænke højt. Det vil sige at vi gerne vil have dig til at fortælle os hvad du oplever, hvilke forventninger du har til de handlinger du foretager, samt hvorledes du opfatter systemets reaktioner herpå. Det kan godt være svært at vænne sig til at tænke højt, så inden vi starter vil vi lige bruge fem minutter på at træne. Samtidig får du en chance for at gøre dig selv fortrolig med den telefon der benyttes under testen. Jeg vil som testleder forsøge at holde mig passiv mens du løser opgaverne, men vil løbende stille spørgsmål hvis jeg gerne vil have afklaret noget.

Efter alle opgaver er gennemført vil vi tage en kort snak om systemet generelt og de problemer du oplevede i forbindelse med testen.

Opgaver til laboratorietest

Du er kommet i gruppe med to personer du ikke har på messenger og vil aftale et møde med dem vha. dette medie

1. Start EQO og log på messenger med din egen konto
2. Tilføj kontakterne Gruppekammerat1 og Gruppekammerat2 med email-adresserne: ____@hotmail.com og ____@hotmail.com
3. Se om de er online og aftal et mødetidspunkt med dem.
4. Du vil ringe til Andreas (som du har i telefonbogen) og høre om han har tid til at drikke en øl senere. Sæt din status til optaget og foretag opkaldet.
5. Vend tilbage til messenger og se om gruppekammeraterne har skrevet noget – hvis ikke, så luk samtalerne.
6. Der er nu gået et halvt år og du vil ikke snakke med de tidligere gruppekammerater længere – slet dem derfor fra din kontaktliste og log af messenger.

Tak ☺

Appendix C - Field

Introskrivelse til felttest

Velkommen og tak fordi du vil deltage i dette forsøg. Forsøget går ud på at vi i den næste halve times tid skal teste et system der hedder EQO, der overfører noget af MSN Messengers funktionalitet til mobiltelefonen. Testen skal hjælpe os til at afsløre brugervenlighedsproblemer i produktet, som på nuværende tidspunkt er under udvikling. Det er således ikke et færdigt system du skal teste og der kan forekomme tekniske vanskeligheder under vejs. Testen foregår på den måde at du får lov at sidde med telefonen og bruge systemet til at løse en række opgaver som vi på forhånd har udarbejdet. Disse opgaver er udarbejdet således at de fører dig rundt i de væsentligste dele af systemet. Systemet der testes indeholder yderligere funktionalitet, men i dag fokusere vi på den del der har med MSN Messenger at gøre.

Du er på et hvert tidspunkt velkommen til at stille afklarende spørgsmål vedrørende opgaverne, men det er meningen at du selv skal forsøge at løse dem. Opgaverne er af varierende sværhedsgrad og det er ikke et mål i sig selv at udføre dem så hurtigt som muligt. Husk det ikke er dig vi tester, men systemet ☺

Under testen vil vi gerne anmode dig om at forsøge at tænke højt. Det vil sige at vi gerne vil have dig til at fortælle os hvad du oplever, hvilke forventninger du har til de handlinger du foretager, samt hvorledes du opfatter systemets reaktioner herpå. Det kan godt være svært at vænne sig til at tænke højt, så inden vi starter vil vi lige bruge fem minutter på at træne. Samtidig får du en chance for at gøre dig selv fortrolig med den telefon der benyttes under testen. Jeg vil som testleder forsøge at holde mig passiv mens du løser opgaverne, men vil løbende stille spørgsmål hvis jeg gerne vil have afklaret noget.

Opgaver

Du er kommet i gruppe med to personer du ikke har på messenger og vil aftale et møde med dem vha. dette medie

1. Start EQO og log på messenger med din egen konto

2. Tilføj kontakterne Gruppekammerat1 og Gruppekammerat2 med email-adresserne: gruppemedlem1@hotmail.com og gruppemedlem2@hotmail.com

3. Se om de er online og aftal et mødetidspunkt med dem.

4. Du vil ringe til Andreas (som du har i telefonbogen) og høre om han har tid til at drikke en øl senere. Sæt din status til optaget og foretag opkaldet.

