IT Licentiate theses
2009-003

Listen to Your Users

The Effect of Usability Evaluation on Software Development Practice

MARTA KRISTÍN LÁRUSDÓTTIR

UPPSALA UNIVERSITY
Department of Information Technology

UPPSALA
UNIVERSITET

Listen to Your Users

The Effect of Usability Evaluation on Software Development Practice

BY
MARTA KRISTÍN LÁRUSDÓTTIR

October 2009

DIVISION OF HUMAN-COMPUTER INTERACTION
DEPARTMENT OF INFORMATION TECHNOLOGY
UPPSALA UNIVERSITY
UPPSALA
SWEDEN

Dissertation for the degree of Licentiate of Philosophy in Computer Science with Specialization in Human-Computer Interaction at Uppsala University 2009

Listen to Your Users

The Effect of Usability Evaluation on Software Development Practice

Marta Kristín Lárusdóttir
marta@ru.is

Division of Human-Computer Interaction
Department of Information Technology
Uppsala University
Box 337
SE-751 05 Uppsala
Sweden

http://www.it.uu.se/

ISSN 1404-5117
Printed by the Department of Information Technology, Uppsala University, Sweden

Abstract

A vast majority of the people in the western world use software systems on daily basis for achieving their goals. To be able to do that each person needs to communicate what he or she wants to do to the system and receive a response. This communication needs to be easy for the user, especially when the system is new to him or her. Otherwise, the user either quits using the system; it takes a very long time or gets very irritated. A software team that is making new software needs to evaluate the usability of the system and various methods have been introduced in the literature to do that.

My research focus in this thesis is on usability evaluation. I study particularly, how usability evaluation methods can be compared, what data should be gathered in usability evaluation to gain knowledge on how the software affects users who are getting new software for their daily work and how useful this data is to the recipients.

Two experiments are reported in this thesis where results from using three different usability evaluation methods are compared. The main result from these two studies is that the think-aloud evaluation method should be used, if the goal of the evaluation is to gather as realistic information as possible on usability problems that the users will have when using the system.

Furthermore four case studies are described in the thesis, in which usability evaluation was done by using the think-aloud method in co-operation with real users in their real work situation. These studies give much richer information on the actual use of the systems involved.

The findings from one of these case studies indicate that the results from user observation done on a system that users have not seen before or used only for few days are rather similar to the results from usability evaluation done when users have used the system for a longer period. So the common practice of doing user observation on a software system that the participants have not seen before and then interpreting that the results will be the same for actual usage of the system when users will use the system for their real tasks for shorter or longer period is adequate.

Preface

This thesis consists of two sections. The first section contains a summary of my work on evaluation methods. The second section contains the full text of the papers that the summary is based on.

List of Papers Included in the Thesis

I Frøkjær, E., Lárusdóttir, M. K. (1999) Prediction of Usability: Comparing Method Combinations. Proceedings of "Managing Information Technology Resources in Organizations in the Next Millennium", Idea group publishing, Hershey, May 16 – 19, 1999, pg. 248 – 257.

II Hvannberg, E. Þ., Law, E. L., Lárusdóttir, M. K. (2007) Heuristic evaluation: Comparing ways of finding and reporting usability problems. Interacting with Computers, March 2007; 19 (2): 225 - 240.

III Hvannberg, E. Þ., Lárusdóttir, M. K. (2000) Usability Testing of Interactive Multimedia Services. Proceedings of the 1st Nordic conference of Computer-Human Interaction, Stockholm, October 23 – 25, 2000.

IV Þorgeirsson, T. and Lárusdóttir, M. K. (2007) Case study: Are CUP attributes useful to developers? Proceedings of the COST-294 open Workshop: Downstream Utility: The good, the bad and the utterly useless usability feedback, Toulouse, November, 6th, 2007, pg. 50 – 54.

V Lárusdóttir, M. K., Ármannsdóttir, S.E. (2005) A Case Study of Software Replacement, Proceedings of the International Conference on Software Development, University of Iceland, Reykjavik, May 27 – June 1, 2005 pg. 129 – 140.

VI Ísleifsdóttir, J., Lárusdóttir, M. K. (2008) Measuring the User Experience of a Task Oriented Software, Proceedings of the COST-294 open Workshop: Meaningful Measures: Valid Useful User Measurement, Reykjavik, June 18th, 2008, pg. 97-102.

My Co-authors

Sigrún Eva Ármannsdóttir	Executive director, Eskill Ltd., Iceland.
Erik Frøkjær	Associate professor, School of Computer Science, Copenhagen University, Denmark.
Ebba Þóra Hvannberg	Professor in Computer Science at the University of Iceland, Reykjavik, Iceland.
Jónheiður Ísleifsdóttir	Software specialist, Decode genetics Inc., Reykjavik, Iceland
Effie Law	Research fellow, University of Leicester, Britain.
Trausti Þorgeirsson	High school teacher, Reykjavik, Iceland.

Acknowledgements

Many thanks to my husband Guðmundur, for motivating me in the period of writing this thesis, making arrangements so this has been possible for me and for reading the thesis and giving me good comments.

Furthermore, I want to thank my supervisor Jan Gulliksen, for being encouraging and constructive. The advice from him, the discussions we have had and our co-operation has been valuable to me. I also appreciate his and his wife's hospitality.

Thirdly, I want to thank the HCI research group at Uppsala University for being very kind to their guest from Iceland. I have had a lot of great discussions with many of them and also nice time together with them during my visits in Uppsala. Special thanks to Åsa, Bengt and Mikael who gave me important feedback on the final draft of the summary and to Elina for arranging things concerning the format of the summary, the printing and the party.

In addition to that I want to thank my colleagues at Reykjavik University for being so positive, especially my dean Ari Jónsson, for making arrangements for me to be able to visit Uppsala every once in a while.

I want to thank my fellow authors for their contribution in this work and the COST-294 action for sponsoring a short term scientific mission for me at Uppsala University.

Finally, many thanks go to my friends and family who have all been very supportive. Especially I want to thank my friend Janet Read for reading the final draft of the summary and giving my valuable feedback.

It is a pleasure for me to experience how all this people have been willing to give of their time and energy and thereby help me to make the writing of the summary a positive experience.

Contents

Introduction

Most software systems that are made today are used by people, who instruct the system to do what they want to achieve by using a computer. The most common ways for the users of giving instructions are by pressing a button on the keyboard or using the mouse to select something on the screen. The system responds in some way, for example by displaying a text, changing the look of a button, printing something out or playing a sound to inform the user that it has processed the instruction. Then the user can make the next instruction to be able to solve his task. This interaction between the human user and the software system is made through the part of the system called the user interface.

A software system is designed by a software development team. This team chooses alternatives for the users to give instructions to the system and lays these out in the user interface. Sometimes the user has many possible ways of interacting with the system and the system can respond by giving more than one type of feedback. The members of the software development team need to take a lot of decisions on how the user interface should look and how it should react to the various instructions. So when the developers have designed one version of the user interface, how can they know that it is a good one? How is it defined what is good or bad in user interface design?

Luckily there is considerable agreement in the software development community that a good user interface is the one that is easy to use for the particular users to achieve their goals in the context of use. This means that the users should be able to complete the work they want to, with relatively little effort and the users should be satisfied after using the system. This also means that the development team needs to have information on who the users are, what they want to do, where they want to use the system and how they want to use it. For gathering this information the development team can make a careful analysis of the people that they presume will be the users of the system. They could analyze the characteristics of the users, the goals the users want to achieve and how these goals have been achieved in the past. When they have a clear idea about that, they start to draw the first sketches of the new user interface. To do that they can use design guidelines, could build on their own experience, and seek good design ideas from other systems, for example from some experts in user interface design. When they have a complete design, they can implement the software and deliver it to

the users. But how can the developers know that this process has ensured a good design?

The only way is to evaluate how easy the user interface is to use. This can be done by asking experts in user interface design to take a careful look at the user interface and list all the problems that the users could have while using the system. Another possibility is to watch real users using the new system and register the problems they really have when using it. While observing users, it can also be registered if the users could solve their tasks or not and logs can be made of how long time it takes to solve each task. Furthermore it is possible to ask the user how satisfied he is after using the system. There are a number of methods or techniques, called usability evaluation methods, which have been defined to measure the ease of use or usability. So how can a developer know what method to use?

To be able to guide developers while choosing an evaluation method, researchers have studied usability evaluation methods to some extent for the last 20 years or so. The main question in these studies is: “What is the best method for measuring usability?” To be able to answer that question, researchers have for example compared the outcome of using more than one method for evaluating one particular software system and compared the outcome from different evaluations. Usually information on usability problems is gathered during the evaluation and afterwards the numbers of problems found by each method are compared. The method that finds the highest number of problem is then generally declared to be the best method.

In recent years researchers have defined new ways of measuring the quality of usability evaluation. Studies have been done on how well the developers understand the information they receive from the evaluation and if the form of the information matters, as for example if they want written text to describe the results or if sketches of the user interface to describe the problems are of more use to them. Some studies have also been done on how much the evaluation is affected by the factors involved, like who is doing the evaluation, who is participating in the evaluation and what systems are evaluated to name a few.

My research focus in this thesis is on usability evaluation. Particularly I study how usability evaluation methods can be compared, what data should be gathered in usability evaluation to gain knowledge on how it affects users to receive a new software for their daily work and how useful this data is to the recipients. To summarize, the following three research questions are studied in the thesis:

1. How different are the results of using various usability evaluation methods for evaluating the same software system?
2. How useful are the results of usability evaluation to the recipients?
3. What impact does a new software system have on users for achieving their goals?

The Background

In this chapter the background to my work on usability evaluation is introduced and references to some literature on the subject are given. First the field of Human-Computer Interaction is explained in a few words, then definitions of usability are introduced and an overview of usability evaluation is given. Most of the commonly used usability evaluation methods are then described briefly and there is some discussion of the ways that these are compared in the current literature. The goal here is to give references to the most relevant literature on each subject in this chapter, but not to give a complete summary of everything that has been written on the matter.

