Department of Computer Science

Experiences with usability testing

Effects of thinking aloud and moderator presence

Sirpa Riihiaho

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 75/2015

Experiences with usability testing:

Effects of thinking aloud and moderator presence

Sirpa Riihiaho

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 26th June 2015 at 12 noon.

Aalto University
School of Science
Department of Computer Science
Strategic Usability Research Group

Supervising professor
Marko Nieminen

Preliminary examiners
Sharon McDonald, University of Sunderland, United Kingdom
Jan Stage, Aalborg University, Denmark

Opponent
Judith A. Ramey, University of Washington, United States

Aalto University publication series
DOCTORAL DISSERTATIONS 75/2015

ISBN 978-952-60-6226-6 (printed)
ISBN 978-952-60-6227-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6227-3

Unigrafia Oy
Helsinki 2015

Finland

Publication orders (printed book):
Sirpa.Riihiaho@aalto.fi

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Sirpa Riihiaho

Name of the doctoral dissertation
Experiences with usability testing: Effects of thinking aloud and moderator presence

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 75/2015

Field of research Computer Science and Engineering

Manuscript submitted 15th April 2014 **Date of the defence** 26th June 2015

Permission to publish granted (date) 8th April 2015 **Language** English

☒ **Monograph** ☐ **Article dissertation (summary + original articles)**

Abstract

Usability testing has become a standard method when evaluating the usability of various systems with real users. Despite this, the factors of usability testing have been given little attention in the academic forums. For example, the use of the thinking aloud method has resulted in quite conflicting effects on the users' performance in the context of usability testing, and the effects of the moderator presence have been studied very seldom in usability research.

This thesis studies methods of usability testing and contextual factors in the test settings that may affect the results of a test, focusing on the effects of relaxed thinking aloud and the presence of a test moderator. It combines an extensive literature review with experiences on usability testing from 22 years covering 143 usability studies. The challenges of usability testing reported in the top academic HCI forums focus on sampling, context of use, use over time, and assessment of utility and value. The methods presented in this thesis provide some solutions to problems related to the context of use and to the assessment of utility and value.

The most significant contribution of this thesis is in the experiment with thinking aloud and moderator presence as its independent variables. The results of relaxed thinking aloud show that it has no significant effect compared to silent performance on the number of usability problems the users face or in their subjective ratings, but it does slow down their performance. A significant effect of the moderator presence is found in the users' subjective rating, as users with a moderator next to them rate the system preferences significantly higher than participants performing alone. Given the benefits of having a moderator next to the user who is able to ask clarifying questions when the experiences are fresh in user's mind, the thesis still recommends this approach in formative usability testing when it is important to come up with practical redesign proposals for the development team.

Thinking aloud in this experiment does not enhance the evaluators' confidence in the detected usability problems and their causes, and the test users mostly report it as an unnatural extra effort. Even so, thinking aloud gives more information about the problems to customers observing the tests, and thereby may motivate designers to make the required changes to the system. Therefore, if performance measurements are required, silent performance or classic concurrent thinking aloud with minimal interventions should be used, but in formative testing, the more explanatory relaxed thinking aloud can be used, as long as its potential effects on users' performance are kept in mind.

Keywords usability testing, thinking aloud, moderator presence, usability evaluation methods

ISBN (printed) 978-952-60-6226-6 **ISBN (pdf)** 978-952-60-6227-3

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Espoo **Year** 2015

Pages 200 **urn** http://urn.fi/URN:ISBN:978-952-60-6227-3

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Sirpa Riihiaho

Väitöskirjan nimi
Kokemuksia käytettävyystestauksesta: Ääneenajattelun ja ohjaajan läsnäolon vaikutuksia

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 75/2015

Tutkimusala Tietotekniikka

Käsikirjoituksen pvm 15th April 2014 **Väitöspäivä** 26th June 2015

Julkaisuluvan myöntämispäivä 8th April 2015 **Kieli** Englanti

☒ **Monografia** ☐ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Käytettävyystestauksesta on tullut vakiintunut käytäntö erilaisten järjestelmien käytettävyyden arviointiin yhdessä käyttäjien kanssa. Tästä huolimatta käytettävyystestauksen osatekijät ovat jääneet hyvin pienelle huomiolle akateemisessa keskustelussa. Ääneenajattelun vaikutuksista on tutkimuksissa saatu hyvin ristiriitaisia tuloksia käytettävyystestauksessa, ja ohjaajan läsnäolon vaikutuksia on tutkittu hyvin vähän käytettävyysalalla.

Tässä opinnäytteessä tutkitaan käytettävyystestauksen menetelmiä ja osatekijöitä, jotka voivat vaikuttaa testauksen tuloksiin, keskittyen rentoutuneen ääneenajattelun ja ohjaajan läsnäolon vaikutuksiin. Työssä yhdistyy laaja kirjallisuuskatsaus ja 22 vuoden kokemukset 143 käytettävyysarviosta. Alan huippujulkaisuissa raportoidut käytettävyystestauksen haasteet keskittyvät testiotoksiin, käyttötilanteeseen, pidempiaikaiseen käyttöön ja hyödyn ja arvon arviointiin. Opinnäytteessä esitellyt menetelmät tarjoavat ratkaisuja näihin ongelmiin käyttötilanteen ja hyödyn ja arvon arvioinnin osalta.

Tämän opinnäytteen tärkein anti on kokeellisessa osiossa, jossa riippumattomina muuttujina ovat ääneenajattelu ja testiohjaajan läsnäolo. Rentoutuneella ääneenajattelulla ei hiljaiseen toimintaan verrattuna ole merkittävää vaikutusta käyttäjien kohtaamien käytettävyysongelmien määrään tai heidän miellyttävyysarvioihinsa, mutta se hidastaa heidän toimintaansa. Testiohjaajan läsnäololla sen sijaan on merkittävä vaikutus, sillä ohjaajan läsnäollessa toimineet käyttäjät arvioivat järjestelmän miellyttävämmäksi kuin yksin olleet. Vaikka ohjaajan läsnäoloon liittyy riski saada positiivisempia arvioita, opinnäyte suosittelee ohjaajan läsnäoloa etenkin tuotekehityksen aikaisessa testauksessa, jossa etsitään keinoja parantaa tuotteen käytettävyyttä. Kun ohjaaja on käyttäjän lähellä, hänellä on erinomainen mahdollisuus kysyä tarkentavia kysymyksiä kunkin tehtävän jälkeen kokemusten ollessa vielä käyttäjän tuoreessa muistissa.

Ääneenajattelu ei lisää tässä koeasetelmassa arvioijien varmuutta löytämistään ongelmista tai niiden syistä, ja käyttäjät kokevat ääneenajattelun lähinnä luonnottomana lisäponnisteena. Ääneenajattelu kuitenkin antaa testejä seuraavalle asiakkaalle lisätietoa ongelmista ja voi täten motivoida tekemään vaaditut muutokset. Niinpä on suositeltavaa käyttää perinteistä ääneenajattelua mahdollisimman pienin keskeytyksin suoritusmittauksia sisältävässä testauksessa, mutta kehityksen aikaisessa testauksessa voidaan käyttää selittävämpää rentoutunutta ääneenajattelua, kunhan sen mahdolliset vaikutukset käyttäjien toimintaan tiedostetaan.

Avainsanat käytettävyystestaus, ääneenajattelu, ohjaajan läsnäolo, käytettävyyden arviointimenetelmät

ISBN (painettu) 978-952-60-6226-6 **ISBN (pdf)** 978-952-60-6227-3

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2015

Sivumäärä 200 **urn** http://urn.fi/URN:ISBN:978-952-60-6227-3

Preface

Gathering the evaluation data and writing this thesis has taken a long time, so there have been several projects and funders that have affected my research, and many persons that have been involved in the process during these years. In the early phases of my research, Leo and Regina Wainstein's Foundation and Tekniikan edistämissäätiö supported my work financially with grants. Also the Academy of Finland, the National Technology Agency (TEKES) and the Finnish Work Environment Fund (Työsuojelurahasto) have funded my research. The experimental part of this thesis, on its part, was conducted in the Mobile Financial Services (MoFS) project funded by TiViT/Digile.

Many colleagues and students have participated in the evaluations that have affected this thesis, and several persons have helped in finding the time for my research and writing. Friends and teammates in Niitti have also been very important for me in finding some balance between my work and hobbies. I am grateful to all of you. I want to especially thank my supervising professor Marko Nieminen for his support, all the active students trying out my ideas in their course assignments, and all my current and former colleagues for interesting discussions on life, the Universe and everything – including usability.

Finally, my warmest thanks go to my family. My mother and my already deceased father have always supported me in my studies and life, and also my sister has always helped me in numerous ways. But most of all, my devoted husband Jukka and lovely daughter Sanni have been the central part of my life: thank you for sharing the joy and misery of life with me!

Sirpa Riihiaho
Vantaa, May 1st 2015

Contents

1 Introduction

Usability is a multifaceted term. In user-centred product development, it usually means the quality of use in a context (Bevan & Macleod 1994), or "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" (ISO 9241-11, 1998). To integrate usability concerns into the real product development processes, a practical usability engineering approach was launched in the 1980's (K.A. Butler 1996). This approach includes measurable usability specifications and objectives; a wide set of usability techniques and tools; and considerations for cost efficiency (Carroll 1997). The costs need to be justified to obtain approval and support from the management, so the ISO standard 9241-210 (2010) lists both the economic and social benefits of human-centred design[1], such as: increment of users' productivity, reduction of training and support costs, and a competitive advantage.

Usability evaluation is an "*essential step in human-centred design and should take place at all stages in system life cycle*" (ISO 13407, 1999). Usability evaluation, and usability testing in particular, have been widely considered and applied as quantitative methods concentrating on effectiveness and efficiency in the work domain (Dumas & Redish 1993, Bevan & Macleod 1994, Rubin & Chisnell 2008), leaving little room for user experience and emotions. For example, the ISO 9241-11 standard (1998) names effectiveness, efficiency, and user satisfaction as the attributes of usability, but user satisfaction here means only "*freedom of discomfort*", *i.e.*, absence of negative feelings and frustration. Even the methods for usability evaluation are often assessed by their effectiveness and efficiency in finding usability problems (*e.g.* Jeffries *et al.* 1991), and considered as recipes that can be followed step by step to end up with similar results every time they are used (Woolrych *et al.* 2011). Indeed, an article by Kathleen Potosnak in IEEE Software in 1988 is titled: "*Recipe for a usability test*". However, the most important goal that usability evaluation has in user-centred product development is to reveal the most disturbing usability problems, to find ways to improve the product and, thereby, to help the development team to fulfil the user requirements (Sawyer, Flanders & Wixon 1996, Wixon 2003).

Usability evaluation can be divided into *formative and summative* evaluation according to the goals of the evaluation. Formative evaluation "*involves*

[1] The terms human-centred design, user-centred design and usability engineering are often used as synonyms. The terms in this thesis are selected according to the terms used in the referred articles.

monitoring the process and products of system development and gathering user feedback for use in refinement and further system development", whereas summative evaluation "*involves assessing the impact, usability and effectiveness of the system*" (Hewett 1986). These evaluations also differ in emphasis between quantitative and qualitative results, as qualitative information is relevant in formative evaluation, whereas quantitative results are more likely required in summative evaluation (Booth 1989, p. 119).

The Technical report ISO/TR 16982 (2002) "Ergonomics of human-system interaction – Usability methods supporting human-centred design" lists some methods that can be used both for design and evaluation, and divides them into methods with direct or indirect user involvement. Similarly, John Karat (1997) divides these methods into empirical user testing methods involving users directly, and usability inspection methods relying on other sources of information about users. The choice of evaluation methods depends on several factors, such as: life-cycle steps, user characteristics, characteristics of tasks to be performed, evaluated product, project constraints, and degree of expertise available (ISO/TR 16982, 2002). Whatever the method is, there are goals for the evaluation, attributes to be evaluated, and a process through which these attributes are assessed (J. Karat 1997). These measurable usability attributes fall into two categories: objective performance measures and subjective user preference measures (J. Nielsen & Levy 1994, Bailey 1996).

The term usability testing is sometimes used as a synonym for user testing and usability evaluation. To make a distinction between a method and a group of methods, this thesis uses the terms as follows:

- *User testing* covers a group of usability evaluation methods that involve user participation, and
- *Usability testing* is a user testing method in which one or more representative users at a time do tasks or describe their intentions under observation.

This definition of usability testing is very close to the definition by Barnum (2011, p. 13): *"the activity that focuses on observing users working with a product, performing tasks that are real and meaningful to them*". Dumas and Redish, on the other hand, name five specific requirements for usability testing: a goal to improve the product; representative users; real tasks; observation and recording; and analysing data and making suggestions for improvements. They also point out that the primary goal of usability testing is to improve the usability of the evaluated product, so the evaluators need to analyse the data, diagnose the usability problems, and recommend changes to fix the problems. (Dumas & Redish 1993, p. 22) Since usability is defined as quality of use in certain context, it is also important to have representative conditions for the evaluation either by having a real use environment or by simulating the essential aspects of the actual use context (Bevan & Macleod 1994).

1.1 Nature and scope of this thesis

This thesis studies the methods and process of usability testing, and the effects of various contextual factors in the test settings that can influence the results. These factors include test users' expertise and expectations; test environment; moderator presence; prototype level and aesthetics; and the use of the thinking aloud method. Juergen Sauer and Andreas Sonderegger have done similar studies using a four-factor framework of contextual fidelity (Sauer, Seibel & Rüttinger 2010) as a model for their studies. This framework builds on four main factors: user characteristics, task scenarios, system prototype and testing environment (Figure 1).

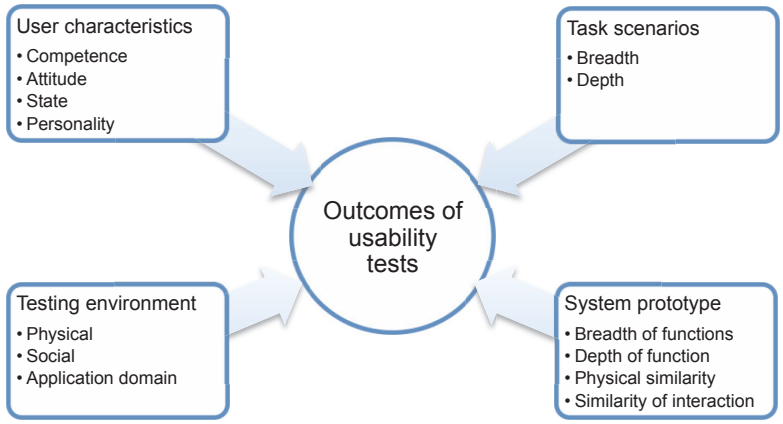


Figure 1: A four-factor framework of contextual fidelity by Sauer *et al.* (2010).