5. Vend tilbage til messenger og se om gruppekammeraterne har skrevet noget – hvis ikke, så luk samtalerne.

6. Der er nu gået et halvt år og du vil ikke snakke med de tidligere gruppekammerater længere – slet dem derfor fra din kontaktliste og log af messenger.

Tak ☺

Appendix D - Diary

Test af mobil messenger – formål og forventninger

Formålet med at lade jer bruge jeres mobil med messenger på over de to uger, som forsøget kommer til at vare, er at teste brugen af messenger i forskellige situationer og finde ud af hvor det eventuelt fejler og kunne gøres bedre.

Forsøget er udformet som et feltstudie, hvor vi lader jer bruge systemet lige som i har lyst til, men vi opfordrer jer til at lade brugen være så realistisk som muligt og ikke lade jer føle tvunget til at bruge det, i situationer hvor i ikke ville bruge ”normalt”. Vi ønsker at opnå en forståelse af brugen af messenger på mobilen. Vi ønsker ikke at teste, hvor gode I er til at bruge jeres mobiler eller hvor mange kontakter I har på messenger.

I får udleveret en dagbog, hvori i bedes føre notater om hvor og hvornår i bruger mobilen til messenger, hvad i lavede samtidig med, hvorfor i brugte messenger i den situation, hvad i synes der var godt og skidt ved brugen. For at samle op på jeres erfaringer og for at vi ikke skal rende og puste jer i nakken, når i bruger messenger på mobilen, har vi planlagt løbende messenger-interviews, hvor vi spørger ind til hvordan det går og hvad I har gjort jer af erfaringer for de ikke går i glemmebogen eller vi overser noget vigtigt.

Forsøget stopper torsdag d.. 23/11 hvor vi samler dagbøgerne ind. Og et par dage efter når alle har tid afholder vi et debriefing-møde hvor vi lige samler op på jeres erfaringer osv.

I får naturligvis jeres udgifter dækket og vi siger tak for jeres villighed til at deltage i det her forsøg og håber på i får noget ud af det også.

Andreas og Lars

Dagbogs Guidelines

Meningen med denne dagbog er at give dig og os et redskab til at fastholde dine erfaringer med den daglige brug af mobil messenger.

Følgende punkter er relevante for os til at opnå en forståelse af din brug af messenger såvel som din opfattelse af brugbarheden. Og du kan derfor gå ud fra følgende, når du fører dagbog.

NB. Hver side bedes påført dato!

- **Hvor og hvornår brugte du det?** F.eks. ”I bussen på vej til skole om morgenen”
- **Hvorfor brugte du det i den situation?** F.eks. ”kommunikation med venner” eller bare for ”at være tilgængelig”
- **Hvad oplevede du som værende positivt/negativt ved brugen?** Både mht. den situation systemet blev brugt i og systemet i sig selv.
- Hvis du oplevede problemer ved brugen af systemet, beskriv problemet og dets konsekvenser.
- Hvis du ikke havde haft messenger på mobilen, hvordan havde du så kommunikeret i situationen – hvis du havde kommunikeret?

Spørgeguide til statussessions:

Spørgsmål der skal afdække brugen af systemet.

1. I hvilke situationer har du brugt systemet? (Få afdækket alle brugssituationer)
2. Hvor lang tid strakte de enkelte brugssessions sig over?
3. Hvad var formålet med de enkelte sessions? (f.eks. Kommunikation, tilstedeværelse)
4. Hvad er din vurdering af systemet som medie/værktøj i disse brugssituationer?
5. Hvad er din overordnede vurdering af systemet og dets anvendelighed?

Spørgsmål der skal afdække problemer i forbindelse med brugen af systemet.