Human-Computer Interaction

The field of **Human-Computer Interaction** (HCI) is defined in the curricula for HCI from 1992 by the Association of Computing Machinery – Special Interest Group for Computer-Human Interaction (ACM SIGCHI, 1992) as:

> "*Human computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and the study of major phenomena surrounding them*".

This definition focuses on the activities of the development team developing software systems, which in this thesis will be called developers. These activities include the design, the evaluation and the implementation of the software. Developers need to include users' perspective in all the mentioned activities. The definition also includes the study of phenomena surrounding users, which gives it a broader scope that only including the activities of the development team. The emphasis in this thesis is on one of the three fundamental activities in the definition: namely the evaluation.

In the ACM curricula for computer science (ACM, 2008) the importance of HCI is described by the following text:

> "*Human-computer interaction is an important area of computing knowledge. As more people conduct more of their daily activities interacting with a computer, the construction of interfaces that ease that interaction is critical for increasing satisfaction and improving productivity. As more software requires a user interface, knowing how to create a usable interface and testing*

the usability of that interface become required skills for all computer science students."

Usability

The term usability has several definitions which will be described in the following.

Usability According to the ISO 9241-11

Usability is defined in the ISO 9241 standard, part 11 (International Standard Organization, 1998) as:

> "*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*".

The definition includes four main elements: the user, the user's goals, the product, and the context of use. **The user** in this definition is always human, and he or she is: *"a person that interacts with the product"*. **The product** in the definition is most often the software that the user is using. But it can also include materials and hardware. It is *"the part of the equipment (software, hardware and materials) for which usability is to be specified or evaluated"*. **A goal** of the specified users is the intended outcome, it is often task oriented, so the users are for example solving their daily tasks by using software but the goal need not be task oriented and can also include activities like entertainment, games, etc. Finally **the context of use** includes *"the user, tasks, equipment and the physical and social environments in which a product is used"*.

There are three measureable elements in the ISO definition of usability: effectiveness, efficiency and satisfaction. **Effectiveness** is *"the accuracy and completeness with which users achieve specified goals"*. This means that it is measured by assessing the extent to which a user can complete his or her goal (completeness) and by noting how accurate the outcome is (accuracy). So if the user wants to insert a particular data into a system, he or she could complete that task, but make typos or select a wrong alternative from dropdown lists without knowing it. In this example, the completeness is fine but the accuracy is bad. **Efficiency** is defined as the resources expended in relation to effectiveness. It is very common to consider the time it takes to complete a task as a good measure of the resources expended, other measures can include keystrokes or mouse clicks. **Satisfaction** is *"freedom from discomfort, and positive attitudes towards the use of the product"*. Satisfaction is often measured by asking the user to rate his or her satisfaction with the

system by using a questionnaire but can also be investigated by looking at the emotional state of the user.

Other Definitions of Usability

There are other definitions of usability which are referred to in current literature. Nielsen (Nielsen, 1993) defined usability by five usability attributes:

Learnability:	*The system should be easy to learn.*
Efficiency:	*The system should be efficient to use for expert user's steady-state level of performance.*
Memorability:	*The system should be easy to remember, so that a casual user is able to return to the system after some period of not having used it, without having to learn everything all over again.*
Errors:	*The system should have low error rate.*
Satisfaction:	*The system should be pleasant to use.*

Nielsen states that learnability is the most fundamental usability attribute, since the first experience many people have with a new system is that of learning how to use it and in most cases one can't afford to spend much money on learning to use a system.

Aside from satisfaction, these five attributes are quite different from the three attributes in the ISO 9241-11 standard. In annex B of the ISO 9241-11 standard, it is described how learnability, error tolerance and memorability can be estimated by using measures of effectiveness and efficiency.

The utility of a system is to what extent the system provides the functions needed for the user. Utility is not part of usability in Nielsen's definition of usability but in the ISO-9241 it is. Nielsen's definition is therefore sometimes called the "small" usability, in contrast to the ISO-9241 definition which is the "big" definition.

Quesenbery (Quesenbery, 2003) defines the five E's for describing usability, which are:

Effective:	*The completeness and accuracy with which users achieve their goals.*
Efficient:	*The speed (with accuracy) with which this work can be done.*
Engaging:	*How pleasant, satisfying or interesting an interface is to use.*
Error Tolerant:	*How well the product prevents errors, and helps the user recover from any those do occur.*
Easy to Learn:	*How well the product supports both initial orientation and deeper learning.*

Quesenbery states that the definition of usability from the ISO 9241-11 standard has done little to help to sell usability. The major criticism is that the ISO 9241-11 definition is too focused on well-defined tasks and goals, so it makes it difficult to talk about how usability applies to products or context where these are less important, like for systems emphasizing pleasure or engagement.

Measuring Usability by Counting Usability Problems

One measure of usability which is commonly used is a count of usability problems. **A usability problem** is defined by Stone and colleagues as (Stone, et. al. 2005):

> *"a difficulty in using the user interface design that affects the user's satisfaction and the system's effectiveness and efficiency. Usability problems can lead to confusion, error, delay or outright failure to complete some task on the part of the user. They make the user interface, and hence the system, less usable for its target users."*

So if the user has problems using the system, it will probably take him or her longer time than expected or even worse, she or he may have to quit using the system before completing the task. If the user has many problems he or she will probably become irritated. The severity of a problem is usually rated according to how much the problem affects the user (Nielsen, 1993) (Molich, 2000) and the frequency with which the problem is encountered. Often three categories for **usability problem severity** are used as defined by Nielsen in (Nielsen, 1993):

a) A minor problem is when the problem has little impact on few users.

b) A medium problem is either a problem that large impact on few users, or a problem that has little impact on many users.

c) A high severity problem is the one that has large impact on many users.

Broader View of "Easy to Use"

Driven by the impression that the definitions on usability described above are too focused on task- and work-related attributes, the term user experience has gained momentum in HCI (Hassenzahl, Tractinsky, 2006). Despite the growing interest in user experience, it has been hard to gain a common agreement on the nature and scope of user experience (Law, et. al. 2009). The study of **user experience** highlights non-utilitarian aspects of user interactions, shifting the focus to user affect, sensation, and the meaning as well as value of such interactions in everyday life. The methods for evaluating user experience are still inadequate according to Obirst and colleagues (Obirst, et. al. 2009).

Caroll (Caroll, 2004) also states that the definition of usability is too narrow. His point is that people have to want to use the system and continue to do so. Part of that is that it has to be fun, he states. In his concluding remarks he points out that the concept of usability is under continuous construction.

Choosing Usability Measures

In his study of usability measures used in practice, Hornbæk (Hornbæk, 2006) divides measures of usability in two categories: *the subjective* usability measures, where the users' perception of or attitude towards the interface, the interaction or the outcome is measured and *the objective* usability measures where the aspects of the interaction not dependent on the users' perception are measured. His advice is to pay special attention to whether objective or subjective measures are appropriate, and whether a mix of those two better covers the various aspects of quality in use.

Usability Evaluation

To **evaluate** is according to Merriam-Webster dictionary either "to determine or fix the value of"; or "to determine the significance, worth, or condition of, usually by careful appraisal and study" (Merriam-Webster online dictionary, ND). Usability evaluation is thus to determine or measure the usability of the software system and as discussed earlier usability can have different definitions. The ISO-9241 definition could be called the "traditional" definition and is widely used.

The term **usability evaluation** is here used to describe the complete test of the UI, including planning the evaluation, conducting the evaluation sessions and presenting the results. An **evaluation session** is when the user interface is evaluated with a single participant either a single user, if the evaluation is done with users participating or a single evaluator, if the evaluation is done through expert analysis. An **evaluation method** describes the process used in each evaluation session.

There are various factors involved in the context of the usability evaluation: the goal of the evaluation, the system evaluated, the evaluation method used, the evaluation data gathered, the data collection, the evaluators, the participants in the evaluation, the recipients of the results, and the environment of the evaluation, as can be seen on figure 1 and will bescribed below. This selection is based on guidelines on how to run a usability evaluation as found in (Mayhew, 1999) (Kwahk, Han, 2002) (Stone et. al. 2005) and on the ISO/IEC 25026 standard (ISO/IEC 25026, 2006).

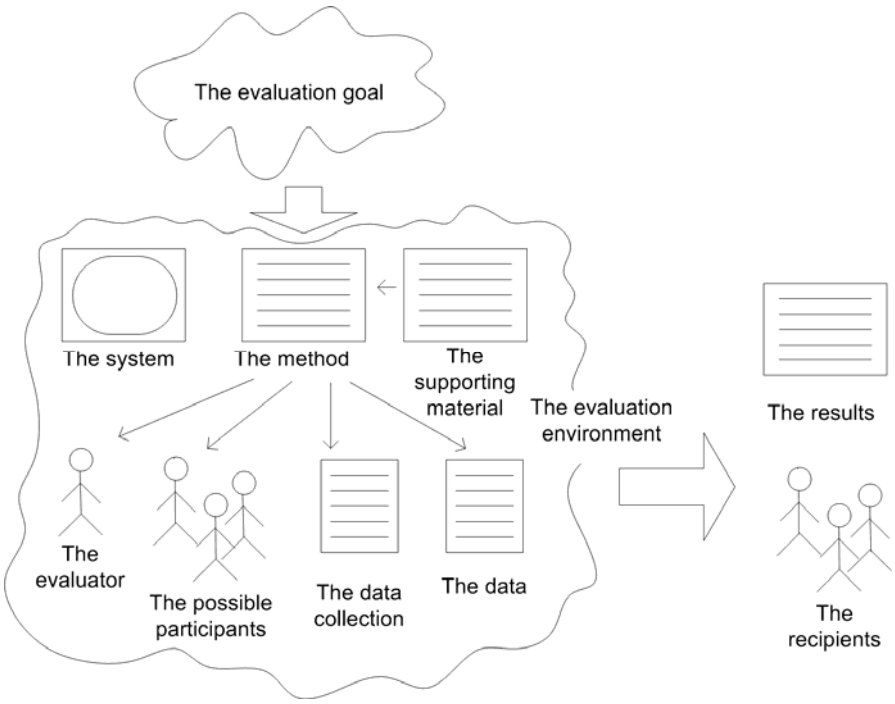

Figure 1. The factors of the context of usability evaluation

The Goal of the Evaluation

The goal of usability evaluation is always to measure the usability or some factors of usability, but there are mainly two categories for how the measurements are used. Either the results are used to give developers feedback on their current version of the user interface design which is still under construction, called **formative evaluation** or the results are used to assess the success of a finished product called **summative evaluation** (Preece et. al., 2002). The choice of the evaluation method depends vastly on the goal of the evaluation.