As Sauer *et al.* (2010) point out: "*The attribute 'contextual' emphasises the wider context and the different aspects of fidelity that are to be considered in usability testing*". Therefore, the term "contextual factors" is used also in this thesis to denominate the factors in usability test setting potentially affecting the results of the test. Compared to the factors addressed in this thesis, the four-factor framework leaves out the characteristics of the selected evaluation methods, such as the use of the thinking aloud method.

The empirical basis of this thesis is built on 143 usability evaluation studies conducted at the Helsinki University of Technology[2] and Aalto University in 1993-2014, including an experimental study on usability testing involving 38 test users. The thesis also includes a literature review of usability testing and on studies of several contextual factors in the testing. The thesis focuses on the phases of designing and conducting usability tests, so the phases of analysis, reporting and communicating the results are given less attention. Methods requiring special equipment, such as eye-tracking systems, are left out, as is remote usability testing in which the interaction with the users is usually through digital media or even asynchronous. As this thesis focuses on usability

[2] The predecessor of Aalto University

testing methods instead of user testing methods in general, also self-reporting methods and methods relying on logging data, are given little attention.

During these 22 years, usability testing has become an established procedure in usability evaluation. The focus on product development has broadened from functionality to the overall experience with the product, and new approaches have emerged, such as: experience-based design (Bødker 2006); customer experience and service science (J.R. Lewis 2014); worth-centred design (Cockton 2006; Gilmore *et al.* 2008); and life-centred design (Holtzblatt 2011). Even so, the conventions of usability testing have remained very much the same as in the textbooks from the early 1990's (*e.g.* J. Nielsen 1993; Dumas & Redish 1993; Rubin 1994).

Thinking aloud is widely used in usability tests world-wide (*e.g.* Dumas & Redish 1993; Gulliksen *et al.* 2004; Dumas & Loring 2008; Rubin & Chisnell 2008; McDonald, Edwards & Zhao 2012). Still, its validity as a research method in usability tests was rarely considered or disputed (Boren & Ramey 2000; Hertzum, Hansen & Andersen 2009) before studies of thinking aloud began to grow in number in 2010 (*e.g.* Cooke 2010; Olmsted-Hawala *et al.* 2010a & 2010b; Zhao & McDonald 2010; Elling, Lentz & de Jong 2012; McDonald & Petrie 2013; Hertzum & Holmegaard 2013; McDonald, McGarry & Willis 2013a; Zhao, McDonald & Edwards 2014). Similarly, a test moderator often sits next to the user in usability tests, but there has been very few studies of the effects of this setting. As these approaches are also used in most of our evaluations, they are in focus in this thesis when studying the effects of contextual factors.

The practical aim of this research is to present a set of flexible usability testing methods applicable in various situations. The methods are not to be considered as step-by-step instructions or as recipes, but more as approaches that can be modified according to the requirements of the evaluation. To be able to plan a good usability test, and especially to make solid interpretations of the results, one needs to understand how various contextual factors of usability testing can affect the results. Therefore, as an academic goal, this thesis goes one level below focusing on the usability evaluation methods as single units, and focuses on the factors of usability testing and its settings, as for example Woolrych *et al.* (2011) suggest.

Most of the usability testing methods presented in this thesis have been presented in my previous papers, such as the modified pluralistic usability walkthrough method (Riihiaho 2002) and the contextual and informal walkthrough methods (Riihiaho 2009). As the references indicate, I have done the analyses of the methods and their reporting mostly on my own. The article on the visual walkthrough and heat maps (Juurmaa, Pitkänen, Riihiaho, Kantola & Mäkelä 2013), on the other hand, is a good example of a publication made together with students on the basis of their course assignments.

1.2 Research problem and questions

According to Ellis and Levy (2008), research must be based on "*an exhaustive understanding of the body of knowledge related to the field or topic of study*" to make an original contribution, and either fill a known gap in the body of knowledge, replicate and expand previous research, or bring out new solutions for some "*specific, identifiable, and documented problems with the currently available solutions*". To strive for these goals, this thesis builds on the knowledge and research on usability testing in the top academic human-computer interaction (HCI) forums, outlining the reported problems with usability testing, and reflecting the possibilities of various usability testing methods and approaches to fix these problems. As another view, this thesis aims to identify contextual factors of usability testing and their potential effects on the results of usability testing, focusing on the moderator presence and on the use of the concurrent thinking aloud method.

The research problem in this thesis is:

Are the established practices of usability testing appropriate and valid, that is they do not change the phenomena they are studying?

This broad problem area is approached through the following five research questions:

1. Which of the top academic HCI forums are active in discussing usability testing methodology?
2. What problems and challenges have been reported on present usability testing methods in these forums?
3. What methods and modifications of usability testing are reported?
4. What contextual factors of usability testing have been identified, studied and reported?
5. What effect does concurrent relaxed thinking aloud method and the presence of a test moderator have in a usability test setting on users' performance and preferences, as well as on the evaluators' analyses?

1.3 Field of studies

This thesis falls into the fields of usability research, user-centred product development, usability engineering and human-computer interaction. For example, Patricia Sullivan (1989) identifies four groups in usability research: sociology; marketing; human-computer interaction ranging from cognitive psychologist to ergonomics engineers; and technical communication including writers, document designers and educators. These groups have their own traditions and methods for research that are brought into usability research contributing a broad variety of methods. Sullivan classifies these methods according to their goals into three philosophical models, namely: the product development model, the cognitive model and the cultural model. The product development model is a pragmatic, engineering model where information is

passed to product development in its various phases. The cognitive model, on its part, aims to model the learning and use of products, and the cultural model aims to describe the social use of products in their real context of use. (Sullivan 1989) When using these classifications, the research in this thesis looks at human-computer interaction, and it adheres to the product development model aiming for pragmatic usability testing methods in human-centred design.

Human-computer interaction is a multidisciplinary field of studies that "*arose as a field from intertwined roots in computer graphics, operating systems, human factors, ergonomics, industrial engineering, cognitive psychology, and the systems part of computer science*" (Hewett *et al.* 1992). Mackay and Fayard (1997) put HCI between natural sciences and engineering along with computer science and human factors, as they all borrow paradigms, techniques and tools from natural sciences and engineering. Similarly to HCI, this thesis borrows from several disciplines, including human factors and ergonomics, cognitive psychology, software engineering, and technical communication. My background from computer science and engineering, and cognitive psychology can be seen in this thesis in the emphasis on research methods and on practical implications, as these are the phases that Lazar, Feng and Hochheiser (2010, p. 13) claim to be in focus in the HCI research process in the disciplines of psychology and computer science (Figure 2).

Figure 2: Research process in HCI and the focus of disciplines in its phases (based on Lazar *et al.* 2010).

2 Research methods and material

This thesis is based on a literature review, empirical usability studies, and an experiment on usability test settings. Table 1 summarises the research questions and the methods chosen to find answers to these questions. This chapter presents the principles of the literature review and the top HCI forums that were the main focus in searching for material. After that, the background of my empirical usability studies is presented along with an overview to the empirical data that gives the basis for my experiences of usability evaluation methods. Finally, the chapter presents the principles of experimental research and gives some examples how these principles have been applied in this thesis.

Table 1: Research questions and methods used to find answers to these questions.

No.	Research question briefly	Research methods
1	Active HCI forums on usability testing methodology?	Literature review
2	Reported problems with present usability testing?	Literature review
3	Methods for usability testing?	Literature review, empirical studies
4	Contextual factors of usability testing?	Literature review, empirical studies
5	Effect of thinking aloud and moderator presence?	Literature review, empirical studies and experiment

2.1 Literature review

An effective literature review is needed to lay grounds for research, *i.e.*, it must motivate the research, connect it to the knowledge already known of the topic, and "*provide justifications for the potential contributions provided by the proposed study*". To be able to do so, it must rely on high quality literature. Therefore, "*literature from leading, peer-reviewed journals should serve as the major base of literature review*" leaving practitioners oriented and not peer-reviewed work only to little attention. (Levy & Ellis 2006)

According to Levy and Ellis (2006), the quality of the sources can be justified by using lists of highly ranked journals and conferences. The depth and broadness of the review, on the other hand, is ensured by using multiple literature databases, and using effective search techniques, such as keyword search, and backward and forward searches. Complying with the recommendations by Levy and Ellis, the related research in this thesis is searched from the high quality journals and conferences on human-computer interaction. Journals are often more appreciated and more highly ranked than conferences, so journals are given more room in this thesis presenting them one by one, whereas

conferences are presented only as a summary. Before going into these specific forums, however, a brief overview of the use of the terms "usability" and "usability testing" in the academic forums is presented.

2.1.1 Usability and usability testing in ACM publications

The ACM Digital Library (ACM DL), published by the Association for Computing Machinery, is highly respected amongst HCI researchers, and has a good coverage on the high quality publications in HCI. Thereby, it served as a source for studying the evolution of usability related publications in this thesis. The term "usability", as such, started to appear in ACM computing articles already in the early 1970's. At first, the term was often used as a synonym for availability, suitability or usefulness, but also the meaning of easy-to-use was attached to the term. For example, in Culpepper's (1975) conference paper "*A system for reliable engineering software*", usability is determined as a requirement for software and its documentation: "*usability – the software must be adequately documented so that it can be easily used and maintained*". Nevertheless, the usability profession is generally considered to have started in the late 1980's, and the works of Whiteside and Bennett at Digital Equipment Corporation and IBM as its one starting point (Dumas 2007).

The use of the term "usability" has increased from the very few in 1970's to nearly 40 000 in the 21st century. Table 2 presents the numbers in more detail in 5 year periods starting from the year 1970. As the amount of material in ACM Digital Library has about doubled every 5 years, the amount of usability material has even tripled, making the proportion of usability material 2,19% in 2010 instead of the 0,03% in 1975.

Table 2: The amount of publications including the word "usability" in ACM DL and its proportion of the whole material in the years 1970-2010[3].

Published before year	Word "usability" used in publication	Amount of ACM publications	Proportion of usability publications (%)
1970	0	15 289	0
1975	11	33 076	0,03
1980	49	57 117	0,09
1985	172	100 043	0,17
1990	483	242 830	0,20
1995	1 336	448 440	0,30
2000	4 185	737 789	0,57
2005	14 585	1 101 604	1,32
2010	39 732	1 812 903	2,19

The first ACM Digital Library entry using the term "usability testing" is in Richard Spencer's book "*Computer Usability Testing & Evaluation*" from the year 1985. A conference paper by Lovie Melkus (1985) on laboratory testing is also available from that year, but it seems to be categorised to the following

[3] Results from ACM DL advanced search on October 7th 2012 for searches: "*Find "usability"*" and "*Find ""*" in any field from publications that have been published before the mentioned year.

year in the ACM DL, as the conference was on December. Table 3 shows the results of the searches the way ACM DL reports them.

Table 3: The amount of publications including the term "usability test" in ACM DL, and proportion of the usability material from the years 1980-2010[4].

Published before year	Term "usability test" used in publication	Amount of "usability" publications	Proportion from usability publications (%)
1980	0	49	0
1985	1	172	0,6
1990	12	483	2,5
1995	22	1 336	1,6
2000	85	4 185	2,0
2005	385	14 585	2,6
2010	1 125	39 732	2,8

Although usability testing is not a common subject in ACM publications as a whole, in HCI related conferences, it is one of the most general subjects in usability. For example, the analysis by Kaikkonen (2009) shows that almost 40% of the full papers in the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI) and the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI) in 1998-2008 refer to usability testing when using the term usability. Barkhuus and Rode (2007) focused on the CHI conference papers, and analysed the type and scope of evaluations in these papers. They divided the papers into analytical *vs.* empirical; qualitative *vs.* quantitative; and papers that focus on the evaluation methods themselves were classified as their own category. The number of these methodological papers has diminished over the years, and in 2006, there were none. The papers including evaluations of new techniques or designs, mostly use empirical and quantitative methods, sometimes backed up with qualitative interviews or surveys. (Barkhuus & Rode 2007)

In their analysis, Barkhuus and Rode (2007) also noticed that currently, almost all accepted papers include some type of usability evaluation. As one reason for this, Barkhuus and Rode propose the CHI review criteria of validating the presented results. Greenberg and Buxton (2008) share this view, and also imply that several researchers evaluate their new techniques in favourable situations making "*existence proofs instead of risky hypothesis testing*", and do testing more "*by rule*" than "*by thought*" to get their papers published.

2.1.2 Top HCI forums

The rankings of academic forums are mostly based on the number of citations to the published articles. Citation analysis has been used since 1873 starting from legal profession. Although it has been criticised to ignore the type of citation and to favour well-known writers and review articles, citation analysis is claimed "*to be the best objective measure available for assessing the impact of journal articles, institutions, and individuals".* (Brown & Gardner 1985) In a citation analysis by Katerattanakul, Han and Hong (2003), 27 academic in-

[4] Results from ACM DL advanced search on October 7th 2012 for searches: "*Find "usability test""* and "*Find "usability""* in any field from publications that have been published before the mentioned year.

formation systems journals were ranked: 2 of the journals were HCI related, namely Human-Computer Interaction and International Journal of Human-Computer Studies, and by various indices, they ranked as high as 7th and 11th.

Surveys are another tool to rank journals, but they have been criticised for their inconsistency and lack of comprehensiveness (Katerattanakul *et al.* 2003). In a large survey by Lowry *et al.* (2004) information systems academics were asked to name their top-four research journals without predefined alternatives. The response rate was 32% corresponding to 2559 responses, but the category of HCI research got only 21 responses. Even so, Human-Computer Interaction was the top HCI journal also in this study (Table 4).

Table 4: Top HCI journals for information systems researchers (Lowry *et al.* 2004). The top choices were given more weight in the results.

Rank	Journal	N (21)	Weight
1.	Human-Computer Interaction	6	10
2.	ACM Transactions on Computer-Human Interaction	3	5
3.	Computer Supported Cooperative Work	3	5
4.	International Journal of Human-Computer Studies	2	3
5.	International Journal of Human-Computer Interaction	1	1
6.	Journal of Computer-Mediated Communication	1	1
	Other journals	5	n/a

Several similar ranking lists are available, such as the lists moderated by the Computing Research and Education Association of Australasia[5] (CORE 2010a & 2010b), and the Publication Forum by the Federation of Finnish Learned Societes[6] (Julkaisufoorumi 2012). CORE has separate ranking lists for journals and conferences, using ranks A, B and C, and also a rank A* for the very best journals. The Finnish Publication Forum, on the other hand, ranks the forums into three numerical levels: 1 (basic), 2 (leading), and 3 (top level).