1. Er du i forbindelse med brugen af systemet stødt på problemer eller uhensigtsmæssigheder? (Få afdækket samtlige problemer)
2. Kan du beskrive problemet/uhensigtsmæssigheden? (Årsagen afdækkes)
3. Hvilke konsekvenser fik problemet/uhensigtsmæssigheden for den opgave du var ved at løse? (Konsekvenserne på kort sigt afdækkes)
4. Hvilke konsekvenser har problemet/uhensigtsmæssigheden haft for din videre brug af systemet? (Problemets konsekvenser for fremtidig brug afdækkes)

Disse spørgsmål skal afklares under hver statussession, men der kan spørges på forskellige måder. Det er således ikke en fast spørgeguide, men snarere en struktur for de enkelte interviewsessions.

Statussessions foregår via. MSN Messenger til mobiltelefonen hver anden dag under hele feltstudiets forløb. Det er muligt at vi efter de første to statussessions vil lade brugerne benytte alm. Messenger for at lette deres arbejdsbyrde. Således kan vi bruge de første statussessions, dels som stimulans til yderligere brug af systemet og dels som en datakilde.

Appendix E - Video Probe

Introduktion og vejledning til video proberne

Velkommen til vores lille forsøg og tak fordi I to gad at deltage.

Over den næste uges tid kommer i til at lære programmet EQO Mobile bedre at kende. Det kommer til at forgå på den måde at i skal gå og bruge jeres telefon med EQO lidt i hverdagen og sammen løse et par opgaver som I finder i kassen hér sammen med en opsætningsvejledning, samt videokamera og bånd.

Første dag skal I løse en introducerende opgave sammen og filme det. Denne opgave finder I i konvolut nr. 1, som også indeholder instruktioner til de næste par dages brug af programmet. Derefter vil i modtage en SMS-påmindelse cirka hver anden dag om at åbne den næste konvolut når I er sammen. I disse konvolutter vil der ligeledes være opgaver og nogle instruktioner til de efterfølgende dage.

Når I løser opgaverne er det meningen at én af jer løser opgaven mens den anden filmer. I kan selv bestemme rollefordelingen – men prøv at skiftes lidt, så I efterfølgende bedre kan diskutere, hvad der er godt og skidt ved programmet. Efter hver opgaveløsning skal I interviewe hinanden og fortælle lidt om jeres erfaringer – dels med den løste opgave og lidt om de problemer I er stødt på mens i har brugt programmet de foregående dage.

Rent praktisk forestiller vi os at en af jer fungerer som kameramand og optager opgaveløsningen fra en vinkel der tillader os at se hvad der foregår på skærmen og høre hvad der bliver sagt. Den anden skal forsøge at tænke højt når han/hun løser opgaven – altså løbende fortælle om hvad de ser ske og hvad de forventer der vil ske. I må meget gerne snakke sammen mens I løser opgaver så det at tænke højt falder mere naturligt – kameramanden er således også velkommen til at stille spørgsmål undervejs.

Opsætningsvejledning

Før du installerer messenger-programmet EQO Mobile på din mobil skal den sættes op til GPRS så den kan gå på nettet – hvis den ikke allerede er sat op til det.

Normalt kan man hente en profil på teleselskabets hjemmeside:

- TDC (inklusiv Unotel og Telmore): http://erhverv.tdc.dk/mobilitet/opsaetning/gprs/
- Sonofon (inklusiv CBB og Debitel): http://dms.sonofon.dk/sonofon/wizard.form
- Telia: http://telia.dk/privat/selvbetjening/

Derefter skal du oprette en profil på EQO's hjemmeside, hvilket er gratis. Når du opretter din konto bliver du spurgt om mobilselskab og telefonmodel så de kan sende en version af programmet til dig som virker på netop din mobil. Når du har gjort det vil du modtage en besked med et link som du skal følge for at downloade og installere EQO.

Lav en konto og følg instruktionerne på: https://www.eqo.com/subprov/signup.html

Ved opsætningsproblemer skal vi nok hjælpe – bare giv lyd.

MVH

Andreas og Lars

60612015 og 24494131

abager@hotmail.com og larsmichael@hotmail.com

Opgaver

Konvolut nr. 1: I denne konvolut finder I en lille opgave som skal sætte jer i gang med at bruge EQO Mobile. Denne opgave skal én af jer løse, mens den anden filmer opgavens udførsel (film gerne over skulderen mens Ii prøver at fokusere på telefonens skærm). Mens opgaven udføres vil vi gerne anmode den der løser opgaven om at "tænke højt".