In a paper by Hertzum and Jacobsen (Hertzum, Jacobsen, 2001) it is stated that "*vague goal analysis prior to usability evaluation leaves many decisions about which aspects of the system to include in the evaluation to the evaluator's discretion*". It is stated that this could lead to considerable variability in the evaluator's final choice of evaluation tasks. The authors recommend that evaluators verify the coverage of their task scenarios in a systematic way to ensure that all relevant system facilities are considered for inclusion in the evaluation.

In a study by Hornbæk and Frøkjær, (Hornbæk, Frøkjær, 2008) an experiment was carried out to determine how business goals would affect the results of think aloud tests. Half of the evaluators were instructed to take business goals into consideration while planning and reporting the results of

the evaluation, so for those evaluators one factor of the evaluation goal was to use the business goals. The results of the study show that usability problems found by the group that included business goal in the evaluation are assessed higher by a company commissioning the evaluation. This indicates that evaluation goals are extremely important in the entire process of evaluation and should be studied further.

The System Evaluated

Software systems are used by diverse user groups to achieve various goals which can range from solving work related tasks to having fun. For getting information on how these systems fit the users' needs, all these types of systems have to be evaluated.

The status of the system in the development process can also be diverse. Paper prototypes of the first ideas of the system can be evaluated with the same evaluation method as running systems that have been installed in the real context of use. Some evaluation methods are difficult to use on paper prototypes though.

A recent study (Lim et. al., 2006) compared the types of usability problems that can be found in three types of prototyping techniques - a paper based, computer-based and a fully functional one. The results show that the unique characteristics of each different prototype affect the usability evaluation in different ways.

One aspect of doing the evaluation is the equipment which is used to do the evaluation. For paper prototypes this could be physical object of foam that the paper prototypes are glued on or it could some equipment like a phone or a computer that the paper prototypes are "used" on, to make the context of use for the participant more realistic. For digital prototype the version of the system running the prototype needs to be described.

The Evaluation Method

The evaluation method is the process by which the evaluation is done. As defined in the Merriam-Webster dictionary, a **method** is: "*a way, technique, or process of or for doing something*" (Merriam-Webster online dictionary, ND). Some evaluation methods are very structured, like the cognitive walkthrough method, other are a less structured, like a method called peer review. Some HCI specialists like for example Stone and colleagues (Stone, et. al., 2005) would like to use the word technique as opposed to the word method when talking about the evaluation process, because many of the evaluation methods are not very structured. In this thesis, the word method is used, because that is commonly used in by researchers in the field, se for example (Cockton, Woolrych, 2009). Evaluation methods will be described in more detail in a separate subchapter.

The Supporting Material

Some methods are based on using supporting material, like usability guidelines, standards or predefined user's tasks. Some method descriptions specify that a particular supporting material should be used, but for other methods the evaluator can choose which supporting material is used in the evaluation.

The Data Gathered

The data gathered depends on the goal of the evaluation and the method selected. Some methods are used to gather information on usability problems, while using other methods it is possible to measure the effectiveness and efficiency of task completion. Information on user's satisfaction is usually gathered through a questionnaire.

The Data Collection

The form of the data collection can be writing the results on paper, using a tool to write the results, video recording everything that happens, voice recording the things that are said during the evaluation, recording what happens on the screen with a specialized software and logging everything that happens in the software to name the most common ones.

After the data has been collected the results need to be analyzed and communicated to the recipients. This can be a very time consuming task in the evaluation. Some researchers have focused on finding low-cost techniques for analyzing data, like Kjeldskov and colleagues (Kjeldskov, et. al. 2004). In their paper they suggest a technique called instant data analysis, which allows evaluations to be conducted, analyzed and documented in a day, where 85% of the critical usability problems in the system were identified by using the method in 10% of the time required to do video data analysis grounded on data from the same user evaluation.

Some studies have been done on what form of feedback fits the development team best. One recent study (Hornbæk, Frøkjær, 2005) compared how developers of a large web application assess usability problems and redesign proposals as input to their system development. The results show that the developers assessed redesign proposals to have higher utility in their work than descriptions of usability problems. The authors suggest that redesign proposals are used as an integral part of usability evaluation.

The People Involved

Usability evaluation needs judgment from a human, so it is not possible to evaluate the usability automatically. In any usability study there are always some human recipients of the evaluation results and these have to understand

and use the results. Furthermore users or representatives of users take part, when experimental evaluation methods are used. In the following the roles of these people are described.

The Evaluators

The evaluator plans the evaluation, conducts it and analyzes the results from the evaluation. During the planning the evaluator prepares all the material that needs to be ready for the evaluation, contacts the people involved and makes sure that all the facilities are available for the evaluation. While conducting the evaluation, his responsibility is to make sure that the process described by the method is followed. Sometimes it is also his responsibility to report the results from the evaluation.

Obviously it is important that the evaluator has good knowledge of the factors involved in the evaluation, particularly the evaluation method used and the user's goals. Many studies have shown that the evaluator's knowledge of the evaluation method affects the results from the evaluation to a large degree. Hertzum and Jacobsen, (Hertzum, Jacobsen, 2001) did a review of eleven studies of the evaluator's effect while using three methods, cognitive walkthrough, heuristic evaluation and the think-aloud method. Their main conclusion is that the average agreement between any two evaluators who have evaluated the same system with the same method ranges for 5% to 65% and no one of the three methods is constantly better than the others. The authors question the fact that think-aloud evaluations with one evaluator are often used as authoritative statements in the literature on evaluation methods.

In a study on heuristic evaluation, (Nielsen, 1992) Nielsen studied how much the evaluator's knowledge of the usability and the evaluator's knowledge of the system's domain affected the evaluation results. Usability specialists were better than novice evaluators in finding usability problems in that study and the double experts that had good knowledge of usability and the kind of interface being evaluated were even better.

In a recent study, (Molich, et. al, 2004) the consistency of usability testing across 9 industrial organizations is reported. The evaluators found 310 different usability problems in total and only two problems were reported by six or more organizations, while 75% of the problems were uniquely reported, that is that were only reported by one team. The authors conclude that the assumption that if evaluators use the same method they will obtain the same results is plainly wrong.

To be able to get reliable results it would be best, if the evaluators using a particular evaluation method were usability experts and domain experts. This is not always possible. For example in Iceland, there are not that many usability experts, so teaching all computer science students how to evaluate has been emphasized. Often evaluation also is part of their other tasks like programming and designing the system.

The Recipients

It depends on the evaluation goal who the recipients of the results from the evaluation are. These recipients can be the developers, the HCI specialists responsible for the design of the interface or others like researchers or managers. In all cases, the recipients need to be informed about the results of the evaluation so they can understand the results, find solutions and prioritize those (Wixon, 2003).

One way of determining the success of usability evaluation is to look at the downstream utility, which is defined by Law (Law, 2006) as:

> *"The extent the improved or deteriorated usability of a system can directly be attributed to fixes that are induced by the results of usability evaluations performed on the system".*

Here the quality of the usability evaluation is determined by how much it improves the actual usability of the system and not by how many problems are found. Researchers do not agree on the scope of usability evaluation, so Cockton (Cockton, 2005) argues that assessing the downstream utility is beyond the scope of pure evaluation methods.

The Participants – the Users

When experimental evaluation methods are used, people are asked to participate. The participants in the evaluation are either the real users of the system being evaluated or representative of the users. Sometimes when system are still being developed it is not quite clear which user group the system will have, but sometimes it is a well defined group.

The expected use of the system can also vary. Some systems will be used for a short time and rarely, so the users will not remember how to use the system from time to time. But some systems will be used all day long, the whole year by particular users, so these will become experts in using the system. In her PhD thesis, Liu (Liu, 2009) studies the effect of involving different user groups in usability tests by analyzing differences in the users' interactions with both simple and complex interfaces. Her main finding is that the effect of user's expertise may be invisible when interacting with a simple user interface, but the expert users outperformed the novice users when interacting with a complex interface.

The Environment of the Evaluation

An evaluation can take place in various surroundings. The most extreme ones are doing an evaluation in the real setting (field), where the system is actually used versus doing an evaluation in a usability laboratory (lab), where the environment is controlled by the evaluator.

In a study by Nielsen and colleagues (Nielsen et. at., 2006) a comparison is made of a field-based evaluation and a lab-based one. The results show that significantly more usability problems were found in the field than in the laboratory setting and the field setting revealed problems with interaction style and cognitive load that were not identified in the laboratory.

Other Aspects

A factor that has not been mentioned yet is the constraints which affect the evaluation. It is stated by Stone and colleagues (Stone et. al. 2005) that formulating the constraints is the most important subject while formulating an evaluation strategy. Some examples of evaluation constraints are: money, timescales and availability of equipment, participants or evaluators.

Usability Evaluation Methods

For the last 20 years or so many methods for evaluating user interfaces have evolved and there are many alternatives to categorize the methods. Some authors (Whitefield, et. al. 1991) (Dix et. al., 2004) (Barkhuus, Rode, 2007) categorize the methods according to whether users participate in the evaluation or not and that is also done here.

Evaluation through User Participation

There are mainly three categories of evaluation methods where users participate in the evaluation (Preece, et. al. 2002), (Dix, et. al., 2004): a) methods for testing users' performance, b) methods for observing users, and c) methods for asking users their opinions.