The lists by the CORE and the Finnish Publication Forum have been selected as criteria for quality forums in this thesis, since they have been used in our university in assessing the quality of publications. Together, they represent both international and local Finnish viewpoints to the academic forums. Table 5 shows the HCI related journals that have the rank A* or A in the CORE list, or level 2 or 3 in the Finnish Publication Forum.

Table 5: Top HCI journals according to the CORE list and the Finnish Publication Forum.

Journal	Rank in CORE (A*/A/B/C)	Rank in Finnish Publication Forum (3/2/1)
ACM Transactions on Computer-Human Interaction	A*	3
Human-Computer Interaction	A	3
Behaviour & Information Technology	A	2
International Journal of Human-Computer Studies	A	2
International Journal of Human-Computer Interaction	B	2

The corresponding ranking list of HCI related conferences is presented in Table 6. Since MobileHCI and NordiCHI conferences are highly respected amongst Nordic HCI researchers, also these conferences are included in the table. The table shows that the CHI conference is one of the most valued con-

[5] http://core.edu.au, October 6th 2011

[6] http://www.tsv.fi/julkaisufoorumi/, December 28th 2012

ferences in the HCI field. Although highly ranked conferences, ACM Conference on Computer Supported Cooperative Work (CSCW), International Conference on Intelligent User Interfaces (IUI) and ACM Symposium on User Interface Software and Technology (UIST) are left to little attention in this thesis due to their distinct focus on cooperative work and user interfaces, without much considerations on usability evaluation methods.

Table 6: Top HCI conferences in the CORE conference ranking list, and their rankings both in the CORE list (A/B/C) and in the Finnish Publication Forum (3/2/1).

Acronym	Conference title	Rank in CORE	Rank in Finnish Publication Forum
CHI	ACM SIGCHI Conference on Human Factors in Computing Systems	A	2
CSCW	ACM Conference on Computer Supported Cooperative Work	A	2
HCI	British Computer Society Conference on Human-Computer Interaction	A	1
Interact	IFIP TC13 Conference on Human-Computer Interaction	A	1
IUI	International Conference on Intelligent User Interfaces	A	1
UIST	ACM Symposium on User Interface Software and Technology	A	1
MobileHCI	International Conference on Human-Computer Interaction with Mobile Devices and Services	B	1
NordiCHI	Nordic Conference on Human-Computer Interaction	C	1

2.1.3 Usability in top HCI journals

The top HCI journals according to the CORE ranking lists and the Finnish Publication Forum were studied one by one to study their activity in usability testing methods. To get an overview, the first search for each journal was simply to find the term “usability” in the abstracts. After this overview, the literature searches were continued with backward and forward searches, author searches and searches with various keywords, such as “usability evaluation” and “evaluation methodology”. Several databases from literature vendors were used, such as: ACM Digital Library[7], IEEE Xplore Digital Library[8], ProQuest ABI/INFORM Complete[9], Elsevier ScienceDirect[10], Taylor & Francis Online[11], SAGE journals[12] and EBSCOhost[13]. The literature searches in top HCI forums for this thesis were started in fall 2011, and the last searches to update the number of hits and to check for new papers were made on February 2nd 2014.

ACM Transactions on Computer-Human Interaction (TOCHI) is the only HCI journal ranked as A* in the CORE ranking list. Also the Finnish Publication Forum ranks it to the top level. Despite its high rankings, ACM TOCHI does not present its impact factors, possibly due to its slight irregularity in the number of issues per year in 2000-2011. The TOCHI Charter includes categories on ”Experimental and Empirical Studies” and ”Analysis and Evaluation

[7] http://dl.acm.org/

[8] http://ieeexplore.ieee.org/

[9] http://search.proquest.com/abicomplete/

[10] http://www.sciencedirect.com/

[11] http://www.tandfonline.com/

[12] http://online.sagepub.com/

[13] http://search.ebscohost.com/

Techniques". These categories include laboratory experiments; field studies; case studies evaluating user interfaces, interaction techniques, tools, and methods; as well as methods for analysing and evaluating alternative designs and their effectiveness. The term "usability", however, is not used in the charter, so it was found only in 42 abstracts (Table 7).

Table 7: Journal information on ACM Transactions on Computer-Human Interaction.

Journal name	ACM Transactions on Computer-Human Interaction
ISSN	1073-0516
Publisher	ACM
Database used	ACM Digital Library
Journal web pages	tochi.acm.org
Rank in CORE list	A*
Rank in Publication Forum	3
Impact factor	Not found
5-year impact factor	Not found
Search terms	"usability" in abstract
Hits	42
Date of search	February 2nd 2014

Human-Computer Interaction was assessed as the most appreciated journal in the HCI field in the studies by Katerattanakul *et al.* (2003) and Lowry *et al.* (2004). Consequently, its impact factor, especially the 5-year impact, is the highest among the HCI journals. The journal presents its readership to be "*professionals with an interest in the scientific implications and practical relevance of how computer systems should be designed and/or how they are actually used*". The subject classifications do not include "usability" as such, but the subject "web usability" is included, and the journal has included two special issues close to usability: an issue on experimental comparisons of usability evaluation methods in 1998; and a section on beauty, goodness and usability in 2004. Outside these special issues, the number of articles containing the term "usability" is distinctly low (Table 8).

Table 8: Journal information on Human-Computer Interaction.

Journal name	Human-Computer Interaction
ISSN	0737-0024 (Print), 1532-7051 (Online)
Publisher	Taylor & Francis
Database used	Taylor & Francis Online
Journal web pages	www.tandfonline.com/loi/hhci20
Rank in CORE list	A
Rank in Publication Forum	3
Impact factor (2012)	2,250
5-year impact factor (2012)	3,039
Search terms	"usability" in abstract
Hits	20
Date of search	February 2nd 2014

Behaviour & Information Technology (BIT) is *"an HCI journal with a leaning towards human factors*" (Valero & Monk 1998). It is targeted at "*human-computer interaction researchers, software and system designers, cognitive ergonomists and psychologists*". Usability evaluation methods are discussed in many articles in BIT, including a special issue on usability evaluation methods in 1997. Consequently, even 156 articles included the term "usability" (Table 9).

Table 9: Journal information on Behaviour & Information Technology.

Journal name	Behaviour & Information Technology
ISSN	0144-929X (Print), 1362-3001 (Online)
Publisher	Taylor & Francis
Database used	Taylor & Francis Online
Journal web pages	www.tandfonline.com/loi/tbit20
Rank in CORE list	A
Rank in Publication Forum	2
Impact factor (2012)	0,856
5-year impact factor	Not found
Search terms	"usability" in abstract
Hits	156
Date of searches	February 2nd 2014

International Journal of Human-Computer Studies publishes *"original research over the whole spectrum of work relevant to the theory and practice of innovative interactive systems*" including evaluation methodologies. Hence, the term "usability" was found from 129 article abstracts (Table 10). There has also been special issues close to usability, such as user experience in virtual learning environments, and usability of health care information systems.

Table 10: Journal information on International Journal of Human-Computer Studies.

Journal name	International Journal of Human-Computer Studies (International Journal of Man-Machine Studies till1993)
ISSN	1071-5819
Publisher	Elsevier Ltd.
Database used	Elsevier ScienceDirect
Journal web pages	www.journals.elsevier.com/international-journal-of-human-computer-studies
Rank in CORE list	A
Rank in Publication Forum	2
Impact factor (2012)	1,415
5-year impact factor (2012)	2,003
Search terms	"usability" in abstract
Hits	129
Date of searches	February 2nd 2014

International Journal of Human-Computer Interaction "*addresses the cognitive, social, health, and ergonomic aspects of work with computers*", and "*emphasizes both the human and computer science aspects of the effective design and use of computer interactive systems*". The journal provides quite many articles on usability even recently, such as the article by Hertzum and Holmegaard (2013) on the effects of time constraints and various interruptions in thinking aloud (Table 11, and Figure 3 on p. 15).

Table 11: Journal information on International Journal of Human-Computer Interaction.

Journal name	International Journal of Human-Computer Interaction
ISSN	1044-7318 (Print), 1532-7590 (Online)
Publisher	Taylor & Francis
Database used	Taylor & Francis Online
Journal web pages	www.tandfonline.com/loi/hihc20
Rank in CORE list	B
Rank in Publication Forum	2
Impact factor (2012)	1,131
5-year impact factor (2012)	1,284
Search terms	"usability" in abstract
Hits	140
Date of searches	February 2nd 2014

In addition to these highly ranked journals, Interacting with Computers is often mentioned as a relevant HCI journal, but its rank was only B in the CORE ranking list, and level 1 in the Finnish Publication Forum. Therefore, it was given only little attention in this literature review.

On the part of journals on ergonomics and human factors, the CORE ranking list did not include these journals, but the Finnish Publication Forum ranked Applied Ergonomics as a top level (3) journal, and both Ergonomics and Human Factors as leading level (2) journals. The term "usability" could be found in the abstract of 67 articles in the Applied Ergonomics, 66 articles in the Ergonomics and 39 articles in Human Factors[14].

Some rank A* journals on cognitive science and psychology were also included in the literature searches, such as Cognitive Psychology and Cognitive Science. In Cognitive Psychology, the term "usability" was used in 4 articles, but none of them used the term in its abstract. In Cognitive Science, the term "usability" was used in 14 articles, from which only 1 used it in its abstract.[15]

During the literature searches, the field of technical communication and its journal IEEE Transactions on Professional Communication came forth. This journal had a special issue on technical communication and usability studies in 2010. In its commentary, Redish (2010) describes the "*intertwined strands and mutual influences*" of technical communication and usability through her own experiences and reflections of these fields in over 30 years: each discipline has its own contribution to usability, and collaboration and respect for other's skills and knowledge are required for the usability discipline to develop further. Although ranked only as a B level journal in the CORE list and as a basic level journal in the Publication Forum, IEEE Transactions on Professional Communication provided quite an active discussion on the use of the thinking aloud method in usability testing.

2.1.4 Usability testing in HCI conferences

The CHI, British HCI, Interact, MobileHCI and NordiCHI conferences were selected for further analysis (See Table 6 on p. 11), and the literature searches focused on full papers in these conferences. The proceedings in 2013 were used in the analysis, except for the NordiCHI conference that is arranged only on even years. Again, the term "usability" was searched from the abstracts to get a general view of the discussions on usability. Table 12 presents the results and the topics that are closest to usability evaluation.

To get a better view of the discussions on usability and usability evaluation methods, the conference programs in 2013 (or 2012) were browsed to check if there were sessions on these issues. INTERACT 2013 included 4 such sessions; CHI 2013 had 3 sessions on evaluation methods; and NordiCHI 2012 had sessions on "Usability & experiences" and "Usability evaluation in context". The programme for the British HCI 2013 conference was not available on its web pages, and MobileHCI 2013 had no sessions on usability or its evaluation.

[14] Searches made on February 2nd 2014

[15] Searches made on February 3rd 2014

Table 12: Top HCI related conferences, their topics closest to usability testing, and the frequency of the term "usability" in the abstracts in 2013.

Acronym	Proceedings in 2013 • topic in call-for-papers or review category closest to usability evaluation	Web pages	"Usability" in abstract (%)	Full papers
CHI	Proceedings of the SIGCHI Conference on Human Factors in Computing Systems • Usability, user experience	chi2013.acm.org	22 (6%)	393
BCS-HCI	Proceedings of the 27th International BCS Human Computer Interaction Conference • Usability engineering • User experience	hci2013.bcs.org	1 (5%)	20
INTERACT	Proceedings of the 14th IFIP TC 13 international conference on Human-computer interaction • Methods and processes for interface/interaction design, modelling and evaluation, including: ○ Evaluation methods / tools ○ Usability / Usability evaluation	www.interact2013.org	NA	NA
MobileHCI	Proceedings of the 15th International Conference on Human Computer Interaction with Mobile Devices and Services • Evaluation and usability of mobile devices and services	www.mobilehci2013.org	4 (12%)	34
NordiCHI	Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design • Evaluation methods • Usability studies	www.nordichi2012.org	NA	NA

2.1.5 Activity of HCI forums in usability issues

From the top academic HCI forums, the most highly ranked ACM Transaction on Computer-Human Interaction and Human-Computer Interaction did not include the term "usability" in their charter or subject categories, and thereby the number of articles addressing usability issues is very low in these journals. As Figure 3 shows, the ACM Transactions on Computer-Human Interaction has included articles with usability considerations only after the year 2003, and their number has constantly stayed low. On the part of the Human-Computer Interaction, the special issues in 1998 and 2004 can easily be spotted from the otherwise very low number of usability related articles.

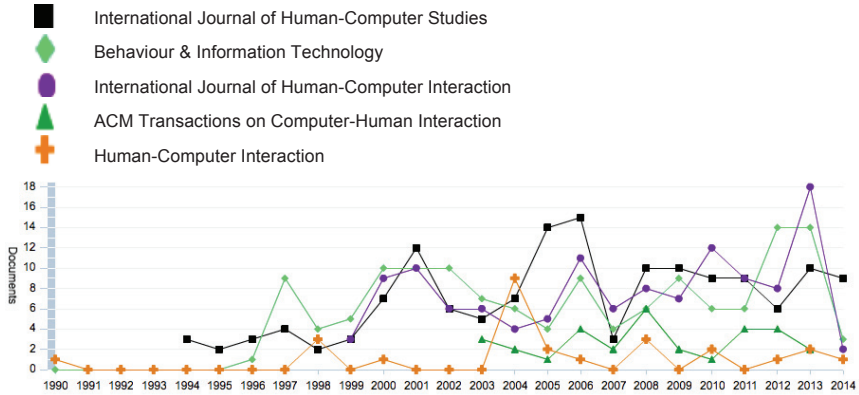


Figure 3: Number of articles including the term "usability" in its title, abstract or keywords in the top HCI forums (Scopus 28th March 2014).

However, if the number of articles including the term usability in their abstract is made proportional to the total number of articles in those journals, the results become somewhat different. When analysing the numbers starting from the year 2010, only the International Journal of Human-Computer Interaction has considerably higher percentage of usability related articles (17,3%) compared to the 6,5%-10,2% in the other journals (Table 13).

Table 13: Number of articles in the top HCI journals from 2010 searched on July 29th 2014, the number of articles including the term usability in its abstract, and the proportion of these articles in the journal.