Opgave: *Opret en MSN Messenger profil i EQO og log på Messenger.*

Når opgaven er udført må I meget gerne kommentere på den, samt på jeres oplevelser med EQO Mobile. I er velkomne til at komme med ris og ros til systemet og evalueringsmetoden.

I løbet af de næste par dage vil vi gerne anmode jer om at bruge/udforske EQO Mobile i jeres hverdag. I bestemmer selv hvornår og hvor i vælger at anvende det og med hvilket formål.

God fornøjelse!

Konvolut nr. 2: I har nu haft to dage til at bruge EQO Mobile på egen hånd. I dag skal i filme hinanden mens I fortæller om jeres erfaringer / oplevelser med systemet. Vi vil gerne have at I fokusere på, hvornår og i hvilken forbindelse I har brugt EQO Mobile, samt hvilke problemer/udfordringer I har oplevet i forbindelse hermed. I bestemmer selv hvor lang tid interviewet skal vare og hvordan I vælger at filme det, men vi vil sætte pris på, at I kommer med eksempler på de problemer I oplever og dokumenterer dem ved at filme telefonen mens I forklarer om problemet.

I løbet af de næste par dage vil vi gerne opfordre jer til at prøve at bruge EQO uden for hjemmet, evt. mens I er undervejs med offentlig transport/går ned til købmanden eller hvor I nu ellers færdes.

God fornøjelse!

Konvolut nr. 3: Denne konvolut indeholder en lille opgave som vi gerne vil have én af jer til at løse, mens jeres partner optager det på video. Desuden vil vi gerne have jer til at fortælle lidt om jeres brug af EQO Mobile i løbet af de sidste par dage. Igen er det op til jer selv hvordan I vil arrangere interviewet, men vi vil gerne anmode om at I forsøger at give eksempler (gengive dem på telefonen og filme den), hvis I har oplevet nogle problemer eller hvis I undres over noget i programmet. I bestemmer selvom I vil tage interviewet eller opgaven først.

> Opgave: *Tilføj via EQO Mobile kontakten eqokammerat@hotmail.com til din liste over kontakter.*

I løbet af i morgen skal I prøve at fange jeres nye kontaktperson på MSN Messenger (Via EQO Mobile) og fortælle ham lidt om hvad du har brugt EQO til og hvad I synes om det.

God fornøjelse!

PS. EQOKammerat kan oftest træffes i tidsrummet 17-22.

Konvolut nr. 4: I dag vil vi gerne bede jer om at løse en lille opgave. Som det var tilfældet med de forrige opgaver vil vi gerne have at den ene af jer filmer, mens partneren løser opgaven (film gerne over skulderen mens I prøver at fokusere på telefonens skærm). Mens opgaven udføres vil vi gerne anmode den der løser opgaven om at ”tænke højt”.

> Opgave: *Slet jeres nye kontaktperson (EQOKammerat) fra jeres liste over kontakter og log herefter af MSN Messenger.*

Når opgaven er løst og kameraet endnu ikke er pakket ned, må I gerne benytte lejligheden til at dokumentere eventuelle nye problemer I er stødt på, såvel som de erfaringer I har gjort jer i forbindelse med de sidste par dages brug.

Der resterer nu to dage af testen, i løbet af dette tidsrum må I bruge systemet som I lyster, men vi vil gerne have jer til at være opmærksomme på eventuelle nye problemstillinger som I kunne tænkes at støde på.

God fornøjelse!

Konvolut nr. 5: Evalueringsforløbet er nu næsten afsluttet, men inden vi lukker og slukker vil vi gerne bede jer om at interviewe jer selv og fortælle lidt om jeres overordnede indtryk af EQO Mobile. I må meget gerne komme ind på hvordan det har været at bruge systemet, og om i har kunnet vende jer til at bruge systemet.