When testing users' performance, the users are asked to perform some predefined tasks in controlled settings and the performance is measured. Typically the time it takes to solve each task is measured, the number of errors made is logged and the navigation path through the interface is tracked. Methods for observing users do always include an observer that observes how a user interacts with a system. Some observation methods require the user to carry out predefined tasks. Of these methods, the best rated by practitioners in Sweden is the think-aloud method (Gulliksen, et. al., 2004), that will be described further below and is studied in all the papers in the thesis. The main methods for asking users for their opinions are interviews and questionnaires (Preece, et. al. 2002), (Dix, et. al., 2004). Sometimes more than one method is used for evaluating through user participation in one evaluation session.

Think-aloud Method

In a think-aloud session one user at a time is asked to solve predefined tasks and talk while he is interacting with the system, so the evaluator can understand how the user thinks about the system while using it. The evaluator conducts the evaluation by handing out the relevant material. In the task solving session, he hands out the tasks one at a time and if the user is not talking the evaluator should encourage the user to say what he is thinking. Sometimes another person observes the think-aloud session too, for gathering observational data and sometimes data is also gathered through recording. There are mainly five steps in a think-aloud session: a) greeting the participant, b) data gathering on the participants background, c) the participant's interaction with the system solving predefined tasks, d) debriefing from the participant and finally e) the participant is thanked for coming.

The think-aloud method was first introduced in software development around 1980, (Lewis, 1982), but still there is no definite definition to the aim and the process of the think-aloud method according to Hertzum and Jacobsen, (Hertzum, Jacobsen, 2001). There are numerous variations of the method that have been employed, but the authors state that it is the single most important method for practical evaluation of user interfaces. That is also confirmed by a survey from Sweden (Gulliksen, et. al., 2004), where usability professionals were asked to rate the usability methods that they had used. There the think-aloud method received the highest score out of 25 methods mentioned.

Actually all three categories of evaluation through user participation could be used in one think-aloud session, that is: a) getting users opinion - the user could be interviewed for getting background information, and after the task solving the user could be asked to answer a questionnaire on his satisfaction with the system, b) observing the user - the user is observed during the task solving and c) measuring the performance - the performance during the task solving could be measured.

A typical usability evaluation using the think-aloud method consists of evaluation sessions with 5 to 10 participants. Each evaluation session lasts typically for one to two hours. To lower the cost, it has been stated by Nielsen that the best results come from the first 5 users (Nielsen, 2000). However, Woolrych and Cockton (Woolrych, Cockton, 2001) argue that the number of participants needed to obtain good results from the evaluation depends on the individual differences between test users, the tool tested and the tasks performed during testing.

Evaluation through Expert Analysis

It can be expensive to carry out usability evaluations with user participation. Consequently, a number of methods have been proposed over the last 20

years or so for evaluation through an expert analysis where users are not involved in the evaluation. The most commonly described methods in HCI literature are heuristic evaluation and cognitive walkthrough (Preece et. al., 2002), (Dix, et. al. 2004), (Stone et. al., 2005). These will be described in detail below.

In evaluation through expert analysis, the experts inspect the interface and assess the impact it would have on the particular users. The experts use interface guidelines, user interface standards, the users' tasks or their own knowledge, depending on the method, to inspire them to find possible problems that the users would have if they were interacting with the system. Because the experts are guessing what problems the users would have, there is a certain risk that they will describe issues as problems that the users do not have trouble with in real use. These issues are called false problems.

Furthermore using models to predict user performance is an expert-based approach to evaluation. These techniques are successful for systems with limited functionality such as telephone systems. The keystroke level model and GOMS are best known in this category (Preece, et. al. 2002).

Heuristic Evaluation

An expert uses a small set of guidelines or heuristics when evaluating the user interface with the heuristic evaluation method. Two sets of guidelines, from Nielsen (Nielsen, 1993) and Gerhardt-Powals (Gerhardt-Powals, 1996) that can be used in heuristic evaluation are shown in table 1.

Table 1. *Overview of Nielsen's and Gerhardt-Powals' guidelines*

Nielsen's guidelines	**Gerhardt-Powals' guidelines**
1. Visibility of system status	1. Automate unwanted workload
2. Match between system and the real world	2. Fuse data
3. User control and freedom	3. Present new information with meaningful aids to interpretation
4. Consistency and standards	4. Use names that are conceptually related to function
5. Error prevention	5. Limit data-driven tasks
6. Recognition rather than recall	6. Include in the displays only that information needed by the user at a given time
7. Flexibility and efficiency of use	7. Provide multiple coding of data when appropriate
8. Aesthetic and minimalist design	8. Practice judicious redundancy
9. Help users recognize, diagnose, and recover from errors	
10. Help and documentation	

Nielsen's guidelines were used in paper I and II in this thesis and Gerhardt-Powals' guidelines were used in paper II.

The procedure of heuristic evaluation involves having a small group of evaluators examine the software individually. Afterwards the group meets and aggregates the results to one aggregated list of usability problems found.

The heuristic evaluation was first introduced by Nielsen and Molich in 1990 (Nielsen, Molich, 1990), and described further by the authors in 1992 to 1994 (Nielsen, 1992), (Nielsen, 1993), (Nielsen, Mack, 1994). Around 65% of usability professional in Sweden rate heuristic evaluation as very good or fairly good for user interface design (Gulliksen, et. al. 2004).

Cognitive Walkthrough

Cognitive walkthrough is an evaluation method that focuses on the user's cognitive activities, goals and knowledge, when he is learning to solve particular tasks by using the software. During the walkthrough the evaluator step through the actions in the user interface that are needed to solve a task and evaluates the possible usability problems that could occur.

Usually the cognitive walkthrough is done in pairs of two evaluators who tell a believable story about why this particular step in the interface is or is not good for the user. The evaluators agree on the tasks, interface descriptions and user background before the evaluation. For each step the evaluators answer four questions: a) Is the effect of the action the same as the user's goal at that point? b) Will users see that the action is available? c) Once the users have found the correct action, will they know it is the one they need? d) After the action is taken, will users understand the feedback they receive?

The method was first described in 1990 (Lewis, et. al. 1990) and evolved through various versions (Polson, et. al. 1992), (Wharton, et. al. 1992) and (Wharton et. al., 1994), that were more or less structured and tedious to apply. The most widely used is the 1994 version (Blandford, A. 2007).

Evaluating Usability Evaluation Methods

Shortly after the first expert-based methods were defined, researchers were interested in measuring the effectiveness of using various methods and finding the advantages and disadvantages of using them. Five comparative studies were published in the years 1990 to 1993, (Jeffries, et. al. 1990), (Karat, et. al. 1992), (Nielsen, 1992) (Desurvire, et. al. 1993) (Nielsen, Phillips 1993). In all these studies the effectiveness of the methods is measured by counting usability problems found and the severity of the problems found is studied. Three of these studies are better described and discussed in paper I. Furthermore a study by Cuomo and Bowen from 1994 (Cuomo, Bowen, 1994) is also discussed there. In these studies an aggregated list of all problems found during user observation is made and used to describe all usability problems that can be found in the system. Then the number of problems found by using another method is compared to the aggregated list.

In 1998 Gray and Salzman, (Gray, Salzman, 1998) published a provocative paper, where the authors found methodological flaws in all the five studies published in 1990 to 1993. They claim that the studies suffer from two basic problems:

a) *It is not clear that what is being compared across the evaluation methods is their ability to assess usability.*

b) *The design of many of the experiments is such that neither the data they produce nor the conclusions drawn from the data are reliable or valid.*

The authors recommend that researchers in the field pay close attention to experimental design in their studies on evaluation methods. They also suggest that usability problems found in an evaluation are categorized in the following four categories:

a) *Problem is found and it is a true problem – called a hit*

b) *Problem is found but no problem exist – called a false alarm*

c) *A problem is not found but it exist – called a miss*

d) *A problem is not found and does not exist – called a correct rejection.*

Then the question is: "*When our usability evaluation method claims that something is a problem, how confident are we that this claim is a hit rather than a false alarm?* It is common to presume that problems found while observing users are true problems that users would have in real use. Problems found by using another method are compared to the list of problems found in user observation to calculate the effectiveness of the method.

Sears (Sears, 1997) describes measures for studying evaluation methods:

Reliability: Evaluators want consistent usability evaluation results, independent of the individual performing of the usability evaluation.

Thoroughness: Evaluators want results to be complete; they want evaluation methods to find as many of the existing usability problems as possible.

Validity: Evaluators want results to be "correct"; they want evaluation methods to find only problems that are real.

Later Hartson and colleagues, (Hartson, et. al. 2001), defined formulas for calculating two of these attributes, the thoroughness and the validity, as:

Thoroughness = Hits/ (Hits + misses)

Validity = Hits / (Hits + False alarms)

They also added a new one, called effectiveness, which:

*Effectiveness = Validity * Thoroughness*

These calculations are used in paper II to compare the results from using different usability evaluation methods.

Overview of the Papers

This summary is based on the following papers, which are referred to in the text by their Roman numerals.

Paper I	**Prediction of Usability: Comparing Method Combinations.**
Authors	Frøkjær, E., Lárusdóttir, M. K.
Publication	Proceedings of "Managing Information Technology Resources in Organizations in the Next Millennium", Idea group publishing, Hershey, May 16 – 19, 1999, pg. 248 – 257.
Short Summary	This paper is a presentation of an experiment where the effectiveness for uncovering and assessing usability problems using 3 evaluation methods, cognitive walkthrough, heuristic evaluation and think-aloud tests were compared.
The Context	The evaluators were computer science students which evaluated in groups of 3 evaluators. For the heuristic evaluation they used Nielsen's guidelines. They made the tasks for cognitive walkthrough and the think-aloud method themselves. In the think-aloud tests there were 3 users, who were also computer science students. They were asked to measure usability by registering usability problems in a running prototype of a GUI system on paper. The evaluators chose the evaluation environment. The recipients of the results were the authors of the paper and one of those was their lecturer.
My Contribution	The experiment was my idea and conducted during my Master studies in Copenhagen University. I made the research plan, conducted the experiment and analyzed the data under the supervision of my co-author. I wrote a Master thesis which the paper is based on. We wrote the paper together. The authors were listed in alphabetic order.