Journal	Articles 2010- mid 2014	Usability related articles	Proportion
ACM Transactions on Computer-Human Interaction	153	10	6,5%
Behaviour & Information Technology	475	40	8,4%
Human-Computer Interaction	84	8	9,5%
International Journal of Human-Computer Studies	394	40	10,2%
International Journal of Human-Computer Interaction	318	55	17,3%

Figure 3 also shows that the journals leaning towards human factors and cognitive ergonomics, such as Behaviour & Information Technology and International Journal of Human-Computer Interaction, have been very active in usability issues. The special issue in 1997 in Behaviour & Information Technology on usability evaluation methods caused a distinguishable peak in the number of articles addressing usability. Soon after that, however, the numbers started to stay on those same levels, and have recently raised even above that. Also the International Journal of Human-Computer Studies (International Journal of Man-Machine Studies before 1994) that mentions evaluation methodology in its subjects has been active on usability discussions. Indeed, in the search for articles, Behaviour & Information Technology, International Journal of Human-Computer Interaction and International Journal of Human-Computer Studies were the most rewarding sources for usability evaluation and testing, as well as for the thinking aloud method.

Figure 4 shows the corresponding numbers in Interacting with Computers, Applied Ergonomics, Ergonomics and Human Factors. Interacting with Computers has been the most active of these journals in discussing on usability issues, as it tries to foster communication between academic researchers and practitioners, and includes usability and user experience design, as well as empirical evaluations and assessment strategies in its topics[16]. Also the human factors and ergonomics oriented journals have been active, and especially the most highly ranked Applied Ergonomics has remarkably raised the number of usability related articles in the recent years. Overall, the number of articles addressing usability issues in the top HCI, human factors and ergonomics journals has increased. Also the conferences have included several paper sessions discussing on usability evaluation. Thereby, it can be concluded that the research on usability, as well as on usability evaluation methods, is still active – open questions still exist, and new themes seem to arise constantly.

[16] http://iwc.oxfordjournals.org/

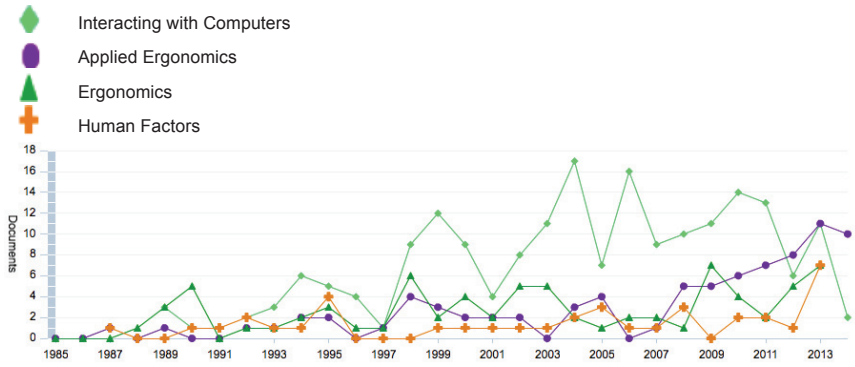


Figure 4: Number of articles including the term "usability" in its title, abstract or keywords (Scopus 28th March 2014).

2.2 Empirical usability studies

Empirical research is about getting knowledge through methods that require direct or indirect contact to the studied phenomenon. The term is sometimes used as a synonym to experimental research relating to the possibility to study through controlled experiments, or as an opposite to theoretical or analytical. Usability testing usually combines several empirical methods including observation, survey, key logging and measuring time on tasks and performance. Therefore, Lazar *et al.* (2010, p. 254) relate it as "*a close cousin of research methods*". The methods enable usability testing to be used as an experimental study following rigorous procedures, as a research method aiming to improve the methods themselves, or as a tool to help the product development team to improve the product. Thereby, it is the goal of the usability testing that determines its type as a research method. (Lazar *et al.* 2010, pp. 254-256)

We have been doing research and giving courses on usability evaluation at the Helsinki University of Technology and Aalto University since 1993 (Koivunen, M.H.T. Nieminen & Riihiaho 1995). The empirical data in this thesis consists of 143 studies that I have participated during these years either as an evaluator in the research team, or as an instructor for the students making the evaluations as course assignments. Each study has had a real customer, such as: GE Healthcare, Kone, Nokia, Polar Electro, Sulake, Suunto, Tekla or Tieto. In addition to the company representatives, the evaluators have usually had a chance to communicate also with the developers of the assessed systems. The students in these courses have been doing their Master level studies or PhD studies in computer science and engineering, information networks, media technology, electrical engineering, communications and networking, industrial design, or cognitive science.

The data for this thesis is gathered from the study reports, and complemented by my own notes. The quality and level of detail of these reports vary considerably over the years and the assignments. Many reports include detailed lists of all the problems found, and some even identify the methods used in finding each problem. However, some of the reports focus only on the most important findings, as for example Molich *et al.* (2004) advice.

Although the goal of a single usability study has usually been to improve the evaluated system, there has also been an overall goal to test and develop new usability evaluation methods for various situations and needs. To get a variety of ideas for improvements, the studies have usually focused on getting qualitative data from the users. This focus is consistent with the results of the survey by Venturi, Troost and Jokela (2006) pointing out the shift from quantitative to qualitative evaluation methods. Also Wixon (2003) emphasises the effect of usability evaluations in the development process, considering evaluations as tools to find and introduce improvements into the product.

Wixon (2003) also suggests that case studies with more precise descriptions of the product and evaluation methods should be reported "*to produce a body of knowledge for applied usability*". However, "*it's not considered a standard practice to publish your data sets or make them available to others*" in the field of human-computer interaction (Lazar *et al.* 2010, p. 12). Instead, the evaluated systems are often mentioned very briefly or left without presentation in the comparisons of usability evaluation methods (*e.g.* C-M. Karat, Campbell & Fiegel 1992; Sawyer *et al.* 1996; J. Karat 1997). If the compared methods have been applied to only one or two systems, the systems have been presented in more detail in some papers (*e.g.* Cuomo & Bowen 1994; J. Nielsen 1994b). Still, there is no established practice in presenting the data sources when several usability evaluation studies are considered. Also Hartson, Andre and Williges (2003) make a comment on this reporting practice by mentioning that "*a significant majority of the comparison studies in the HCI literature on UEM [Usability Evaluation Methods] effectiveness did not provide the descriptive statistics needed to perform a meta-analysis*".

With this background, the empirical data in this thesis is presented quite briefly in a table in Appendix A including columns for the identifier of the study (ID); the type of the evaluated system; the year when the study was made; indicators whether the system was a software system (SW) or smart product (SP); as well as indicators for professional use (Pro) or recreational (Re) use. The summary also shows if the evaluation was made in our research project (RP) indicating that the evaluators were experts in usability, or as a course assignment (CA) with master level or postgraduate students as evaluators. Finally, the summary lists the methods used in the evaluation; whether thinking aloud method (TA) was used; and the number of users in the study including users in the contextual interviews and pilot tests. Table 14 shows a sample of this data.

Table 14: A sample of the empirical data from the usability evaluations included in this thesis.

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
1	Elevator control system	1993	SP	Pro	CA	Pre-test interview, visual walkthrough, usability test, post-test interview, heuristic evaluation	X	4
3	Reverse vending machine	1993	SP	Re	CA	Usability test, post-test interview, observation	X	11

The identifier of the study is used in this thesis to refer to a certain study. For example, Study 3 refers to the course assignment in 1993 in which a reverse vending machine was evaluated. The presentation of the data is very similar to the meta-analyses by Hartson *et al.* (2003) and Hertzum (2006) on the part of presenting the methods used and number of users in their analyses. Also the brief presentation of the type of the system is very similar to Hertzum (2006). However, the presentation does not show the number of evaluators or their level of expertise, as the teams have usually included 3-5 evaluators, and their level of expertise can be drawn from the information whether the evaluations were made in our research projects or as course assignments. The reported number of problems, on the other hand, has depended on the level of detail in reporting the results, and therefore, is left out of the table.

2.2.1 Usability evaluation process in Aalto University

The process of usability evaluation in Aalto University follows very much the basic process presented in Chapter 3.1, but is supplemented with contextual methods when getting familiar with the system. Also the part of communicating the results is emphasised in our process by making it a separate phase after the analyses. When familiarising with the system, we usually make one or two usability inspections to the system to get some early notions on the usability problems. We also interview some users preferably in a real use context to get ideas for suitable test tasks, and to get a deeper understanding of the users' goals. This information is of special value when generating redesign proposals.

Our evaluation process has evolved in these 22 years we have done and taught usability evaluations. In 1993 (Studies 1-8), for example, we started the evaluations in our courses with user testing and made inspections only after them if at all. Only two out of the eight studies included inspections, and the method in both of them was heuristic evaluation. Then 10 years later in 2003 (Studies 94-102), we started the evaluations with inspections to get familiar with the systems and to get a baseline for comparing the effectiveness of various evaluation methods, and to get some experience of the level that inspections can reach without user testing. Heuristic evaluation was still the most popular inspection method used in all the studies, whereas cognitive walkthrough was used only in one study. The variety of the user testing method had spread at that time to include informal walkthrough (Study 96), peer-tutoring and probes (Study 97), symbol and terminology tests (Study 98), paired-user testing (Study 100), pluralistic usability walkthrough and use diaries (Study 102) in addition to the common usability testing. Validation tests with one or two users were also included into the process to get feedback of the redesign proposals before presenting the results to the development teams. Starting from the year 2011 (Studies 128-143), contextual interviews have been used in the familiarising phase, thereby, leading to the process presented in Figure 5.

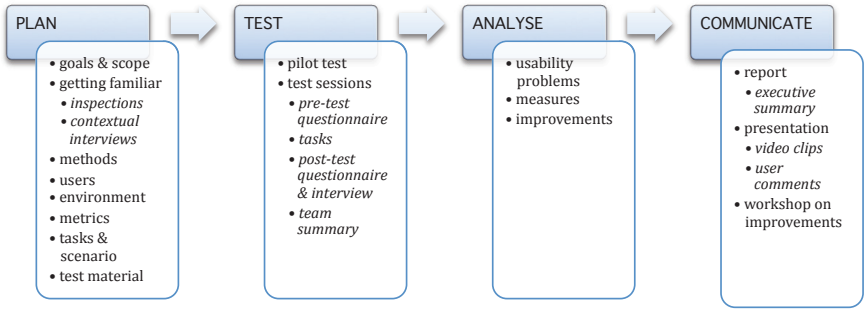


Figure 5: The process of usability evaluation in Aalto University.

In the contextual interviews in the planning phase, current or potential users of the evaluated systems have been observed and interviewed in real use contexts. The results of these interviews have helped to make realistic test plans with motivating test tasks. Understanding the users' goals and expectations has also helped to make realistic redesign proposals. These proposals have then been validated with a few users.

Each evaluation that we have done has entailed several methods including usability inspections, usability tests, interviews and questionnaires. Although we have several variations of the traditional usability test, the most common usability testing method in our studies has been a usability test with predefined test tasks in a controlled test environment. However, evaluations in the field, in users' workplaces or homes, and tests on the move are getting more and more common.

We try to make the rapport in the test sessions relaxed and open. Therefore, we have usually applied concurrent relaxed thinking aloud method to get information about the users' intentions and hesitations. The moderator has been close to the participants to serve as a listener for the thinking aloud (Figure 6). We also want to show that we honour the users' skills and knowledge, so the roles of a test user and a moderator have been very similar to a master and an apprentice used in the contextual inquiry by Beyer and Holtzblatt (1995).

Figure 6: A moderator with a participant in two different test settings. (All the persons in the pictures are members of the evaluating team).

Communicating the results is a crucial phase in the evaluation process if improvements are desired to the system. The results should not depress the developers but give ideas for making the system even better. Therefore, we usually give some concrete redesign proposals that we have validated with a few users, and present these in workshops to the developers so that we can refine the proposals together or generate totally new ideas to fix the problems.

2.2.2 Empirical data

The evaluated products have included both software systems and *smart products*, *i.e.*, appliances that integrate embedded information technology with physical products. Unlike traditional computers, smart products are not general platforms, but usually intended for a certain set of tasks, and, therefore, provide only minimal controls for interaction (Keinonen *et al.* 1996). These smart products have included professional appliances and consumer electronics, such as: anaesthesia monitors, patient monitors, televisions and gaming slot machines. Software systems, on their part, have included various mobile and web services for recreational use, and several applications for professional use. Figure 7 shows some examples of the systems that we have studied.

Figure 7: Systems assessed in our usability evaluations include a mobile financial application (Study 138), a travel and expense management system (Study 127), a calendar service for heart rate monitors (Study 95), an entertainment system for big crowds (Study 132), heart rate monitors (Studies 30, 58, 66, 72, 80, 81, 88, 89, 97, 98 & 121), gaming slot machines (Studies 12, 65, 70, 76, 77 & 96), a building information system (Study 128), and anaesthesia monitors (Studies 44 & 125).

A majority (66%) of the evaluated systems have been software systems, and only 34% have been smart products. Figure 8 shows this ratio for each year in 1993-2014. Most of the software systems have been intended for professional use (70%), whereas the smart products have been mostly for recreational use or intended for a wide set of users (75%). Figure 9 shows this division into professional and recreational systems in 1993-2014. In Studies 115 and 142, the portal for building permissions and the mobile conference program applications were intended both for professional and novice users, so the total numbers in the following figures are not identical in 2009 and 2014.

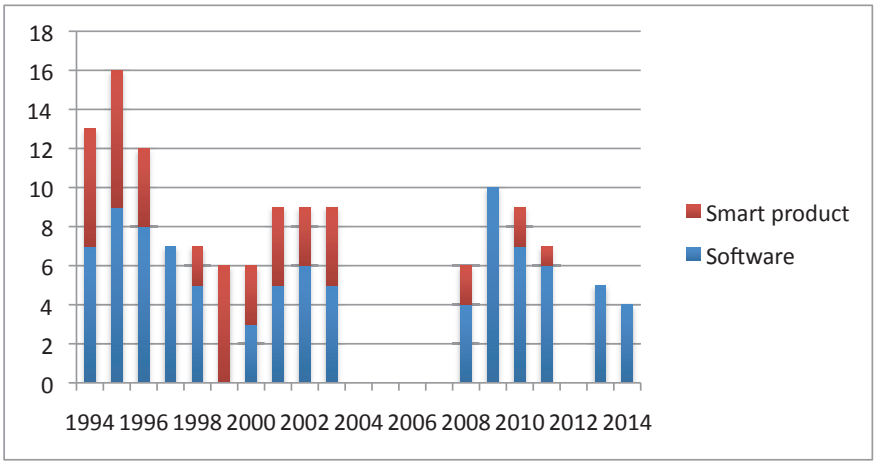


Figure 8: Division of the evaluated systems into smart products (49 out of 143) and software systems (94 out of 143) in 1993-2014.