Når I har afsluttet dette sidste interview har I udspillet jeres rolle i dette forsøg, og vi vil gerne benytte lejligheden til at sige tak for hjælpen. Som tak for hjælpen / kompensation /plaster på såret kvitterer vi for jeres hjælp med et par flasker god rødvin.

Mange tak for hjælpen!

Mvh. Lars og Andreas

Appendix F - Top10

EQO Mobile - Top10 Usability problems

The purpose of this exercise is to identify and rank the 10 most significant usability problems with EQO Mobile. In order to do this, we need you to identify and rank the 10 problems with the system that you feel would be most important to fix. The exercise consists of two overall tasks.

Task 1: You will be provided with a list of all the usability problems that we have identified with the EQO system (83 in total). **Firstly,** we want you to read and understand all problems on the list (you are welcome to ask us if something needs to be clarified). **Secondly,** we want you to pick out the 10 most significant problems and rank them from 1 to 10 (1 being most important). For instance if you think that fixing problem number 33 would have the greatest impact on the usability of the system, you should rank this problem as #1. While you are doing this, another usability expert will be creating his top 10 list from the same 83 problems.

Task 2: When both of you have produced a top 10 of usability problems, we want you to merge them into one shared list of ranked 10 problems, which you both agree on represent the 10 most important usability problems with the system. **Firstly,** we want you to agree on which 10 problems should be on the final list. **Secondly,** we want you to rank these problems from 1 to 10 (1 being most important).

After the exercise, we need you to hand the following information back to us:

1. Your two individual ranked top 10 lists
2. Your merged ranked top 10 list

Thanks in advance ☺

Andreas & Lars

Appendix G - List of Identified Problems

	Problem description (Standards)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
1	The "back" button doesn't cancel or go back when the user is prompted to confirm the deletion of his profile.	Claus				
2	The user expects to be able to exit the program by pressing the "back"button in the EQO main menu, but nothing happens when he tries.	Claus, Esben		Anne, Caroline, Kenneth, Nils, Mads		Katrine & Allan, Nanna & Niels
3	Against expectation the number-buttons can't be used to navigate the contact list.		Camilla, Julie		Søren	
4	The user expects to be able to use the cancel button to delete elements like contacts and IM profiles, but the button has no functionality besides when typing text.		Camilla			
5	The user assumes that he is able to fill in the form fields without having to enter them by pressing "Rediger"/"Compose". For instance by just starting to type when adding a new MSN account, a new contact person or when chatting.	Claus	Julie, Lars, Mads, Troels	Anne, Kenneth, Nils		Signe & Peter, Nanna & Niels
6	The user expects to be able to change the feedback sound when receiving a new message. Like the ability to change the sound of an incoming SMS on the cell phone.					Signe & Peter
7	The user expects to be able to cancel to login process, but can't.					Nanna & Niels
	Problem description (Mental Model)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
8	The user expects to be able to group the contacts - as in the PC-version.	Thomas	Camilla			Signe & Peter, Nanna & Niels
9	Despite having chosen to hide contacts that are offline - so they aren't shown in the contact list - they are shown the next time the user logs on to EQO mobile.				Christian	Signe & Peter, Nanna & Niels
10	The user expects to be able to chat with more than one contact at a time - in the same session.		Julie, Troels, Helle			
11	The user finds that the lacking functionality of being able to click on links, share files etc in chats is			Nils		Nanna & Niels