Paper II	**Heuristic evaluation: Comparing ways of finding and reporting usability problems.**
Authors	Hvannberg, E. Þ., Law, E. L., Lárusdóttir, M. K.
Publication	Interacting with Computers, March 2007; 19 (2): 225 - 240.
Short Summary	The aim of this paper is to refine a research agenda for comparing and contrasting evaluation methods, heuristic evaluation and the think-aloud method. To reach this goal, a framework is presented to evaluate the effectiveness of different types of support for structured usability problem reporting. This paper reports on an empirical study of this framework that compares two sets of heuristics, Nielsen's heuristics and the cognitive principles of Gerhardt-Powals, and two media of reporting a usability problem, i.e. either using a web tool or paper.
The Context	The evaluators using the heuristic evaluation were novice evaluators and they selected the environment for the evaluation. The evaluators applying the think-aloud method had good knowledge of HCI. The think-aloud sessions were done with real users in their real environment. Both groups were asked to measure usability by registering usability problems in a running web application. The recipients of the results were the authors of the paper.
My Contribution	I supervised the students that gathered the data from the evaluations and took active part in planning the experiment and analyzing the data. I was active in writing the paper.
Paper III	**Usability Testing of Interactive Multimedia Services.**
Authors	Hvannberg, E. Þ., Lárusdóttir, M. K.
Publication	Proceedings of the 1st Nordic conference of Computer-Human Interaction, Stockholm, October 23 – 25, 2000.
Short Summary	Two service trials were conducted where thirty families had access to video-on-demand, news-on-demand and worldwide web services for three months. The users had access to the services via a set-top-box connected to a television or via a personal computer. The paper describes how three methods: logging, the think-aloud

	method and questionnaires were applied as well as their results. The experiences from using the methods are discussed.
The Context	The think-aloud sessions were conducted by a usability expert and data was gathered by observation and note taking. The 10 participants were representative of the user group and the evaluation was done in their own environment. In the think-aloud sessions usability problems were registered as a measure of usability. Usage was recorded by logging and users satisfaction by using a questionnaire. The recipients of the results were the scientists and the developers of the system.
My Contribution	I planned and conducted the evaluation using the think-aloud method and the use of the questionnaires. We analysed the data and wrote the paper together. The authors were listed in alphabetical order.

Paper IV	**Case study: Are CUP attributes useful to developers?**
Authors	Þorgeirsson, T. and Lárusdóttir, M. K.
Publication	Proceedings of the COST-294 open Workshop: Downstream Utility: The good, the bad and the utterly useless usability feedback, Toulouse, November, 6th, 2007, pg. 50 – 54.
Short Summary	In this paper a case study of a classification scheme for usability problems called CUP (Classification of Usability Problems) is described. The individual attributes are analyzed according to how helpful they are for developers to understand, prioritize and fix a defect. Additionally factors are analyzed that determine whether developers decide to fix a defect or not.
The Context	The 10 think-aloud sessions were conducted by a usability expert in the users own environment on a version of a web application which should be installed two weeks later. Another usability expert observed the evaluation sessions and took notes. The usability was measured by registering usability problems, task completion and task time. The recipients of the results were the developers of the application.

My Contribution	I planned and conducted the evaluation using the think-aloud method and the use of the questionnaires. I also conducted the interviews with the developers. Both co-authors took active part in analyzing the data and writing the paper.

Paper V	**A Case Study of Software Replacement**
Authors	Lárusdóttir, M. K., Ármannsdóttir, S.E.
Publication	Proceedings of the International Conference on Software Development, University of Iceland, Reykjavik, May 27 – June 1, 2005, pg. 129 – 140.
Short Summary	This paper describes a case study, where the impact of introducing to users a new Windows software system to replace an existing one was measured in the users' own environment. First the old system was evaluated, then the new one shortly after introducing it to users and again after six months usage.
The Context	The think-aloud sessions were conducted with the real users in their environment by a usability expert. The applications were running software that the users had used for some period. Another usability expert observed the sessions and took notes. Usability was measured according to the ISO 9241 definition. The recipients of the results were the manager at the company involved.
My Contribution	I planned and conducted the evaluation using the think-aloud method and the use of the questionnaires. I wrote the paper but my co-author contributed with comments on the paper.

Paper VI	**Measuring the User Experience of a Task Oriented Software**
Authors	Ísleifsdóttir, J., Lárusdóttir, M. K.
Publication	Proceedings of the COST-294 open Workshop: Meaningful Measures: Valid Useful User Measurement, Reykjavik, June 18th, 2008, pg. 97-102.
Short Summary	In this paper a study on a web based tool is described that is used to keep track of attendance and work schedules by employees and managers in large companies. Ten users participated in the think-aloud sessions measuring the us-

ability of a new version of the software and the user experience was measured before and after each think-aloud session.

The Context	The context was the same as described in paper IV. In this paper measurements on user experience were gathered before and after the solving tasks in think-aloud sessions.
My Contribution	This experiment was my idea. I planned and conducted the measurement of the user experience. My co-author analyzed the data and wrote the paper under my supervision and I contributed with comments.

Method

In the papers various research methods are used, as can be seen in table 2. In this chapter the use of each method will be described.

Table 2: Overview of the research methods used in the papers

	Paper I	Paper II	Paper III	Paper IV	Paper V	Paper VI
Experiment	x	x				
Questionnaires	x	x		x	x	x
Observation in think-aloud sessions			x	x	x	
Informal interviews			x	x	x	
Structured interviews				x		

Experiment

In paper I and II an experimental design was used to compare the results from evaluation using three different usability methods.

In paper I, 51 computer science students used first either heuristic evaluation according to Nielsen's guidelines (Nielsen, 1993) or cognitive walkthrough for one week in groups of 3 evaluators. Two systems were evaluated; an experimental text retrieval system called TeSS was evaluated by 33 students and a graphical text editor Asedit was evaluated by 18 students. After that each group handed in a usability problem list. Then all the groups evaluated the TeSS for one week using the think aloud method and handed in the results. This experiment was part of a grading in a HCI course that the students were having, therefore they were quite motivated to do a thorough evaluation. The evaluators had not used the methods before but obtained the detailed material describing the methods. The students made the tasks for the think aloud sessions themselves and they chose the evaluation environment. The organization of the think aloud sessions was in that way that one group of students help another group by participating as users and vice versa. This experimental design included various opportunities for comparing the results

both from using one method and combining the results from using one method after the other.

In paper II results from using two types of guidelines in heuristic evaluation, Nielsen's guidelines (Nielsen, 1993) and Gerhart-Powals' guidelines (Gerhardt-Powals, 1996) are described. Furthermore results from reporting usability problems using paper and with the help of a web tool are compared. As a result there were four combinations of context of the evaluation, because there were two variations of the supporting material for the heuristic evaluation used and two variations of registering the problems. The results from these four types of evaluation context for the evaluation through expert analysis were compared to the results from the user observation.

The 20 evaluators using the heuristic evaluation were computer science students who had not used the method before. They were asked to participate in a research project, but did not obtain any grading or reward after taking part in the study. The think aloud tests were done by 2 other students who had more knowledge of the method than the evaluators doing the inspection. They ran 10 think aloud sessions with predefined tasks that were defined by the researchers specially to be able to find the problems found in heuristic evaluation. The evaluation project was part of a course the students were taking and they were highly motivated to do the study.

To be able to evaluate how many problems were found by using each method, all the usability problems found were gathered in one joint list in paper I, II and IV. The process of making the joint list is not trivial. First all the problems found are gathered in one list and then for each of the problems, it is decided if it is the same as some other problem on the list, or if it is a unique problem. In paper II a systematic approach to make the joint list was taken as described in Connell and Hammond (Connell, Hammond, 1999); in the other papers it was systematic even though a pre-described process was not chosen. The researchers had to make their own judgment of which two problems were the same in all the papers. Furthermore the severity of each problem was judged by the researchers. When comparing results from one study to another the process of making the joint list and judging the severity has to be similar to make a fair comparison of the effect of using one method or the other.

Questionnaires

Questionnaires were used in five out of the six papers with various purposes. In paper IV, the purpose was to gather information on how useful usability problem lists were for the software developers, but in paper V and VI the survey was used to gather information on the user's subjective opinion about the user interface. In paper VI the standardized questionnaire Attrakdiff 2.0 (Hassenzahl, 2004), was used before the think-aloud sessions to

gather the expectations of how it would be to use the system and the same questionnaire was used to measure the user experience on using the system after those sessions. In papers I and II questionnaires were used to gather information on the evaluator's satisfaction on using the evaluation method.

Using a questionnaire is rather formal approach. The users are given a sheet of paper to fill in; they stay quiet and concentrate on giving the right answers, which is a bit like in an exam. Some of the users' comments pointed in that direction.

Observation in Think-aloud Sessions

I was the evaluator in the think-aloud sessions in the studies described in paper III to VI. The evaluator sat beside the users, handed out user's task one by one and observed their use of the system. During the evaluation session, I encouraged the users to describe what they were doing while solving the tasks. Before and after these evaluation sessions, the conductor handed out questionnaires and made an informal interview with the users at the end of the evaluation. In all these studies this was done in co-operation with real users in their real environment. Another usability specialist was taking notes in all these studies and the sessions were audio recorded.

Sitting beside the users many sessions in a row and observing their use of the same software system solving the same tasks gives invaluable information. Sometimes that contextual information is hard to describe in detail but it gives extensive insight on the work situation and the implications for the users. It can certainly bring new research ideas into play having such a close contact with the actual users.

Two key elements of evaluating with the think-aloud method is choosing participating users and selecting the tasks for the evaluation. In the studies described in paper III to V, the tasks were selected by the authors of the papers in close co-operation with domain experts which knew the users tasks and the software well. The users were recruited by domain experts that selected representative users in all the studies.

Informal Interviews

As recommended in many textbooks (Molich, 2000), (Preece, et. al. 2002) (Stone, et. al., 2005) a debriefing session in the form of an informal interview was made after each think-aloud session in the studies described in paper III to V. There was considerable variation in how much information the users were willing to give in the informal interviews, but the interviews gave the users opportunity to explain their thoughts and ask questions to the

evaluator which added to the understanding of their work situation for the researchers.