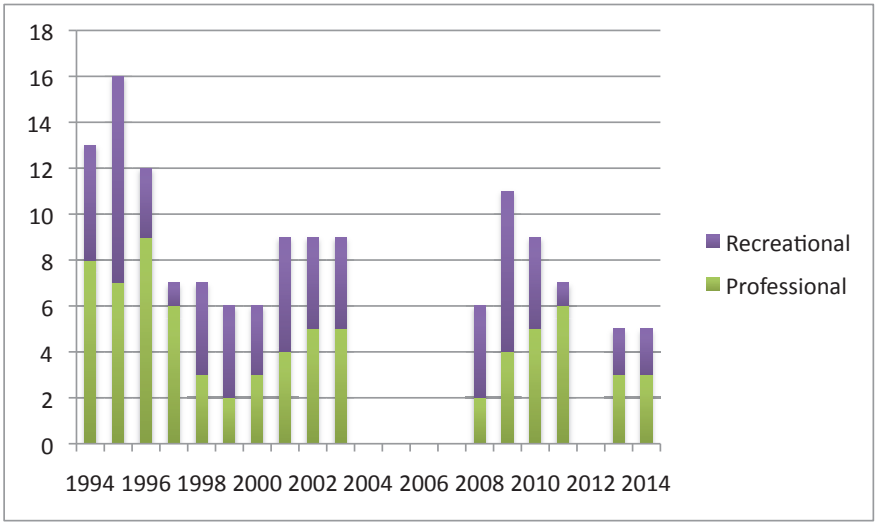


Figure 9: Division of the evaluated systems into professional systems (78 out of 143), and systems for recreational use or a wide set of users (67 out of 143) in 1993-2014.

We have usually applied at least one usability testing method with the users, and one inspection method without user involvement. Only two studies (Studies 112 and 119) did not include user testing as they were part of our research projects in which no resources were available for user testing at that point. Usability tests have been conducted in as many as 122 studies, and heuristic evaluations in 103 studies. Thereby, usability tests and heuristic evaluations are the most common pair of evaluation methods in our studies in addition to interviews and questionnaires before or after the tests. Figure 10 demonstrates the numbers and their ratios of how often we have applied different usability testing methods in our studies.

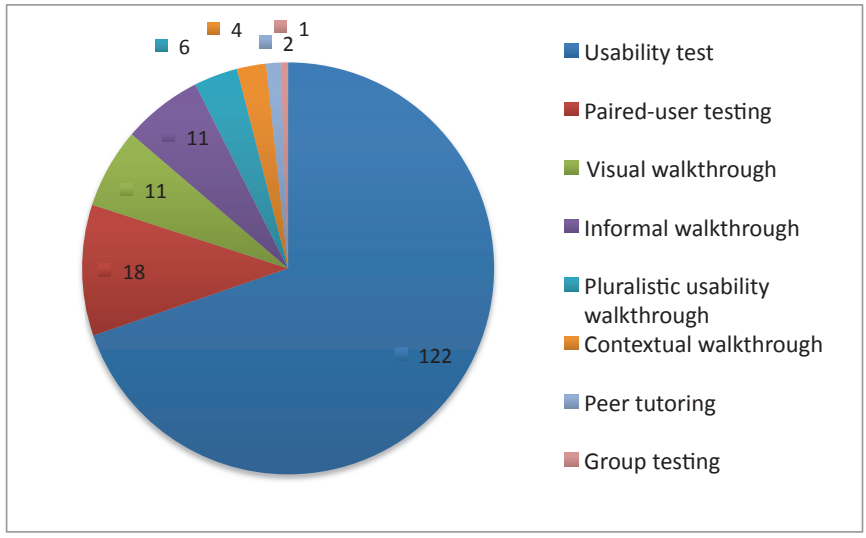


Figure 10: Usability testing methods used in our evaluations and the number of studies in which they have been used.

The figure does not show the number of times that thinking aloud method has been used, since it has been used in all of our usability tests, informal walkthroughs, contextual walkthroughs and visual walkthroughs with the exception of the studies in call centres (Studies 59 and 60), the summative evaluation of a patient monitor (Study 44), and a comparative test of two research systems with a focus on efficiency and effectiveness (Study 141). Thereby, thinking aloud has been used in 128 of our studies.

The thinking aloud in our tests has not complied with the classic instructions by Ericsson and Simon (1980, 1984), as we have asked the test users also to explain why they are doing something instead of just verbalising what they are doing. This way, we have aimed for more information about the users' experiences during the task performance. Interviews and questionnaires have also been left out of the figure, since these methods are not categorised as usability testing methods but as more general user testing methods. The average number of test users in our studies has been from 7 to 8 users.

We have had a test moderator next to the user in all the studies except for my own experiment (Study 123) in which half of the test users performed alone. Most of our studies have been made in controlled settings in usability laboratories or seminar rooms; the contextual walkthroughs make almost the only exceptions in this matter. On the part of the fidelity of the assessed system or prototype, most of the systems have been finished or nearly finished products with the exception of paper prototypes in the pluralistic usability walkthroughs.

From the selection of usability inspection methods, cognitive walkthroughs (N=36) have been quite frequently used to supplement the studies in addition to heuristic evaluations, but other methods have been quite rarely used. Figure 11 shows the number of times that various usability inspection methods have been used in our studies.

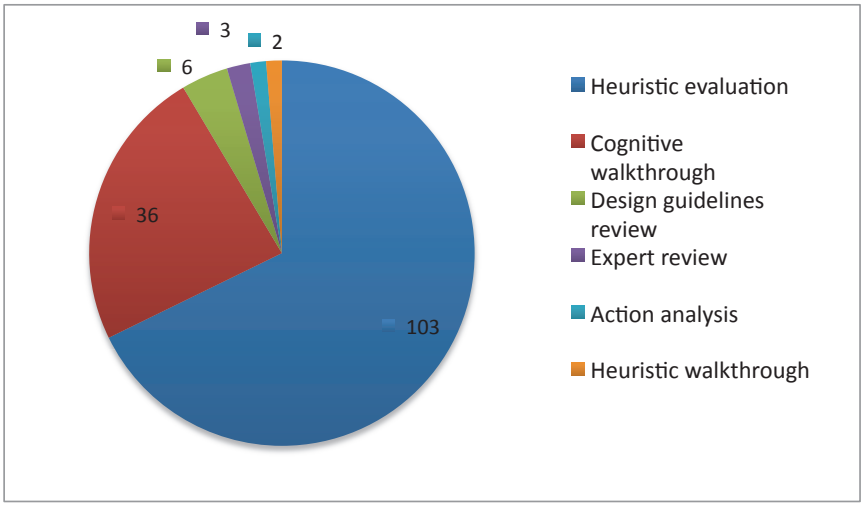


Figure 11: Usability inspection methods used in our evaluations and the number of studies in which they have been used.

2.3 Experimental research

Experimental research is about testing research hypotheses through controlled empirical tests. A hypothesis is a focused statement predicting a difference between experimental treatments; a *null hypothesis* claims that there is no difference. The empirical tests try to find statistical evidence to confute the null hypothesis and to support the alternative hypotheses. The hypotheses should state both the *independent and dependent variables* of the study. Independent variables refer to the experimental treatments that are modified in the tests either between subjects or within subjects, and the dependent variables refer to the measurable outcomes of the test. The statistical significance of a difference in treatments is reported by means of a *probability value p*. This value represents the probability of claiming a difference when there is none. It is important to be confident of the existence of an effect and a difference in treatments, so the *p*-value should be less than 0,05, and in critical contexts even less (Lazar *et al.* 2010, p. 34; Sauro 2006).

2.3.1 Designing experiments

When designing an experiment, two basic decisions have to be made: the number of independent variables, *i.e.*, variables that are modified in the tests, and the comparisons between subjects or within subjects. When more than one independent variables are used, a *factorial design* is needed. A factorial design gives means to investigate the effect of all independent variables at the same time, as well as their interaction effects. (Lazar *et al.* 2010, pp. 44-56) In this thesis, there are two independent variables in the experiment, *i.e.*, the moderator presence and the use of the relaxed thinking aloud, thereby leading to a

factorial design. Similarly, the amount of different values for each independent variable has to be considered. In this thesis, both the variables have two alternatives, as a moderator either is next to the test user or not, and the test user either thinks aloud while doing the tasks or performs silently.

Another important decision is whether to use between-group or within-group design. *Between-group* or *between-subject* design means that one user is exposed only to one value of each independent variable, so that the results of the groups are compared to each other. In *within-group* or *within-subject* design, each test participant is exposed to every value of each independent variable. Thereby, within-group design enables statistically significant results even with relatively small number of participants, but the order in which the participants are exposed to different treatments requires special attention to compensate possible learning effects. (Lazar *et al.* 2010, pp. 44-52) In this thesis, the between-group design was more appropriate, because the use of the thinking aloud method before acting silently, or a lengthy presence of a moderator before acting alone could have easily biased the users' performance and the test results. Furthermore, finding usability problems from the assessed system was presented as the main goal to the test users, making different test settings and repetition of same test tasks in one test session quite inexplicable.

2.3.2 Analysis of experiments

In statistical analysis, there are descriptive statistics that illustrate the data, and inferential statistics that show how representative this data is for the general population (Hughes 1999). The most common indicators of the data include means, medians, modes, variances, standard deviations and ranges. The statistical significance in the difference of means between treatments can be calculated by *t* tests and analyses of variance. A *t* test can be used when there is only one independent variable that has only two conditions. Otherwise, analysis of variance (ANOVA) is needed. There are several types of *t* tests and analyses of variance for different experimental designs. (Lazar *et al.* 2010, p. 73-91) For example, in this thesis, as there are two independent variables with two conditions in a between-group design, a factorial analysis of variance must be used. Other common statistical methods include correlation, regression and chi-square (χ^2) tests.

The use of parametric tests, such as *t* tests or analyses of variance, assume that the data is collected from a normally distributed population, the measures are scaled by equal intervals, and the variance in different groups is close to equal (Lazar *et al.* 2010, p. 91-92). For example, in the analyses in this thesis, the evaluators used a scale from 1 to 3 to assess the certainty they had on the existence of a usability problem. The value 1 indicates that the evaluator assumed a problem, value 2 indicates that the evaluator was almost sure about a problem, and value 3 indicates that the evaluator was definite of a problem. Although the numbers have equal intervals, their meanings do not, and therefore, a non-parametric statistical test, such as a chi-square test, needed to be used to compare the results of different treatments.

In a chi-square test, the results have to be outlined in a frequency count table summing the amount of answers for each alternative for each treatment group. Then, with a factor called a *degree of freedom* and other data, the chi-square test can be calculated to get a probability value *p* that indicates the statistical significance of the differences between treatment groups. The chi-square test does not make as many assumptions on the data as parametric tests, but it does assume that each participant has selected only one alternative, such as ”yes” or ”no”. (Lazar *et al.* 2010, p. 92-94) Although the expected cell frequencies are recommended to be at least five, the Pearson’s chi-square test has been found to be very robust with small cell frequencies, “*even when the expected frequencies in one or two cells are as low as 1 or 2 when N is 20 or more*” (Camilli & Hopkins 1978).

To analyse the answers to open-ended questions in questionnaires or user comments in usability tests, qualitative data analysis is required. If the material to be analysed consists of text, content analysis is the favoured method (Silverman 2000, p. 128). Content analysis starts with getting familiar with the material, and defining the focus of the analysis. After that, the text content is “coded”, *i.e.*, classified to either predefined or emergent categories. When classifying the data, interesting patterns are sought, as well as connections within and between these patterns. (Lazar *et al.* 2010, p. 289-301)

2.3.3 Reliability and validity

Experiments should be replicable and produce similar results to be reliable. Human subjects, however, make it challenging to get recurring results even with same subjects (Lazar *et al.* 2010, pp. 57). For example, when a test participant repeats a task several times, the performance time varies causing *random errors*. The effect of these random errors can be minimised by enlarging the sample size. *Systematic errors*, on the other hand, cannot be faded by larger sample sizes, since they cause a bias to only one direction. Test procedure, moderator behaviour, test environment, participants and measurement instruments are major sources for these systematic errors (Lazar *et al.* 2010, p. 59). A team of evaluators with different perspectives can enhance objectivity, as well as the use of an outside evaluator (Hughes 1999).

Validity, on its part, concerns with the question whether the measurements and conclusions correspond to the real world. Validity is often divided into internal validity asking if the measured thing really is an attribute of the studied phenomenon, and into external validity referring to how well the measurements apply in the field outside the experimental setting (Hughes 1999). The validity of qualitative research can be improved by respondent validation, as well as method and data triangulation (Silverman 2000, p. 177). Respondent validation refers to evaluators checking their interpretations and conclusions from the original subjects. *Triangulation*, on the other hand, means using several research approaches to study a phenomenon. For example, Jick (1979) presents a continuum of triangulation design from scaling to between methods approach: *scaling* quantifies qualitative measures; *within methods approach* uses multiple techniques of the same method; and *between methods*

approach uses several distinct methods for convergent validation. Chauncey Wilson (2006) adds also facilitator, observer, user group, geographic and qualitative-quantitative triangulations to this continuum.

The observer or investigator triangulation is emphasised in qualitative data analysis. The investigator reliability, on its part, has two dimensions: *stability* and *reproducibility*. Stability refers to the reliability of the same investigator coding the same phenomena in the same way if coding is repeated, *i.e.*, intra-coder reliability. Reproducibility then, refers to inter-coder reliability, *i.e.*, several investigators interpreting and coding the phenomena similarly. Measures, such as Cohen's Kappa (κ), can be used to measure "*the proportion of agreement beyond what would be expected on the basis of chance*" when the coders use given categories, so it ranges between 0 and 1 (Hartson *et al.* 2003). If Cohen's Kappa is over 0,60 the inter-coder agreement is satisfactory, and over 0,80 is near-perfect (Lazar *et al.* 2010, p. 299).

The coders' participation in the design of the experiment is also a potential source of bias. Subjective or inside coders, *i.e.*, evaluators who have participated in the design of the study or in the data collection easily understand the terms and concepts in the data, but can neglect some new viewpoints as they are so familiar with the current ones. Outside or objective coders, on their part, may need more guidance in coding, and may miss some interesting phenomena if they are not experts in the research domain, but are more open-minded to new viewpoints. Therefore, outside coders are often considered as more reliable than inside coders. (Lazar *et al.* 2010, pp. 296-299)

On the part of assessing the evaluator effect, Hertzum and Jacobsen (2001) propose a measure called any-two agreement. It is "*the number of problems two evaluators have in common divided by the number of problems they collectively detect, averaged over all possible pairs of two evaluators*". This measure takes into account the effects of a small number of evaluators both in the expected overlap of the findings and in the overall number of detected problems. (Hertzum & Jacobsen 2001)

In the experiment of this thesis, both quantitative measures and qualitative data were analysed to support triangulation. Two outside coders in addition to myself analysed the test sessions with the help of the recordings, so also investigator triangulation was used. Since the evaluators did not have a common set of problems to be categorised, or predefined categories for the problems, the measure of any-two agreement was used in assessing the evaluator effect in the experiment.