	annoying.					
12	The user doesn't understand why their account must be saved before it can be used.	Claus	Camilla, Julie, Lars, Mads, Troels, Helle			
13	The user expects to be able to change the display name and their associated comment					Signe & Peter, Nanna & Niels
14	The user expects to be able to send offline messages.					Signe & Peter, Nanna & Niels
15	The user expects to be able to add new smilies/emoticons as in the PC version.					Kamilla & Dennis
16	The user expects to be able to create a new MSN account.					Katrine & Allan, Nanna & Niels
17	When receiving a message while composing one, the user expects to be able to read the new message and return to the message he was composing afterwards. But the message he was composing is not saved so he has to start typing over.					Katrine & Allan
	Problem description (Semantics)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
18	The user doesn't understand the menu items in EQO home ("Friends" and "Messages").	Claus, Thomas	Camilla, Julie, Lars, Troels, Mads	Caroline		
19	It is unclear to the user what the menu item “Home” is and surprised to be returned to EQO Home when pressing it.	Esben		Caroline		
20	The users can’t make sense of the “ok” functionality of the right soft-button when signing on to an MSN account.	Claus	Camilla, Mads, Troels	Nils, Anders	Søren	
21	The user is puzzled over the "Compose"-functionality since it usually is associated to mail functionality not chatting.	Thomas, Jakob				
22	In the IM Services "mere- menu the user notices that "add new" should have been phrased "add profile" making it easier to understand.	Thomas				
23	When exiting EQO the user is confused about the cancel/exit option instead of a yes/no when being asked	Thomas, Claus		Mads, Nils		Nanna & Niels

	if he wants to exit.					
24	The status icons in the contact list are hard to decode and the user doesn't know who is away and who is busy.	Jakob	Helle	Anne, Caroline		
25	When chatting, the user doesn't understand "clear" in the "mere" menu.			Kenneth		
26	The Hide functionality in the EQO Home "mere" menu is unclear to the user, who doesn't understand that it minimizes EQO.			Kenneth		
27	The language used on the soft-buttons is inconsistent. They change between english and danish: Compose/Rediger, Select/Vælg.	Claus	Lars			
28	When composing a message the 'quick notes' aren't translated into the default language of the phone.					Kamilla & Dennis
	Problem description (Physical)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
29	The small sized screen limits the users overview of the system.					Katrine & Allan, Kamilla & Dennis, Nanna & Niels
30	The user is frustrated over the slow input rate of text.			Anders, Mads, Nils	Christian, Klaus, Søren, Peter	Katrine & Allan, Kamilla & Dennis, Nanna & Niels
31	After a while the user feels a bit motion sick from using it while 'on the move'.			Anne, Nils		
	Problem description (Feedback)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
32	The user is logged off MSN without any notice.	Claus	Helle	Anne		Kamilla & Dennis, Nanna & Niels
33	When exiting EQO the user is confused whether or not he is still online.			Kenneth		
34	When adding a new contact the "invitation" notification disappears before the user gets to read it.	Esben	Helle			
35	When the contact list updates the user can't interact with the system.	Claus	Julie, Mads, Troels	Kenneth		Kamilla & Dennis,
36	When signing on to his MSN account the user is returned to EQO Home			Nils		

	without any feedback.					
37	The user has to confirm closing an empty chat.					Nanna & Niels
38	The user receives no verifications when sending messages, setting status, deleting contacts or signing in.		Julie, Mads, Helle			
39	When "hiding" EQO the connection is lost and EQO must reconnect when maximized. The users expect to be able to maintain their online status while EQO is minimized, since they are not informed of anything else.		Camilla, Julie, Mads, Helle	Kenneth		
40	When starting up EQO the user doesn't understand why he can't interact with the "IM Services" menu while it is connecting.	Esben	Lars, Julie, Troels	Kenneth, Mads, Nils		
41	The user doesn't understand the "Data network error"-message that they are presented with when thrown off the EQO network or failing to connect.		Camilla, Helle			
42	When adding an account there is no validation on the input fields and the user is not notified about illegal characters making the entered email address unusable for signing in (the user gets a "time out" notification after a while).		Camilla			
43	EQO doesn't connect after several minutes and the user receives no feedback.			Anne, Anders		
44	The user experienced sent messages not being delivered when chatting and wasn't notified about it.				Peter	
	Problem description (Navigation)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
45	After minimizing the program and made a call the user is forced to either maximize EQO or quit it. It is not possible to make another call for instance before you return to EQO.		Troels			
46	When adding a new contact the user expects that pressing "ok" brings him back to the contact list and is surprised that he has to confirm his	Claus, Thomas, Esben	Camilla, Julie, Lars, Mads	Anne, Kenneth, Nils		Kamilla & Dennis, Nanna & Niels