Structured Interviews

Structured interviews were used in paper IV to gather information on the usefulness of usability problem reports for the software team. The researchers met with the two developers both at the same time and asked prepared questions and asked them to fill in questionnaires. One of the researchers was the conductor of the meeting while data was gathered by the other researcher by note taking and audio recording.

In retrospect it would have been a more systematic approach to meet the developers one by one for structural interviews and data gathering with the questionnaires. However, conducting the interview like we did gave opportunities for a more open communication, as the interview drifted into being semi-structured in some period.

Results

The results in summary are presented in two parts. In the first part the goal is to answer the first research question: “How different are the results of using various usability evaluation methods for evaluating the same software system?” Results from paper I and II give some answers to this question.

In the second part the goal is to answer the other two research questions, which are: “How useful are the results of usability evaluation to the recipients?” and “What impact does a new software system have on users for achieving their goals?” Three exploratory studies described in paper III to VI give some answers to these questions.

There are also other more detailed results in the papers that are described in the papers only.

Part I: Using Various Evaluation Methods

In paper I, the goal was to study:

- If the results of using either heuristic evaluation or cognitive walkthrough were different from the results of using the think aloud method for evaluating the same system,
- If the results from heuristic evaluation was different from the results of the cognitive walkthrough,
- If combining the results from using two evaluation methods would give better results than using only one method.

In summary, this study shows that a group of 3 evaluators finds 19% in average of the total usability problems by using the heuristic evaluation while group of 3 evaluators using the cognitive walkthrough method find 10% of the problems in average. The mean percentage of usability problems found using the think-aloud method with 3 users was 18%. Here the evaluators were all computer students in their third year; therefore they could all be grouped as novice evaluators. They all evaluated the same software, which was a running prototype of an information retrieval system, which was rather simple. The students chose the users for their user observation and the evaluation environment.

The main contribution in this paper was to study the effect of using one usability method after the other. The results show that if heuristic evaluation is used before the think-aloud method is used, it improves the results of the think aloud evaluation substantially. This approach has not been described in other studies to my best of knowledge.

In paper II the goal was twofold:

- To compare the number and seriousness of problems found per evaluator in heuristic evaluation with two different sets of usability heuristics, Nielsen's (Nielsen, 1993) and Gerhardt-Powals (Gerhardt-Powals, 1996)
- To compare two different ways of reporting usability problems, on paper and with the help of a web tool.

The main conclusion was that more than 60% of the usability problems discovered in think aloud sessions were undetected by the evaluators in the heuristic evaluation. This result does not match with the results from paper I. The reason here could be the evaluator effect, (Hertzum, Jacobsen, 2001). The evaluators doing the user observation were more skilled and much more motivated than the evaluators doing the heuristic evaluation, but in paper I the evaluators had similar background and the same motivation.

The results also show that the validity was almost the same from using a paper and using a tool to report the problems, even though many more problems were reported using the tool than the paper. The findings show also that 60% of the effort has been wasted using paper and about 55% of the effort has been wasted using a tool. The ineffectiveness of our tool in enhancing the validity can be attributed by the cognitive load of switching between the tool and the system being evaluated, hasty data entry resulting in false problems and biased use of certain classification values, because the evaluators did maybe often pick the default value in the drop-down menu for classifying the problems.

In the two studies in part I evaluation methods were studied in experimental settings, asking several evaluators to evaluate the same system and deliver in results. The whole evaluation context was not completely realistic. The evaluators were computer science students that did not have much experience in preparing and conducting the evaluation nor describing the results from the evaluation. The users were also computer science students and there were no constrains on the evaluation environment. To be able to tell how the results would be when experienced evaluators evaluated with real users in the real context of use, three more exploratory studies were made and the results are described in part II.

Part II: The Effect of Usability Evaluation

The results in part II are described in four papers. Two of the papers study how useful the results are for the recipients and thereby give answers to the second research question and the effect on users getting a new software system for achieving their goals answering the third research question is studied in all the four papers.

The Usefulness of the Results for the Recipients

In paper III usability evaluation was done on a multimedia service system in 30 homes in Iceland. There were 2 trials, people from 10 homes participated in the first trial and people from 20 homes in the second. In the first trial logging, questionnaires and the think-aloud method were used to measure the use and usability of the software. In the second trial logging and questionnaires were used. After both the trials the usability specialists that conducted the evaluations made their subjective judgment on the usefulness of the methods. The usability specialists were also members of the development team.

The main conclusion in this paper was that these three evaluation methods complement each other but it is difficult to answer questions about the users' successes and problems using the software, if the think-aloud method is not used.

In paper IV, the think-aloud method was used in the context of 10 users on a second version of a software that was delivered two weeks after the evaluation. The results were classified according to the CUP classification scheme (Hvannberg, Law, 2003) (Vilbergsdóttir, et. al. 2006) and handed over to the developers and their project manager. The goal of the study was to measure how useful the results from using the think-aloud method for evaluating the software were for the developers of the software. The results had been analyzed and described in detail by using seven attributes called the CUP attributes.

The results show that the developers chose to fix only 13% of the registered problems. The reason was that they did not have time to fix more and did not prioritize fixing these problems highly. The developers were asked to respond on what of the seven CUP attributes where useful for them in the three steps of working with the usability problems: when understanding the problem; prioritizing the problem and fixing the problem. Both of the developers involved in the study agreed on that information on the user's task, the context of the problem in the user interface and a text description of the problem was useful in all the three steps of working on solving the problems. Out of the four other attributes the developers agreed that the frequency of the problem was useful for prioritizing the problem. One of the two developers said that the severity of a problem and the analysis of what caused the

problem were also useful when prioritizing the problem, but the other developer did not agree. Furthermore one of the developers found information useful when fixing the problem on the development phase that the problem was expected to have it origins in. For all the other attributes developers did not say that the information was useful.

The Impact on Users

The effect of introducing a new software system to users on how they achieve their goals is described in paper V. Usability evaluations were carried out, first on the old system with 6 users, secondly on the new software shortly after introducing it to users with 8 users and finally on the new one again after six months of use with the same 8 users. All the evaluation sessions were conducted in the users' own environment. This experimental design was selected to be able to measure how much and in what way the users were affected by the new system for solving the tasks they needed to achieve their goals in the real work environment.

The main findings of the study were that the new GUI interface did not in all cases evolve into a more effective system for users to solve their tasks with. Half of the tasks were less effective to solve than in the old character-based system and only one task was clearly more effective to solve. The new system did however benefit users in that sense that they eventually used supplementary systems much less than before and were more satisfied with the new system than the old one.

When looking at how much the users improved their usage between the evaluations done two weeks after installation of the new system and the one done six months later, the results show that the usage did not improve much at all. This was especially clear for solving the tasks that proved difficult for users from the start. Six months later some of the users refused to even try to solve those. For the tasks that were moderately difficult to solve to begin the usage had improved to some extent but much less than expected. Therefore, it seems extremely important that each part of the system is usability tested and redesigned if necessary before handed to users. The users will not adjust easily to usability problems in systems, so tasks that are really difficult to solve on the first days of usage will continue being difficult. These results also indicate that usability evaluation done in connection to the installation of a system gives good indication of how much users will be affected by the new system.

In paper VI a study on measuring the user experience is described. Ten users that participated in a think-aloud evaluation also described in paper IV, answered the questionnaire Attakdiff 2.0 (Hassenzahl, 2003) to measure their expectations to the new system before using it and their user experience after using the system in the think-aloud evaluation. The results show that users had higher expectation in all four attributes measured, the attraction of

the system, the hedonic stimulation, the hedonic identification and the pragmatic manipulation.

Concluding Remarks

In papers I and II the main result was that the think aloud method should be used if the goal of the evaluation is to gather as realistic information as possible on usability problems. In papers III, IV, V and VI the usability evaluation was done by using the think aloud method in co-operation with real users in their real context of use. In papers IV and VI this was done on software that the users had not seen before. In paper V the think aloud sessions were conducted both two weeks after the installation of the software and 6 months after the installation and in paper III the think aloud session were conducted 6 weeks after the installation.

The results from paper V indicate that the results from user observation done on a system that users have not seen before or used only for few days are rather similar to the result from user observation done when users have used the system for longer period. Consequently the common practice of doing user testing on software that users have not seen before and interpret that the results will be the same for actual usage of the system when users will use the system for their real tasks for shorter or longer period is adequate.

Discussion

My motivation for doing this research is twofold:

- I want to guide practitioners on what method to choose for their evaluation.
- I want to study how it impacts users in achieving their goals to get a new software system to use.

In this chapter the discussion will be structured according this motivation.

Finding the Best Method for the Practitioner

Wixon claims (Wixon, 2003) that the literature where usability evaluation methods are compared fails the practitioner. He summarizes that there are shared set of premises in the literature, namely:

1. *Number of problems detected is the most appropriate criterion for evaluating a method.*
2. *Methods can be evaluated in relative isolation from the practical goals of the method and the context in which the method is used.*
3. *A quasi-scientific framework is the most effective approach to resolve disputes about the best method.*

He argues that the three premises above severely limit the usefulness of the literature for the practitioner.

First he states that, problem detection is only the first step in improving a system, and it is not sufficient for product improvement or for method evaluation. There are two studies described in paper I and II in this thesis, where the number of problems found by using different evaluation methods are compared. In both these studies, the assumption was made, that finding a list of all the real problems in the system is possible. If enough think-aloud sessions are conducted, then a list of the true problems in the system can be made. That was the approach in paper II. In paper I there was a slightly different approach taken. The list of all usability problems in the system was made on the basis of the results from the think-aloud session, but the authors added some problems to the list that were found by the using the other methods and the authors found important. When this list of all real usability prob-

lems is there, and we trust that the list is right, calculations can be made on how many real problems the evaluation through expert analyses returns and how many false alarms and misses. Then the results from paper II, indicate that practitioners should use the think-aloud method when ever possible because they would find three times more problems than using heuristic evaluation.