3 Usability testing

The International Standard ISO 9241-11 (1998) defines usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. The *effectiveness* means the accuracy and completeness with which users achieve specified goals. *Efficiency* means the resources expended in relation to the effectiveness. *Satisfaction* means freedom from discomfort, and positive attitudes to the use of the product. (ISO 9241-11, 1998)

The usability of a product is not an attribute of the product alone, but an attribute of interaction with the product in a context of use (J. Karat 1997). A product can therefore have very different levels of usability when used in different contexts, so the context should be clearly defined for design and evaluations. The *context* in the definition includes (ISO 9241-11, 1998):

- The users and other stakeholder groups,
- Their characteristics, such as: knowledge, skill, experience, education, training, physical attributes, habits, preferences and capabilities
- Their goals and tasks, and
- The environments of the system including the technical environment, as well as the physical, social and cultural environments in which the product is used.

Maguire (2001) divides these factors of context into more detailed components, and also Alonso-Ríos *et al.* (2010) present a very detailed taxonomy for the context-of-use in usability studies. For example, Alonso-Ríos *et al.* (2010) divide the user's experience into previous experience with the system and similar systems, whereas Maguire (2001) divides the same factor into product experience, related experience, task knowledge, organisational knowledge, training, input device skills, qualifications and language skills.

Usability can be defined also with other parameters. For example, Jakob Nielsen (1993) defines usability by five attributes: learnability, efficiency, memorability, error prevention and subjective satisfaction. Also ISO/IEC 9126-1 standard (2001) "Software engineering – Product quality – Part 1: Quality model" includes learnability in usability but additionally lists understandability; operability; attractiveness; and compliance with standards, conventions and style guides as characteristics of usability. Furthermore, Han *et al.* (2000) make a very detailed division of usability into 48 usability dimensions, and Seffah *et al.* (2006) list 10 factors with 26 measurable criteria and 127 specific metrics for usability.

The attributes can also be viewed on the basis of the level of interpretation or effect that the use of the system can have. Nielsen (1993) presents usability as a part of system usefulness together with system utility. These, on the other hand, are part of practical acceptability, and finally, part of system acceptability (Figure 12).

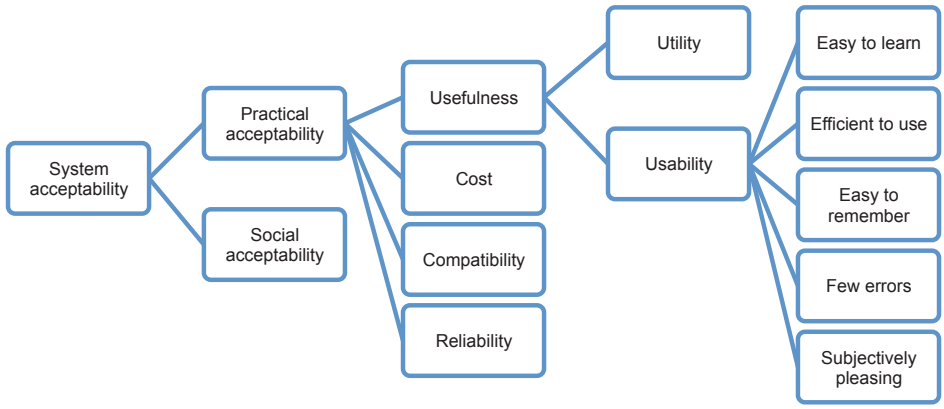


Figure 12: Jakob Nielsen's (1993) framework for usability.

In a very similar way, Sengers and Gaver (2006) identify various levels of interpretation about the use of a system. The lowest levels are related to the operation and control of the system from recognizing a button and understanding its functionality to figuring out how to do a certain task. At the middle levels, the utility of the system is assessed including the purpose of the system and the role it can have in the users' life. At the highest levels, the value of the system is interpreted considering the system value to the users and their social groups and culture – similarly to Nielsen's (1993) system acceptability and Cockton's (2006) value-centred design.

Sengers and Gaver (2006) also point out that evaluation traditionally assumes new systems to produce a single authoritative interpretation. However, new systems are used in varying contexts by different kinds of users with different backgrounds, needs and social contacts, so most likely, also the use of the system will differ considerably in these situations. Therefore, already the design of new systems should support multiple interpretations, and also usability evaluations should take this into account. (Sengers & Gaver 2006)

These examples show that even if usability is considered only as situational ease of use and usefulness, it entails many factors affecting to its assessment and measurement. However, there are several other viewpoints to usability as research groups from different research fields have their own view of usability entailing their own goals, research questions and research methods (Sullivan 1989). For example, Hertzum (2010) lists six perspectives to usability that he calls images of usability, and emphasises that the list is not an exhaustive one.

This list includes universal, situational, perceived, hedonic, organisational and cultural usability (Hertzum 2010):

- Universal usability aims for systems that are accessible and usable for everyone,
- Situational usability refers to the quality in use in a specified context,
- Perceived usability refers to the user's subjective experience when using the system,
- Hedonic usability refers to the joy of use, focusing on pleasure and emotion,
- Organisational usability refers to the collaborative activities and the match to the work practices in an organisation, and
- Cultural usability refers to the differences in users' cultural backgrounds that affect to the way users perceive and experience a system.

There is also a more comprehensive approach to usability related to business organisations, their decision-making and priority setting (*e.g.* Happ 1994, Bloomer, Croft & Kieboom 1997; Rosenbaum, Rohn & Humburg 2000). This *"strategic usability" "embeds usability engineering in the organisational processes, culture and product roadmaps"* contributing to all decisions in the system development (Rosenbaum *et al.* 2000). Although there are these different viewpoints to usability and variations in the way usability professionals construe usability, the most general view tends to emphasise the utilitarian constructs over the experiential, and individual level over the organisational or environmental fit (Hertzum & Clemmensen 2012). From these various perspectives, this thesis concentrates on the situational and perceived usability.

3.1 Process of usability testing

Usability testing, as it is presented in textbooks and used in practice, is strongly based on carefully prepared test tasks that the test users are asked to perform in controlled conditions (*e.g.* J. Nielsen 1993). Thereby, all participants perform the same tasks under as similar conditions as possible to enable easy and reliable data gathering, focusing on quantitative data. This approach is appropriate for settings with well-known tasks and outcomes, and is highly effective in influencing decision making in organisations, but is limited in its ability to gather data on true user tasks, task flows, user profiles, and contexts of use (Rohn *et al.* 2002). As Greenberg and Buxton (2008) state, quantitative empirical evaluations give us something that appears to be scientific and factual instead of only expressing opinions.

The process of usability testing has three major phases: design and preparation of the test, conducting the test sessions, and analysing and reporting the results. Instructions for planning and conducting a usability test are presented in several articles and textbooks, such as: Hansen (1991); Dumas and Redish (1993); Nielsen (1993); Rubin (1994); Wixon and C.E. Wilson (1997); Dumas and Loring (2008); Rubin and Chisnell (2008); Tullis and Albert (2008); and Barnum (2011). Therefore, this chapter gives just an overview of the basic process and its phases. The focus is on planning and conducting the tests, but also

the phases of analysis, reporting and communicating the results are addressed to some extent. The engineering side of me comes forth in this chapter as various lists and steps of activities required for various phases of usability testing. Despite of their simplicity, these lists are not meant as recipes, but only as practical checklists of things to consider in these phases.

3.1.1 Planning usability evaluations

As in theatre or movie production, preparation of a detailed manuscript and setting the stage are important steps also in usability testing. In the testing, the evaluators build up a stage and props suitable for the use context to minimise distractions of the laboratory equipment. The moderator then works as a director leading the test users through their own roles as naturally as possible, at the same time honouring their skills and knowledge. (Salzman & Rivers 1994)

The purpose of a test plan is to get an approval of the management and other people involved, and to organise and clarify the goals of the test (Hansen 1991). When talking about goals for usability evaluation, there are two different levels of goals: the overall goal and motivation for doing the evaluations, and the more specific usability goals for selected attributes. Common reasons to do testing include: ensuring that the product reaches its minimum level of usability; getting feedback of how well the objectives are met; and identifying potential usability defects in the product (Bevan & Macleod 1994). Testing with users also gives an opportunity to collect new information about the users' needs, and getting their opinions on the strengths and weaknesses of competing design solutions (ISO 9241-210 2010). Comparisons to competitive products, finding issues for training, and simply educating user interface designers to be more empathic towards users are also viable reasons for testing.

Within these general goals, there usually are some more specific goals for a single evaluation. These goals should be specific to ensure that the tests really measure and assess the attributes of interest. These goals, thereby, also affect the selection of evaluation methods, test participants, the number of test users, and the set of test tasks.

3.1.2 Test moderator

As usability testing started, it was often considered as an experimental research method rather than an exploratory study (*e.g.* Preece *et al.* 1994, p. 652). Therefore, care was taken to keep the test sessions as similar as possible for all the test participants, and interaction with the test users was minimised to avoid biasing the users' performance. Dumas and Redish (1993, p. 30), for example, advice to leave the test user alone "*to simulate what will happen when individual users get the products in their offices or homes*". On the other hand, having a moderator near the user was also recommended in some references already in the early 1990's, but still with minimal interaction with the test user (*e.g.* J. Nielsen 1993, p. 190).

Dumas and Loring (2008, pp. 125-131) summarise some advantages and challenges of having a moderator in the test room: the integrity of the test data

is less biased if the moderator is not present, but the user may feel uncomfortable alone in the test room, and this in turn, may bias the results. Being close to the user also makes it possible to observe users' first reactions and impressions when they explore new systems. Dumas and Loring (2008, pp. 128-129) conclude that there are situations in which the decision whether to have a moderator beside the users or not stems straight from the objectives of the test, interactivity of the product, and characteristics of the participants. For example, if the objective of the test is to explore design alternatives, or do other kind of formative testing, having a moderator next to the user and probing on users' expectations and perceptions gives lots of valuable information for further development. Also, if the system has very restricted interactivity, or the users are likely to require special help in using the system, the moderator should always sit next to the users. (Dumas & Loring 2008, pp. 128-129)[17]

Guidelines for interacting with the test users emphasise the risk of biasing the users' behaviour by the moderator's tone of voice or body language. For example, raising or lowering the pitch of voice may be interpreted as agreement or disapproval (Rubin & Chisnell 2008, p. 203). These instructions also include hints for giving positive feedback on success and failures (Barnum 2011, p. 208-209), reminding to think aloud, and probing for insights on interesting issues (Dumas & Loring 2008, pp. 72-75).

3.1.3 Test participants

The participants in a usability test should represent the real users as well as possible. Friends, family or co-workers are not preferred (Hansen 1991), because a close relationship between a participant and a moderator can easily bias the results. It is usually easier for a test user to interact with a moderator and observers, and criticise the product after potential problems, if the observers are not acquainted (Schrier 1992). However, in East Asian cultures, it is recommended that the test users are familiar with the moderator and that they are of higher status than the moderator to feel comfortable in giving negative comments on the evaluated system (Yeo 2000).

Recruiting participants is sometimes very time-consuming and hard, so the evaluators should make a specific plan, how to recruit the participants and how to reward them. Selecting the participants has the following steps:

1. Developing user profiles,
2. Selecting user subgroups for a test,
3. Defining and quantifying characteristics for each subgroup, and
4. Deciding how many participants to include in a test.

Developing user profiles means thinking broadly about the users of the product, what characteristics they share and which characteristics distinguish one user group from another. In this phase, several user groups are usually identified, so the next step is to select the subgroups for the planned test and to set

[17] The effect of having a moderator near the test users are further discussed in Chapter 3.7.5 and in the experimental part of this thesis.

requirements for the identified characteristics including the level of expertise in various aspects. (Dumas & Redish 1993, pp. 119-129)

Nielsen (1993, p. 43) divides users' previous experience into three categories: experience with the evaluated system or its previous versions; experience with computers, information technology and technical devices in general; and experience with the task domain. Maguire (2001) adds to this list organisational knowledge, training, input device skills, qualifications, and language skills. Bødker and Madsen (1998) also remind of the effect of the organisational situation, since managerial users, secretaries and case workers all have different requirements for the system. For example, when Bødker and Madsen studied a hospital system to be used by nurses, they noticed that also the shift the nurses were taking changed their requirements. Hence, there are many factors that should be considered and prioritised when selecting a suitable set of test participants and also the situations to be evaluated.

Especially in evaluating occupational systems, the test users should be experts in the task domain, but the suitable level of experience with similar systems and information technology in general should be considered according to the test goals. Similarly, the level of expertise should be considered when assessing products for recreational use. Novice users are good at revealing problems with learnability and affordance, whereas expert users are good at revealing inconsistencies with similar products and in finding illogical features.[18]

If the evaluated system is intended for children, Markopoulos and Bekker (2003) recommend to pay special attention to the test participants' capacity and tendency to verbalise; capability to concentrate; motivation; ability to adjust to strange environments; trustworthiness of self-report; and knowledge and skills. On the part of participants with cognitive disabilities, Lepistö and Ovaska (2004), advice to put special effort on gaining the participants' trust already before the evaluation sessions.

3.1.4 Number of test users

The required number of test participants depends on many issues, such as: how many subgroups should be covered; how much resources are available for the tests; and how important it is to get statistically significant results (Dumas & Redish 1993, p. 127). The skills and experience of the moderator as well as the number of iterations planned for the design also affect the decision. According to the studies by Virzi (1992), Nielsen (1994a) and J.R. Lewis (1994), the first few participants give the most information, and additional participants are likely to reveal less and less problems. For discovering about 80% of the usability problems, 4 or 5 users were required in the studies by Virzi (1992), and 5 in the studies by J.R. Lewis (1994) and Nielsen (1994a). In Virzi's study, moderators were skilled usability experts, whereas in Nielsen's studies, the moderators were novices in usability testing. As a premise for these results, J.R. Lewis sets an average likelihood of problem detection between 0,32 and

[18] The effect of the test users' level of expertise is further discussed in Chapter 3.7.2.

0,42, and for more rare problems, more users are required to have the same coverage (J.R. Lewis 1994).

Virzi's (1992) studies also show a correlation between the severity of a usability problem and the likelihood to find it: as problem severity increases, the likelihood that it is found within the first few participants also increases. As Nielsen (1994a) mentions, this follows by definition if the severity of a problem depends on the frequency with which the problem is met, and on the effort that the user has to make to overcome the problem. Thereby, a severe problem is more likely to occur, and the moderators will easily detect the situation in which the user has the problem (J. Nielsen 1994a). The study by J.R. Lewis (1994), however, does not support these findings of the correlations between the problem severity and the likelihood to detect it.