	action by pressing “mere/add”.					
47	The "mere" menu seems to be inconsistent. Some times the menu items are specific for the highlighted item and sometimes it is general. The user is confused.	Jakob, Claus	Camilla	Caroline, Kenneth		
48	The users have troubles understanding the purpose of the “sign in automatically” checkbox in the “add new MSN account” window.		Camilla, Julie, Lars, Mads, Troels, Helle	Anne, Caroline, Kenneth, Mads, Nils		
49	When deleting a contact the user thinks he must choose the contact before it can be deleted and therefore presses “vælg” instead of going through mere to delete it.			Anders, Caroline, Nils		
50	The user can’t edit the email address of his MSN account and has to delete it and add a new one in order to change it.		Camilla	Anders		
51	When pressing ok to the error message about being unable to send offline messages the user is returned to IM Session instead of where he came from (the contact list).					Nanna & Niels
52	In the 'mere'-menu in the IM Sessions window the menu item 'settings' has no funtionality when active sessions is highlighted.					Kamilla & Dennis
53	The user has to leave the chat window to write a message for the chat so he can't see what the contact writes while he writes himself.	Claus, Jakob, Thomas	Troels	Anne	Christian	Signe & Peter, Nanna & Niels, Kamilla & Dennis
54	The user is returned to "IM Services" after having set his status in the contact list instead of returning to the contact list.	Claus				Nanna & Niels
55	The user can sign in from "IM Services" but not sign out - which causes problems when the user wants to sign out where to user has to go through "set status".	Claus, Thomas, Esben, Jakob	Julie, Lars, Mads	Anders, Caroline, Nils		Signe & Peter
56	The user is not returned to the last place of origin when pressing back.	Claus, Thomas, Esben, Jakob	Camilla, Julie	Caroline, Kenneth, Nils		Kamilla & Dennis
57	Troublesome reconnection to EQO, where the user doesn't understand		Camilla, Lars, Troels			

	why he has to press "goto" to reconnect after losing the connection.					
	Problem description (Information)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
58	It can be hard to tell the participants in a chat apart because there is no difference in font or color.		Troels			
59	The user finds it strange that nothing on the contact list indicates that she is currently having a conversation with a contact. She expects an icon or some other kind of representation of the activity.			Anne		
60	The user is annoyed over the screensaver hiding the status of the application.	Thomas	Lars	Nils		Katrine & Allan, Nanna & Niels
61	The status of the user isn't apparent unless he is in "set status" or is offline in ”IM Services".		Camilla, Julie, Lars, Mads, Helle	Caroline		
62	Dependent on the color scheme of the phone the text is sometimes very hard to read because of the font and background colors are similar.			Mads		Signe & Peter, Kamilla & Dennis
63	In IM-services the number of active sessions are represented erroneously, so the user thinks he has more conversations going that he actually has.	Claus, Jakob				
64	In longer chat sessions the scroll bar blocks some of the text to the far right.				Klaus	
65	In the chat windows the display name of the contact is shortened to 7 characters, causing the user to be confused about who he or she is talking with.		Julie, Troels			Signe & Peter
66	When receiving a new message in another chat windows there is no visual feedback about it in the active chat window. Preventing the user to know about activity in other chats when the phone is set to 'silent'.					Katrine & Allan
67	The user can't tell contacts with the same display name apart (because the email and comments associated to each contact isn't visible).				Klaus	Signe & Peter, Nanna & Niels