Chattratichart and Lindgaard (Chattratichart, Lindgaard, 2009) state that it is unlikely that a usability test will reveal all problems that exist. If that is true, a researcher that observers hundreds of think-aloud sessions can't be sure that he has the list of all real problems in the system and then counting how many of those problems are found by other methods is not as reliable as it seems. Further research is needed on how reliable usability methods really are. What is significant is probably only the relative effectiveness of the usability evaluation methods within each study.

In the study described in paper VI the user experience was measured to see if that gives more valuable information for the practitioner than counting usability problems. The user expectation was measured before each think-aloud session and the user experience of using the system right after the session. The results show that the user experience was somewhat lower than the expectations in all the four factors measured by the questionnaire Attrakdiff 2.0 (Hassenzahl, 2004). But this result is even harder for the practitioners to use than a list of usability problems.

Secondly, Wixon (Wixon, 2003) claims that:

> *"Isolating a method from the broader context in which it is used renders any purported evaluation to be of little practical significance because it eliminates important consideration such as team buy-in, resources available, relative ease of making a change and numerous other practical considerations."*

His advice is to focus on factors of success, such as how effectively the method introduces usability improvement into the product. Researchers should study whether a method in its very practice encourages participation, buy-in, and collaboration by the development team.

In paper IV the think-aloud method was used in a real user context by usability experts. There the goal was to measure how much and in what way the software developers could use of the results from the think-aloud sessions. The results were disappointing because the evaluation was done very thoroughly and the problems were described in a very detailed way. Still the developers only corrected a small proportion of the problems reported to them. Why? I have had private conversations with one of the developers involved over and over again to try to understand why they did not want to correct more problems and my conclusion is that there was not enough team buy-in in this study. The developers did not ask for this evaluation, it was not included in their development process to use the usability evaluation results, and therefore they claimed they did not have time. My question is

also: Would we find the same results, if the method was used in more realistic setting for the practitioners? My guess and hope is that they would prioritize the problem descriptions higher.

The Real Impact on Users Achieving their Goals

Wixon (Wixon, 2003) suggests that the methods should be studied in their real context, not on simulated systems or hypothetical models. Many researchers share his opinion, for example Cockton and Woolrych (Cockton, Woolrych, 2009) recently describe this need in the final report of working group 2 in the COST-294 action.

In paper V the actual users of two systems were observed in the real work environment using systems that were already installed and these users had been using. The predefined tasks that the users were asked to solve were defined by their project manager at the company and not by the usability experts so that these would be as realistic as possible. So all the planning was carefully done, to make the results as realistic as possible.

In this study we found that two tasks were particularly difficult for the users to solve in the new system. After two weeks of usage, half of the users could not complete one of the tasks and 75% of the users could not complete the other task. Six months later this had not changed much, 63% of the users could not complete each of those tasks. Since I did this study, my question has been: But what did they do then? How would they solve these tasks in real life? when I was not there doing a user observation? Would they have the same problems as they had when I was there? These questions remain open and need further exploration.

Some Reflections on the Research Work

The studies in this thesis were done independently and over a period of nine years. In 1999, when paper I was published, it was common to count usability problems as a measure of effectiveness of evaluation methods and then the formulas to calculate thoroughness, validity and effectiveness, which are used in paper II, did not exist. Now 10 years later measuring the downstream utility of usability evaluation, as was done in paper IV, is regarded as a better way to estimate the effectiveness of usability evaluation. The need for more case studies to understand the complexity of usability evaluation was first expressed by Wixon in 2003 (Wixon, 2003). The case studies described in paper III to VI contribute to that understanding.

The studies in this thesis do not give complete answers to the research questions. Continuing research work would benefit by picking up the identified lessons to provide more thorough answers to the research questions.

Future Work

As pointed out by Sears (Sears, 1997) and described in the background chapter evaluators want reliability, that is consistent usability evaluation results, they want thoroughness, that is they want the results to be complete, and they want validity, that is the results to be correct.

Wixon (Wixon, 2003) suggested studying the actual use of evaluation methods and how these fit in the whole working context of the practitioners. These two approaches will be described further below.

Studying the Reliability, Thoroughness and Validity

The impact of variations of the factors of the evaluation context needs to be studied in more detail, in order to better understand reliability and thoroughness of evaluation methods. Each of the factors, the goal of the evaluation, the processes of the methods, the supporting material, the people involved, the systems evaluated, the data gathered and the evaluation environment could be studied in more detail while the other factors should be kept as realistic as possible. One issue to study is, if the methods deliver different results if used on systems from various domains, for example games or work-related systems. This will actually be one of the main goals of a newly accepted COST action, IC0904 called: “Towards the Integration of Transectorial IT Design and Evaluation”, which I will be a member of.

Another issue to study in more detail is the combination of some of those factors and getting better knowledge on how the factors correlate to one another, for example studying if difference in age of the participants in user observations does affect the results of usability evaluation of systems from different domains.

It is extremely important to study the validity of using evaluation methods, to extend our understanding on if evaluation methods are actually measuring the real problems users have when using the system in real life. The trouble is that it is very hard to find ways to do this.

I agree with Wixon (Wixon, 2003) when he suggests that we do more case studies because, as he states: *“the development of real products is the only context sufficiently rich to produce the kind of nuanced examples that are needed to develop a differentiated and contextualized understanding of methods and techniques needed by practitioners”*.

Researchers could also go back to the users that participated in an evaluation of a particular version of a system before installation and obtain feedback from them after a period of actual use to check they have had the same problems with the system that were registered in an evaluation and not corrected. Another possibility is to watch users or video record their use on a particular system for some period of time and compare those results to results from usability evaluation.

The Actual Use of Evaluation Methods

The third aspect to study is how the usability evaluation methods are actually used today by practitioners to understand strengths and weaknesses of the methods from a practitioner's point of view. This has been done in survey studies in some countries in Europe and USA, for example by Gulliksen and colleagues (Gulliksen, et. al., 2004) in Sweden, by Mao and colleagues (Mao, et. al. 2001) in the USA, and Bystad and colleagues in Norway (Bystad, et. al. 2008). However, this has not been done on a wider scope, for example doing the same survey throughout Europe. Actually, a group of five experts, including myself, that were members of the COST-294 action started working on this idea a year ago, but that task is not finished.

Another research aspect is to study to what extent the methods encourage participation, buy-in and collaboration by the development team, like suggested by Wixon (Wixon, 2003). This could be done by interviewing practitioners or observing their work and their whole work context. An interesting aspect here is how the practitioners can be motivated to use the usability evaluation methods to a greater extent. Would it help to integrate the evaluation methods in the current processes, especially the ones that are really popular today, like the agile processes? How could that be done? I find these questions extremely interesting for understanding the actual use of evaluation methods in real settings and how the practitioners can be encouraged to use usability evaluation methods to a greater extend in the future.

Final Words

As a final remark I would like to relate to the title of this thesis through the advice that developers should listen to their users and use the think-aloud method for evaluating their software. I have experienced in my studies that the evaluators get rich and valuable information by observing users, listening to their comments and experiencing their context of use. To understand how the software would be used in reality, the think-aloud method should be used with real users participating, using their goals to make tasks for the evaluation and conducted in the users real environment.

References

ACM Special Interest Group on Computer-Human Interaction (SIGCHI) Curriculum Development Group 1992. Curricula for Human-Computer Interaction. Technical Report, ACM, New York.

ACM Curriculum Reports, 2008. Computer Science Curriculum 2008: An Interim Revision of CS 2001, ACM, IEEE Computer Society. Accessed on 15/08/09 at http://www.acm.org//education/curricula/.

Blandford, A. 2007. Cognitive Walkthrough. In Proceedings of the COST-294 workshop called R^3UEMs: Review, Report and Refine Usability Evaluation Methods, Athens.

Barkhuus, L., Rode, J. 2007. From Mice to Men - 24 Years of Evaluation in CHI. ACM CHI'07 - Alt.CHI. http://www.viktoria.se/altchi/

Bygstad, B., Ghinea, G., Brevik, E. 2008. Software development methods and usability: Perspectives from a survey in the software industry in Norway. *Interacting with computers*, 375-385.

Carroll, J. M. 2004. Beyond fun. *Interactions* 11, 5 (Sep. 2004), 38-40.

Chattratichart, J., Lindgaard, G. 2009. Is the 'Figure of Merit' Really That Meritorious? T. Gross et al. (Eds.): *INTERACT 2009, Part I, LNCS 5726*, pp. 235–238.

Cockton, G. 2005. I can't get no iteration. *Interfaces*, 63, 4.

Cockton, G.,Woolrych, A. 2009. Comparing Usability Evaluation Methods: Strategies and Implementation – Final report of COST294-MAUSE Working Group 2. In *Maturation of Usability Evaluation methods: Retrospect and Prospect.* Law, E. L., Scapin, D., Cockton, G., Springett, M. Stary, C. and Winckler M. Eds. IRIT Press. Toulouse, France.

Cuomo, D.L., Bowen, C.B. 1994. Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers*, 6, 86-108.

Desurvire, H. W., Kondziela, J. M., Atwood, M. E. 1993. What is gained and lost when using evaluation methods other than empirical testing. In *Proceedings of the Conference on People and Computers VII*. A. Monk, D. Diaper, and M. D. Harrison, Eds. Cambridge University Press, New York, NY, 89-102.

Dix, A., Finlay, J., Abowd, G., Beale, R. 2004. *Human-Computer Interaction*. Third edition, Prentice-Hall, Inc.

Gerhardt-Powals, J. 1996. Cognitive Engineering Principles for Enhancing Human-Computer Performance. In *International Journal of Human-Computer Interaction*, 8 (2) pp. 189-211.

Gray, W.D., Salzman, M.C. 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, 3, 203-261.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., Herulf, L. 2004. Making a difference: a survey of the usability profession in Sweden. In *Proceedings of the Third Nordic Conference on Human-Computer interaction* (Tampere, Finland, October 23 - 27, 2004). NordiCHI '04, vol. 82. ACM, New York, NY, 207-215.

Hassenzahl, M. 2004. The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, 19, 319-349.

Hassenzahl, M., Tractinsky, N. 2006. User experience – a research agenda, *Behavior and Information Technology*, Vol. 25, No. 2, pg. 91 – 97.