The average likelihood of detecting a problem has a substantial effect to the required number of test users if a certain proportion of the total usability problems is reached for (J.R. Lewis 1994). The basic formula for estimating the proportion of problems discovered is $1-(1-p)^n$, where p is the problem-discovery rate, and n is the number of test users. Thereby, the required sample size may vary substantially according to the complexity of the system, the level of the development and other issues. For estimations made from small samples, J.R. Lewis recommends to make an initial estimation of the required sample size with the data from the first two participants, and then adjust it with the data from two additional users. (J.R. Lewis 2001b) Formulas for more detailed estimations can be found for example in the article by Kanis (2011).

Various studies have shown that 5 test users are not always enough for the 80% coverage. For example, the meta-analysis by Hwang and Salvendy (2010) shows that 9 users were often required to detect 80% of the problems. Molich *et al.* (2004), on the other hand, found no relation between the number of test users and the number of detected usability problems. Molich and Dumas (2008) repeated a similar study 4 years later, and noticed that even a combined list of 9 testing teams did not cover all the problems detected in the system, even though each group had from 5 to 15 participants in their tests. From another perspective, Caulton (2001) reminds that the more subgroups are identified from the users, and the more distinctive they are, the lower the expected proportion of detected problems is for a single user. All in all, the advice by Dumas and Redish (1993, p. 128) to include at least 3 participants from each subgroup, and to have 6 to 12 participants in 2 to 3 subgroups, is a practical generalisation of the recommendations.

3.1.5 Test environment and use context

Usability tests can be conducted practically anywhere. As Anna Wichansky pointed out already in 2000: "*Today, usability testing is being conducted in simulated homes, classrooms, cars and virtual reality environments. There are portable lab systems that can be carried to remote user sites to collect data, so usability engineers can go to their users if their users cannot come to them.*" The real use context with tasks emerging from the users' work reveals problems that would be hard to detect in laboratory settings with predefined

tasks. For example, McDonald, Monahan and Cockton (2006) estimated from their data from Contextual Interviews that about 2/3 of the problems identified in these interviews were related to the context of use instead of the evaluated system. Consequently, the real context, tasks emerging from the users and rich data set are considered as the main advantages of the field methods, whereas their cons include the potential of being laborious, requirements of long timescales and problems in data analysis (Monahan, Lähteenmäki, McDonald and Cockton 2008).

The customer site is more familiar to the participants making it easier for them to relax. However, customer site is challenging for the evaluators, because interruptions are hard to control and the equipment varies from site to site, or has to be brought along. Specific laboratories, on the other hand, offer dedicated equipment and peace for testing (Bawa 1994), but the participants must travel to these laboratories. In addition, the artificial environment can produce unrealistic results (ISO 9241-11). Still, testing in laboratories gives greater control of the variables critically affecting the level of usability, and the measurements are more precise than in the field tests.

Despite the physical location of the test, the context of the test should include the most critical components of the actual use context (Bødker & Madsen 1998). Ethnographic and contextual methods can be used to identify these components, and to assess their relevance. Issues to be considered include realistic test data; having everyday materials and tools available; need for simulated interruptions and time pressures; need for cooperation between users; placement of the system and other relevant material; as well as the organisational situation affecting to the role of the user (Bødker & Madsen 1998). To simulate interruptions, Bødker and Madsen (1998) suggest that specific ”situation cards” could be used, and videotapes from real situations could be used to simulate a particular alarm.

The use context has numerous attributes in several levels of detail as the context-of-use taxonomy proposed by Alonso-Ríos *et al.* (2010) very well demonstrates. Despite this wide range of possible contexts, the most common contexts easily cover over half of the use situations, as in the studies by Kim *et al.* (2002). In their studies, Kim *et al.* studied the use contexts of mobile Internet with 37 participants, and categorised the use contexts with 8 parameters: reason to use the system; emotional status; one or two hands used; moving or staying still; visual distractions; auditory distractions; number of people around; and the level of interaction with others. The results show that just 2 types of use contexts out of the 256 theoretical possibilities covered over 20% of the reported use sessions, and 14 contexts covered over 50% of the sessions. Although these results cannot be generalised to all systems, they show that some contexts are often substantially more common than the others, and are thereby worth more attention in the design and evaluation. The results also revealed that the availability of hands, movement of legs and the number of people around the user had the greatest effects in the types of usability problems found in the mobile Internet. (Kim *et al.* 2002)

3.1.6 Usability metrics

The goals of a test should be specified right from the beginning to be able to focus the evaluation and limit the scope to certain user groups, attributes and tasks. For each attribute of interest, there should be measures set to be able to evaluate the level of usability, such as: time to finish a task; time to find information in the manual; number of wrong menu choices; number of repeated errors; or observations of frustration or confusion (Dumas & Redish 1993, p. 185). These measures are important especially in summative testing, but can be useful also in formative testing to give quantitative results for comparisons with former tests or competing products. However, in formative testing, the measurements should emphasise metrics other than performance times, such as in Table 15, especially if thinking aloud method is used in the tests.

Table 15: Examples of usability requirements (modified from Whiteside, Bennett & Holtzblatt 1988).

Attribute	Measuring concept	Measuring method	Current level	Accepted level	Planned level	Optimal level
Installability	Installation task	Proportion of successful installations without manual	<50%	50%	90%	100%
Affordance	Video programming task	Proportion of users finding the function without manual	10%	30%	80%	100%
Error rate	Data entering task	Frequency of incomplete data inputs	50%	30%	10%	5%
Comparison with competitors	Questionnaire	Relation of answers	Equal	60% prefers	70% prefers	90 % prefers

There are numerous possibilities in measuring different components of usability and widen the scope of metrics from the general performance times, as the reviews on usability measures by Hornbæk (2006) and Seffah *et al.* (2006) clearly demonstrate. However, also single score metrics for usability have been presented to help in comparing products, and assessing the effects of changes in a system over time. For example, Sauro and Kindlund (2005) present SUM, a single, standardised and summated usability metric that combines the task completion ratios, error rates, satisfaction scores and task times into a single score. Yet, to give better means to assess the reasons for the results, the results for each metric should be presented separately before combining the measures (Hornbæk 2006).

The usability metrics should include both objective and subjective measures, but they should be clearly separated, and analysed for correlations. For example, the time to complete a task does not generally vary a lot between the test users, but the way the users perceive the required time may vary much more, and, thereby, can reveal frustrations and rigidity in the use of the system. The review also shows that too many of the usability measures focus on usability at a micro level. This means that the studied tasks are brief, and mostly focus on measuring perceptual and motor aspects. Realistic tasks, however, are more on the macro level; require problem solving and critical thinking; and last from several hours to even months. Overall, this type of measures assessing the evolving use and usability over time are too scarcely used. (Hornbæk 2006)

3.1.7 Test tasks and scenario

Test tasks should be realistic and represent the actual use expected in the field as well as possible (J. Nielsen 1993, p. 185). The tasks should cover the most important parts of the product, and reflect the attributes selected as the focus of the test. The tasks should not last more than an hour to keep the test users focused. However, the coverage of the tasks should be wide enough, and the variety of the tasks big enough to give the users a chance to try the system out, and learn to control and use it. On the part of the number of test tasks, a meta-analysis by Lindgaard and Chattratichart (2007) revealed that the number of problems found was significantly correlated to the number of test tasks.

Additionally, Skov and Stage (2012) found the quality and relevance of the test tasks to considerably affect the number of problems found, especially when usability experts acted as evaluators. The tasks should, thus, be meaningful and also presented in a logical order (Hansen 1991). They should have a clear and unambiguous goal but no instructions on how to complete them with the system. Therefore, the wording should avoid terms from the system, so that it does not give distinct hints for the required actions. If it is possible, the participants should get an easy task first to get familiar with the test settings, and to get an early experience of success (J. Nielsen 1993, p. 184).

The tasks should be independent from each other, and handed out one at a time to the participants, so that some tasks could be skipped if the time agreed with the user is running out. Scenarios can be used to combine the tasks and to give the participants a role they can relate to. In the context of usability testing, *scenario* means "*a personalised, fictional story with characters, events, products and environments*" (Preece *et al.* 1994, p. 462). These scenarios are especially important for people from Chinese culture, as they pay more attention to contextual information compared to Western users, and thereby may find isolated tasks as artificial and hard to understand (Clemmensen *et al.* 2009). Scenarios also help to create the organisational context for the test, for example a specific shift in nurses' work. Similarly to the test tasks, also the test data should be as realistic as possible to give the test users an opportunity to assess the utility of the system in addition to its usability. (Bødker & Madsen 1998)

3.1.8 Questionnaires and interviews

The test users may be quite nervous at the beginning of a test, so it is recommended to give them a chance to relax and get acquainted with the situation by asking them to fill in a brief questionnaire concerning their background information as well as expectations and attitudes toward the system. After the test, both questionnaires and interviews are often used to get feedback and comments from the users.

There are several established questionnaires available on usability. System Usability Scale (SUS), for example, was made freely available in 1986, and has thereafter become almost a *de facto* standard in usability evaluation (Brooke

2013). Other potential standardised questionnaires include Software Usability Measurement Inventory (SUMI)[19], Website Analysis and MeasureMent Inventory (WAMMI)[20] and NASA Task Load Index (NASA-TLX)[21]. It is recommended to use these standardised questionnaires instead of designing own questionnaires due to their better reliability (Hornbæk & Law 2007).

The SUS questionnaire consists of ten statements, and the results are presented as a single number ”*representing a composite measure of the overall usability of the system being studied*” (Brooke 1996). This score ranges from 0 to 100 making 50 points the theoretical mean score, but the actual realised average score is close to 70 (Bangor, Kortum & Miller 2008). Brooke (1996) instructs not to use the scores of individual statements alone as they are not meaningful on their own. Even so, the factor analyses by J.R. Lewis and Sauro (2009) revealed that the scores of statements 4 and 10 together could be used to assess the learnability of a system.

The SUS questionnaire consists of five positive and five negative statements (Brooke 1996). A mixture of positive and negative statements is generally used to balance various biases, such as respondents' tendency to agree with most of the statements, and to make respondents think about each statement before answering. However, negative statements make mistakes more likely both in answering the questions and in interpreting the results, so Sauro and J.R. Lewis (2011) tried out an all positive version of the SUS questionnaire. As their results proved to be almost identical with the original one, they recommend to use this positive version especially in evaluations without moderators when it is hard to check the answers. The original and modified versions of SUS are presented in Table 16.

Table 16: Original and modified versions of the SUS questionnaire.

Original SUS questionnaire (Brooke 1996)	Modified all positive version (Sauro & J.R. Lewis 2011)
1. I think that I would like to use this system frequently	1. I think that I would like to use the website frequently
2. I found the system unnecessarily complex	2. I found the website to be simple
3. I thought the system was easy to use	3. I thought the website was easy to use
4. I think that I would need the support of a technical person to be able to use this system	4. I think that I could use this website without the support of a technical person
5. I found the various functions in this system were well integrated	5. I found the various functions in the website were well integrated
6. I thought there was too much inconsistency in this system	6. I thought there was a lot of consistency in the website
7. I would imagine that most people would learn to use this system very quickly	7. I would imagine that most people would learn to use the website very quickly
8. I found the system very cumbersome to use	8. I found the system very intuitive
9. I felt very confident using the system	9. I felt very confident using the website
10. I needed to learn a lot of things before I could get going with this system	10. I could use the website without having to learn anything new

In the positive version, the word ”system” has been replaced with ”website”. Also the other terms have been assessed in several studies, and the word ”cumbersome” in statement 8 has proven to be unclear to some users. There-

[19] http://sumi.ucc.ie/

[20] http://www.wammi.com/

[21] http://humansystems.arc.nasa.gov/groups/TLX/index.html

fore, Bangor *et al.* (2008) has replaced it with a more common word ”awkward”, whereas Finstad (2006) recommends to use both words “cumbersome” and “awkward” in parallel. In the positive version, word “intuitive” is used.

Considering interaction with the test users in general, Barnum (2011, p. 208) reminds that test users are eager to please. She refers to the studies by Reeves and Nass (1996) in which test participants' post-test ratings were significantly higher if they used the same computer for performing the tasks and answering the questions than if they switched the computer. To ask potentially sensitive questions, Nielsen (1993, p. 213-214) recommends to use computers rather than in person to minimise the bias. He also recommends to use the value of 3,6 as an estimate for neutral mean instead of 3 when using a scale from 1 to 5 (p. 37). Furthermore, the test users' subjective quality judgements in post-test questionnaires do not necessarily reflect the whole test but only its most recent incidents (Hassenzahl & Sandweg 2004). Therefore, it is recommend to use task specific post-task questionnaires if task-relevant information is required (Sauro & J.R. Lewis 2009).

3.1.9 Conducting tests

A pilot test is needed to check the test tasks, instructions, equipment and placements before the actual tests. The pilot user does not have to be from the target group, but someone outside the test team. Dumas and Redish (1993, p. 268) recommend to make the pilot test two days before the actual tests so that the preparations are finished but the test team still has enough time to make changes if needed.

Gomoll (1990) gives ten steps for conducting a usability test:

1. Introduce yourself,
2. Describe the purpose of the observation in general terms,
3. Tell the participants that they may quit at any time, and still get the fee,
4. Explain the purpose of the equipment in the room,
5. Explain how to think aloud, and give an example if desired,
6. Explain that you cannot provide help during the test,
7. Describe the tasks and introduce the product,
8. Ask if the user has any questions, and then begin the observation,
9. Conclude the observation, and
10. Use the results.

Usability tests are typically conducted to make the products less stressful to use, but the testing process itself can still be very stressful for the test participants (Schrier 1992). Therefore, several ethical considerations and recommendations have been listed to reduce this stress, for example by Schrier (1992) and Nielsen (1993, p. 184). The main ethical considerations before the test session include that everything is ready before the participant arrives; the participant is informed about the state of the system and of the confidentiality of the results; and the participant knows that it is the product that is tested, not the user (J. Nielsen 1993, p. 184). The number of observers should be limited into two or three if the observers are in the same room with the participant, and friends or relatives as test participants should be avoided so that the

participants are "*less likely to feel embarrassed if they have difficulty using a product*", and feel more relaxed to state negative comments on the product (Schrier 1992). Sonderegger and Sauer (2009) even recommend to avoid all non-interactive observers in the test room, as they have a significant effect in several usability measures, such as the stress level and the task times.

To give the test users an early feeling of success, the first task should be an easy one (J. Nielsen 1993, p. 184). The moderator should explain and hand out the tasks one at a time so that the users do not get overwhelmed of the task list. This way, the moderator may also skip some pre-selected tasks if time is running out. In no circumstances, the moderator may indicate that the participant is making mistakes or proceeds too slowly. Instead, the moderator should "*maintain a positive attitude throughout the entire test session, no matter what happens*" (Schrier 1992).