68	When deleting a contact it is not removed from the contact list.	Esben	Mads, Helle, Julie			
69	Contacts become invisible on the contact list when they are being added and when they participate in a chat session.		Camilla, Lars, Mads			
70	When a new message is received the sound is too low to notice (if there is any?), forcing the user to keep an eye on the phone at all times.				Christian, Søren	Nanna & Niels
71	When losing network coverage the connection to the EQO-Network is lost and all active chat sessions are lost (experienced at home, in a train, in a car and in a bus)				Klaus, Peter, Søren	Signe & Peter
	Problem description (Conceptual)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
72	The user is nervous about using EQO because of high priced data traffic rates and the fact that EQO possibly produces traffic all the time while being online.					Kamilla & Dennis, Nanna & Niels
73	The user also doesn't see the point in using messenger on the mobile when she only is able to chat with the contacts that are online.			Caroline		
74	The user feels stressed by being online on mobile messenger and being accessible as if he were sitting in front of his PC.			Mads	Christian	Kamilla & Dennis
75	The long time it takes to connect to the EQO and messenger networks irritates the user.		Julie, Camilla, Lars, Troels	Anders, Caroline, Nils, Mads,		Nanna & Niels, Katrine & Allan, Kamilla & Dennis
76	The user compares the use of the system with texting (sms), but fails to see any advantages of EQO. They prefer texting because it is a well known, fast and inexpensive form of communication.			Anders, Anne, Caroline, Kenneth,		
	Problem description (Bugs)	**H. I.**	**Laboratory**	**Field**	**Diary**	**Video Probe**
77	When signed on there are no status icons in the contact list, which makes it impossible to know who is online and who is offline.				Klaus, Peter	

78	The "exit" button has no functionality when being asked to confirm the exit of EQO.	Esben				
79	After pressing "resume" when receive a new message the user was unable to access the new message under "active sessions" or being able to delete it in "active sessions" or access it through the contact list.	Esben				
80	When trying to update EQO the phone gets a white screen and crashes.				Christian, Klaus	
81	EQO can sometimes affect the functionality of the phone outside EQO so that a reboot is necessary for instance the phone can't show pictures after EQO has been active.				Peter	
82	Sometimes the system becomes unresponsive randomly when the user tries to: choose account, close a session, deleting a contact, and setting status. The system hangs for a while before it becomes responsive again.		Camilla, Lars, Mads			
83	The application freezes randomly and forces the user to reboot the phone.		Camilla, Lars	Nils, Mads	Christian	Kamilla & Dennis, Nanna & Niels

Summary

Inspired by prior research regarding the value of contextual usability evaluations compared to non-contextual, the initial goal of this thesis was to examine whether or not it is of any value to venture into the field when performing usability evaluations of mobile systems. Thus this thesis seek out to answer the question of whether or not field evaluations can provide evaluators with additional knowledge and possibly a better insight into problems that arise when users apply mobile system in the "wild".

The work presented in this thesis serves to shed some light on the current practices of the HCI research community regarding how and why researchers evaluate the way they do. Furthermore, a comparative study of the methods when applied to a mobile system, serves to outline the strengths and weaknesses of the most commonly used usability evaluation methods.

A literature review was conducted, in order to identify how and why professionals in the field of Human-Computer Interaction evaluate mobile systems as they do. Based on the results from the review it was concluded that researchers typically conduct usability evaluations of mobile systems by *expert evaluations, laboratory evaluations, field evaluation* or *longitudinal evaluations.* The first two methods are typically conducted without including the context of use while the latter includes the context of use.

Based on the knowledge provided by the literature review, a set of evaluation sessions were arranged, utilizing the four most commonly applied evaluation methods. By doing so the necessary first-hand knowledge was obtained and an analysis of the individual methods paved the way for a thorough comparison of the characteristics of the evaluation methods. It was eventually concluded that all methods possess different strengths and weaknesses and that no superior method could be identified. Thus in thesis researchers and practitioners will find not find an unambiguous answer to what single method they should apply.

Stimulated by the analysis of the applied methods, a new method was constructed in order to investigate if and how the strengths of different methods could be combined in order to overcome their weaknesses. The novel method was inspired by cultural probes and based on the experiences gained with the longitudinal and contextual evaluations. The resulting method – *the video probe evaluation* – was applied, evaluated and compared to the original four. In the end it was concluded that by overcoming known obstacles by combining known techniques, new obstacles arose. Thus the novel approach suffered the same sentence as the original four and was labeled as a valuable but not superior method.

Thus this thesis does not provide an answer to what methods should be applied to evaluate the usability of mobile systems. However the data that was collected and presented in this thesis heavily indicates that contextual evaluations are capable of providing evaluators and in the end developers, with valuable knowledge.