Hartson, H. R., Andre, T. S., Williges, R. C. 2001. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction* 13, 4, pg. 373-410.

Hertzum, M., Jacobsen, N.E. 2001. The evaluator effect: A Chilling Fact about Usability Evaluation Methods. In *International Journal of Human Computer Interaction*, 13(4), 421-444.

Hornbæk, K. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies* 64, 2 (Feb. 2006), 79-102.

Hornbæk, K., Frøkjær, E. 2008. Making use of business goals in usability evaluation: an experiment with novice evaluators. In *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 903-912.

Hornbæk, K., Frøkjær, E. 2005. Comparing usability problems and redesign proposals as input to practical systems development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA, April 02 - 07, 2005). CHI '05. ACM, New York, NY, 391-400.

Hvannberg, E. T., Law, E. L. 2003. Classification of Usability Problems (CUP) Scheme, In *Proceedings of the Interact conference 2003*, pg. 655-662.

International Organization for Standardization, 1998: ISO 9241-11 Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability, Geneva, Switzerland.

International Organization for Standardization, 2006. ISO/IEC 25062. Software Engineering - Software product Quality Requirements and Evaluation (SQuaRE)-Common Industry Format (CIF) for Usability Test Reports.

Jeffries, R., Miller, J. R., Wharton, C., Uyeda, K. 1991. User interface evaluation in the real world: a comparison of four techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology* (New Orleans, Louisiana, United States, April 27 - May 02, 1991). S. P. Robertson, G. M. Olson, and J. S. Olson, Eds. CHI '91. ACM, New York, NY, 119-124.

Karat, C., Campbell, R., Fiegel, T. 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, United States, May 03 - 07, 1992). P. Bauersfeld, J. Bennett, and G. Lynch, Eds. CHI '92. ACM, New York, NY, 397-404.

Kjeldskov, J., Skov, M. B., Stage, J. 2004. Instant data analysis: conducting usability evaluations in a day. In *Proceedings of the Third Nordic Conference on Human-Computer interaction* (Tampere, Finland, October 23 - 27, 2004). NordiCHI '04, vol. 82. ACM, New York, NY, 233-240.

Kwahk, J., Han S. H. 2002. A methodology for evaluating the usability of audiovisual consumer electronic products, *Applied ergonomics*, vol. 33, 419-431.

Law, E. L. 2006. Evaluating the Downstream Utility of User Tests and Examining the Developer Effect: A Case Study. *International Journal of Human-Computer Interaction*, 21 (2), 147-172.

Law, E. L., Roto, V., Hassenzahl, M., Vermeeren, A. P., Kort, J. 2009. Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the 27th international Conference on Human Factors in Computing Systems* (Boston, MA, USA, April 04 - 09, 2009). CHI '09. ACM, New York, NY, 719-728.

Lewis, C. 1982. Using the 'thinking-aloud' method in cognitive interface design. IBM Research Report RC 9265(#40713). Yorktown Heights, NY: IBM Thomas J. Watson Research Center.

Lewis, C., Polson, P. G., Wharton, C., Rieman, J. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People* (Seattle, Washington, United States, April 01 - 05, 1990). J. C. Chew and J. Whiteside, Eds. CHI '90. ACM, New York, NY, 235-242.

Lim, Y., Pangam, A., Periyasami, S., Aneja, S. 2006. Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices. In *Proceedings of the 4th Nordic Conference on Human-Computer interaction: Changing Roles* (Oslo, Norway, October 14 - 18, 2006). A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh, and D. Svanæs, Eds. NordiCHI '06, vol. 189. ACM, New York, NY, 291-300.

Liu, Y., 2009. Usability evaluation of medical technology – Investigating the effect of user background and user's expertise, Department of product and production development, division design and human factors, Chalmers university of technology, Gothenburg, Sweden.

Mao, J., Vredenburg, K., Smith, P. W., Carey, T. 2001. User-centered design methods in practice: a survey of the state of the art. In *Proceedings of the 2001 Conference of the Centre For Advanced Studies on Collaborative Research* (Toronto, Ontario, Canada, November 05 - 07, 2001). D. A. Stewart and J. H. Johnson, Eds. IBM Centre for Advanced Studies Conference. IBM Press, 12.

Mayhew, D. J. 1999. The usability engineering lifecycle – a practitioner's handbook for user interface design, Morgan Kaufmann publishers, San Francisco.

Merriam-Webster Online (ND), last accessed 12/08/09 at http://www.merriam-webster.com/dictionary.

Molich, R., Ede, M. R., Kaasgaard K., Karyukin, B. 2004. Comparative usability evaluation, Behavior and information technology, Jan-Feb 2004, vol. 23, no. 1, pg. 65-74.

Molich, R., 2000. Brugervenligt webdesign, Teknisk forlag, Copenhagen.

Nielsen, J. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, United States, May 03 - 07, 1992). P. Bauersfeld, J. Bennett, and G. Lynch, Eds. CHI '92. ACM, New York, NY, 373-380.

Nielsen, J. 1993. Usability Engineering. Academic Press, New York.

Nielsen, J. 2000. Why You Only Need to Test With 5 Users. Alertbox March 19, 2000, at http://www.useit.com/alertbox/20000319.html, accessed on 15/08/09

Nielsen, J. and Mack, R. (eds.) 1994. Usability inspection methods, John Wiley & Sons, Inc., New York.

Nielsen, J., Molich, R. 1990. Heuristic evaluation of user interfaces. In *Proceedings ACM CHI'90 Conference*, ACM, Seattle, WA.

Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J., Stenild, S. 2006. It's worth the hassle!: the added value of evaluating the usability of mobile systems in the field. In *Proceedings of the 4th Nordic Conference on Human-Computer interaction: Changing Roles* (Oslo, Norway, October 14 - 18, 2006). A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh, and D. Svanæs, Eds. NordiCHI '06, vol. 189. ACM, New York, NY, 272-280.

Nielsen, J., Phillips, V. L. 1993. Estimating the relative usability of two interfaces: heuristic, formal, and empirical methods compared. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands, April 24 - 29, 1993). CHI '93. ACM, New York, NY, 214-221.

Polson, P. G., Lewis, C., Rieman, J., Wharton, C. 1992. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies.* 36, 5 (May. 1992), 741-773.

Preece, J., Rogers, Y., Sharp, H. 2002. Interaction design, John Wiley & Sons.

Quesenbery, W. 2003. The Five Dimensions of Usability. In Albers, M. J., and Mazur, B. (Eds.), *Content and Complexity: Information Design in Technical Communication.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sears, A. 1997. Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction*, 9(3), pg. 213-234.

Stone, D., Jarrett, C., Woodroffe, M., Minocha, S. 2005. User Interface Design and Evaluation, Morgan Kaufmann, San Francisco.

Vilbergsdóttir, S. G., Hvannberg, E. T., Law, E. L. 2006. Classification of usability problems (CUP) scheme: augmentation and exploitation. In *Proceedings of the 4th Nordic Conference on Human-Computer interaction: Changing Roles* (Oslo, Norway, October 14 - 18, 2006). A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh, and D. Svanæs, Eds. NordiCHI '06, vol. 189. ACM, New York, NY, 281-290.

Wixon, D. 2003. Evaluating usability methods: why the current literature fails the practitioner. *Interactions* 10, 4 (Jul. 2003), 28-34.

Wharton, C., Bradford, J., Jeffries, R., Franzke, M. 1992. Applying cognitive walkthroughs to more complex user interfaces: experiences, issues, and recommendations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, United States, May 03 - 07, 1992). P. Bauersfeld, J. Bennett, and G. Lynch, Eds. CHI '92. ACM, New York, NY, 381-388.

Wharton, C., Rieman, J., Lewis, C., Polson, P. 1994. The cognitive walkthrough method: a practitioner's guide. In *Usability inspection Methods*, J. Nielsen and R. L. Mack, Eds. John Wiley & Sons, New York, NY, 105-140.

Whitefield, A., Wilson, F., Dowell, J. 1991. A framework for human factors evaluation. Behavior & Information Technology, 10 (1), 65-79.

Woolrych, A., Cockton, G. 2001. Why and When Five Test Users aren't Enough. In *Proceedings of IHM-HCI 2001 Conference*: Volume 2, eds. J. Vanderdonckt, A. Blandford, and A. Derycke, Cépadèus Éditions: Toulouse, pg. 105-108.

Recent licentiate theses from the Department of Information Technology

2009-002 Elina Eriksson: *Making Sense of Usability: Organizational Change and Sensemaking when Introducing User-Centred Systems Design in Public Authorities*

2009-001 Joakim Eriksson: *Detailed Simulation of Heterogeneous Wireless Sensor Networks*

2008-003 Andreas Hellander: *Numerical Simulation of Well Stirred Biochemical Reaction Networks Governed by the Master Equation*

2008-002 Ioana Rodhe: *Query Authentication and Data Confidentiality in Wireless Sensor Networks*

2008-001 Mattias Wiggberg: *Unwinding Processes in Computer Science Student Projects*

2007-006 Björn Halvarsson: *Interaction Analysis and Control of Bioreactors for Nitrogen Removal*

2007-005 Mahen Jayawardena: *Parallel Algorithms and Implementations for Genetic Analysis of Quantitative Traits*

2007-004 Olof Rensfelt: *Tools and Methods for Evaluation of Overlay Networks*

2007-003 Thabotharan Kathiravelu: *Towards Content Distribution in Opportunistic Networks*

2007-002 Jonas Boustedt: *Students Working with a Large Software System: Experiences and Understandings*

2007-001 Manivasakan Sabesan: *Querying Mediated Web Services*

2006-012 Stefan Blomkvist: *User-Centred Design and Agile Development of IT Systems*

2006-011 Åsa Cajander: *Values and Perspectives Affecting IT Systems Development and Usability Work*

2006-010 Henrik Johansson: *Performance Characterization and Evaluation of Parallel PDE Solvers*

2006-009 Eddie Wadbro: *Topology Optimization for Acoustic Wave Propagation Problems*

UPPSALA
UNIVERSITET

Department of Information Technology, Uppsala University, Sweden