The participants should indicate themselves, when they have completed a task, because with some systems, it may be possible to do things by mistake without noticing, or not to notice that the system has actually completed the task in hand. After the test, the moderator should thank the participants for their help, and remind that they stay anonymous in the results. The recordings are presented outside the testing team only if the participants give a permission for this (J. Nielsen 1993 p. 184).

3.1.10 Analysis of test sessions

The goal of the test analysis is to structure the findings as interpretations and descriptions of user experience (Whiteside *et al.* 1988). The analysis gives interpretations of what happened in the test sessions, and what problems and successes emerged. The analysis should thereby lead to recommendations for improving the usability of the product.

Dumas and Redish (1993, p. 309) list five activities for analysing data from usability tests focusing on quantitative results:

1. Classify and summarise quantitative data,
2. Search for trends and surprises in data,
3. Examine the data for usability problems,
4. Use statistics, and
5. Organise problems by scope and severity.

The data analysis begins with describing the characteristics of the data (Dumas & Redish 1993, p. 318). The analyser computes measures, such as the number of errors occurring in a task, or average times for performing a task. The statistical analysis of the test results is quite often neglected, as Sauro (2004) points out. Therefore, confidence intervals, minimum and maximum values, standard deviations, *p*-values, power and other statistics should be used to put a number on the confidence statements (Sauro 2004, 2006).

The problems found should be organised by their importance, *i.e.*, scope and severity (Dumas & Redish 1993, p. 322). The *scope* of a problem refers to the locality of the problem, *i.e.*, how widespread the problem is. The *severity* of a usability problem, on the other hand, refers to the frequency with which the problem occurs; the impact of the problem when it occurs; and the persistence

of the problem (J. Nielsen 1994b). Several scales are available to rate these problems. For example, Dumas and Redish (1993, p. 324) give a four level scale with a clear reference to the impact on users' tasks:

- Level 1 problems prevent users from completing a task,
- Level 2 problems significantly slow down the users' performance and frustrate them,
- Level 3 problems have a minor effect on usability, and
- Level 4 problems point to potential upgrades in the future.

Instead of traditional video data analyses, Kjeldskov, Skov and Stage (2004) suggest using Instant Data Analysis[22]. This method utilises the resources already used in testing, *i.e.*, the moderator and a note taker, and requires only one more hour after the test sessions to make a summary of the findings with the help of a specific facilitator. The Instant Data Analysis was able to identify 85% of the critical usability problems in only 10% of the time required for the video data analyses made for comparison. (Kjeldskov *et al.* 2004)

To support the matching of the detected usability problems as well as their classification and analysis, various frameworks for usability problem reports have been developed. For example, Lavery, Cockton and Atkinson (1997) propose a structured problem report including a description of the context of the problem, its cause, potential breakdown and its outcome, as well as a recommendation for a redesign. The Usability Problem Taxonomy of Keenan *et al.* (1999), on the other hand, divides the description of usability problems into two major components, namely artefact and task component. These components are further divided into more specific subcategories, such as: visualness, language and manipulation (Keenan *et al.* 1999).

Despite these frameworks, the analyses in practice are mainly informal and rely on professional experience (Følstad, Law & Hornbæk 2010, 2012). The generation of recommendations for redesign is also unsupported. However, to improve the reliability of the results, Følstad *et al.* (2012) recommend to explore the analyses in groups after individual analyses.

3.1.11 Communicating results

The results of the analyses should be communicated to the development team along with the redesign proposals. The International Standard 13407 (1999) gives an example structure for a usability evaluation report in formative testing, and ISO/IEC 25062 (2006) defines a Common Industry Format dedicated to summative testing. These report formats have much in common, but the part of recommendations, for example, is essential in formative testing, but not as relevant in summative testing. The main structures of these report formats are presented in Table 17. Also Theofanos and Quesenbery (2005) give quite a detailed format for formative test reporting gathered from several practitioners.

[22] The term rapid reflection is also used for the method in a later article by Kjeldskov *et al.* (2005a).

Table 17: Contents of usability test reports as instructed by ISO 13407 (1999) and ISO/IEC 25062 (2006).

ISO 13407 (Annex B): Formative test	ISO/IEC 25062: Summative test
Executive summary	Title page
Product evaluated	Executive summary
Objectives of evaluation	Introduction
Context of use	Full product description
Measurement plan	Test objectives
Users	Method
Methods	Participants
Sequence	Context of product use in the test
Results	Tasks
General	Test facility
Video analysis	Participant's computing environment
User interface design	Test administrator tools
Workflow and process	Experimental design
Training	Procedure
User debriefing	Participant general instructions
User perception questionnaires	Participant task instructions
Recommendations	Usability metrics
Appendices	Effectiveness
	Efficiency
	Satisfaction
	Results
	Data analysis
	Presentation of the results
	Performance results
	Satisfaction results
	Appendices

Dumas, Molich and Jeffries (2004) give four general guidelines for usability test reports: emphasise the positive comments; express negative findings tactfully; avoid usability jargon; and be as specific as you can when describing the problems and presenting redesigns. Usability test reports tend to emphasise the negative sides of the system, since the evaluations try to reveal as many problems as possible. However, the positive comments and successful implementations should also be reported so that the designers know what to preserve and what to modify. Therefore, Molich recommends to present at least one positive finding for each three problems (Redish *et al.* 2002), and to keep the total number of usability problems in a manageable scale, such as 15-60 problems, focusing only on the most important ones (Molich *et al.* 2004).

In addition to presenting the problems, reasons for the problems should be indicated (Høegh *et al.* 2006), as well as recommendations for fixing these problems (Dumas & Redish 1993, p. 332). The studies by Hornbæk and Frøkjær (2005) show that the developers prefer having both the problem descriptions and the redesign proposals to get a good understanding of the problems as well as ideas for fixing them. Also multimedia reports and annotated screen dumps have been valued by the developers (Nørgaard & Hornbæk 2009).

If possible, the test team should evaluate the usability of the recommended changes with a few users to minimise the risk of introducing new usability problems (Dumas & Redish 1993, p 338). Keeping the business goals in mind, and reflecting the proposed changes to these goals have also improved the utility of the test reports, so that the developers have rated these reports significantly higher than reports without such business connections (Hornbæk & Frøkjær 2008a).[23] Similarly to the usability problems, also the redesign propo-

[23] Impact on development process discussed also in Chapter 3.6.

sals have to be justified and presented clearly. The problems and solutions should still be presented separately, so that the developer can easily consider also other solutions for the problems. (Jeffries 1994)

On the part of communicating the results to the development team, Dumas and Redish (1993, p. 341) recommend to communicate the results in multiple ways: in written report; verbally in a meeting with the developers; and visually with video clips from the usability tests. Also Ehrlich, M.B. Butler and Pernice (1994) praise for the videotapes, and having the developers attending the usability test sessions in the first place.

3.2 Thinking aloud method

Thinking aloud is one of the most direct methods to gain information about participants' internal states (Ericsson & Simon 1980). Thereby, thinking aloud is applied in several research areas, and is one of the central methods also in usability testing (Dumas & Redish 1993 p. 278; J. Nielsen 1993 p. 195; Rubin & Chisnell 2008 p. 204). It has been used for a long time in psychological studies as a method to study the cognitive processes needed in problem solving. John Watson (1920) was the first to report on using thinking aloud as he tried to learn more about the psychology of thinking (Fox, Ericsson & Best 2011). Karl Duncker (1945; original German version 1935) then was among the first researchers to utilise thinking aloud in empirical studies of mathematical problem solving in 1925-40. Duncker made a clear distinction between thinking aloud and introspection: in thinking aloud, the subjects "*allow their activity to become verbal*" instead of explaining the reasons for their actions. He encouraged the subjects "*not to leave unspoken even the most fleeting or foolish idea*" to make the subjects verbalise all their ideas. (Duncker 1945)

Ericsson and Simon (1980) categorise verbal reports by three criteria: time of verbal reporting; level of thinking aloud; and form of probing. The time is very important in verbal reports, because it is the working memory contents that are desired (J. Karat 1997). The participants may verbalise the information at the same time as they attend to it, *i.e.*, verbalise concurrently, or describe their cognitive processes afterwards, *i.e.*, verbalise retrospectively. Retrospective reports are viewed as less useful and less reliable, since they rely on participant's memory of what he has been thinking some time ago. (Ericsson & Simon 1980) Even so, comparisons of concurrent and retrospective reports have shown that retrospective reports give more explanations for the reasons of the actions and recommendations for improvement, whereas concurrent reports give procedural information of the actions participants are taking with the elements of the user interface (*e.g.* Bowers & Snyder 1990).

For the level of thinking aloud, Ericsson and Simon (1980) name three levels according to the need of processing of thoughts:

1. When the information is told in the same way as it is processed in the short-term memory[24], it is called direct, *i.e.*, level 1 verbalisation.
2. In level 2 verbalisation, the original information is not in verbal form but, for example an image, and it has to be translated into verbal form.
3. In level 3 verbalisation, the subjects are asked to do something more than just to tell their thoughts aloud. The subject may be asked to filter or select information according to given instructions, or to generate new principles or strategies. If the subjects are asked to describe their motor activities or routine actions that they would not otherwise pay attention to, the verbalisation falls into level 3.

If the test participants use level 1 or 2 verbalisation, Ericsson and Simon claim that the cognitive processes remain the same as if the participants acted silently. However, level 2 verbalisation may slow down the performance. In level 3 verbalisation, the subjects may alter their normal behaviour, and pay more attention to information that can help them to be more efficient in the following tasks. (Ericsson & Simon 1980)

The third criteria for categorising verbal reports is the form of probing. Ericsson and Simon (1980) name three categories also for this criteria. In the first category, the subjects are asked to articulate their thoughts at the same time they are processing the information. In the second category, the subjects may perform their tasks silently, but the experimenter probes *concurrently* for some specific information avoiding leading questions. In the third category, the information is asked only after the subject has completed the task, *i.e.*, *retrospectively*. In addition to these, there is also *interpretive probing* that takes place only after the subjects have performed a number of tasks. For this last version, Ericsson and Simon are quite sceptic on the quality and accuracy of the subjects' memories of their cognitive processes after a set of tasks. (Ericsson & Simon 1980)

In addition to the time and level of verbalisation and the type of probing, Ericsson and Simon (1984) bring forth the distinction between cases where the subjects are verbalising their thoughts to themselves and cases where the verbalisations are directed to someone else. Werner and Kaplan (1963) call these cases as inner speech and external speech. In their study, one sentence long written descriptions are compared with two different recipients: the instructions are directed to either the writer himself or to "any other person". The study is a between-subject study, and the order of the settings remain the same. First, the participants write the descriptions for themselves, then they practice explaining visual and oral stimulus to another person, and after this

[24] The terms short-term memory and working memory are used variously in HCI literature. Dix *et al.* (2004, p. 29) and Benyon, Turner and Turner (2005, p. 104) use the terms as synonyms, whereas Preece *et al.* 1994 (p. 65) claim that the term working memory has replaced the term short-term memory. Shneiderman (1998, p.355), on his part, says that *"people use short-term memory in conjunction with working memory*" indicating them to be separate entities. Then, in a textbook on cognitive science by Eysenck and Keane (1990, pp. 134-148), the terms are presented to origin from competing theories of memory structure, and also Alan Baddeley (1992) uses both these terms in his article "Working Memory" differentiating the concepts. Therefore, this thesis uses the same terms as the original papers.

rehearsal, they write down the one sentence long descriptions to someone else. The results show that the inner speech is more compact having significantly less words than the external speech especially concerning the visual stimulus. The external speech also uses more common referents than the inner speech. (Werner & Kaplan 1963)

3.2.1 Effects on performance in problem solving

The effects of thinking aloud in problem solving, learning and performance in succeeding tasks have been studied in several experiments, such as those by Gagné and Smith (1962); Davis *et al.* (1968); and Stinessen (1985). Gagné and Smith (1962) present an experiment in problem solving using the Tower of Hanoi as the problem. Half of the participants are asked to "*state verbally a reason for every move at the time they made it*", and the other half to "*search for a general principle which could be stated verbally after the tasks were solved*". At first, the participants verbalising their reasons are slower than the silent ones, but as they proceed to the more demanding problems that everyone performs silently, they perform much better than the silent ones, and make profoundly less excess moves. Searching for general strategies, on the other hand, does not have a significant effect on the succeeding performance. As a conclusion, the researchers summarise their experience of verbal reporting and transfer affect by stating: "*It would appear that requiring verbalization somehow "forced the S[ubject]s to think*"." (Gagné & Smith 1962)

With a very similar settings, James Davis *et al.* (1968) continue these studies on problem solving. They keep the variable of searching for a general strategy similar to the previous study, but divide the variable of thinking aloud into two variables: thinking aloud and experimenter presence. Even with the participants working alone, the experimenter instructs the participants and helps them to familiarise themselves with the test situation before she leaves the room. As the experimenter leaves, she tells that she would be in the next room and observe the subjects through a one-way-mirror. (Davis *et al.* 1968)

The results of Davis *et al.* (1968) are somewhat conflicting: the results with a five-tile Tower of Hanoi show that the presence of an experimenter significantly facilitates the participants' performance, but with the final, silently solved six-tile problem, these differences are no longer detected. Instead, the ones that have thought aloud previously, perform significantly better, similarly to the results by Gagné and Smith (1962). In the five-tile problem, however, this effect of thinking aloud is not yet significant. The conclusions of the study state that "*the mere presence of others [in a test setting] is a complicated variable*", and that "*talking while working apparently improved performance potential*". (Davis *et al.* 1968)

Also the study by Stinessen (1985) supports the findings of verbalisation having a positive effect on performance when solving the Tower of Hanoi problem as compared to participants performing silently. As a possible reason for this effect, Stinessen suggests that the requirement to state reasons for every move disrupts the process that otherwise could be automated. Thus, he concludes that the models of problem solving that are based on verbal reports

do not always "*adequately represent problem solving process which normally will take place*". (Stinessen 1985)

Several studies have thereby shown that verbal reports may improve the participants' performance and even change the cognitive processes in problem solving. Furthermore, already Duncker (1945, p. 11) points out that the subjects do not report everything in their mind, as the mediating phases leading directly to the solution tend not to be verbalised. Also Maier (1931) makes similar observations, and states that the factor that sets off a complete solution is lost to consciousness without paying attention to it. Duncker (1945, p 11) summarises these remarks: "*A protocol is relatively reliable only for what it positively contains, but not for that which it omits.*"

Consequently, Nisbett and T.D.C. Wilson (1977) review a number of studies made on verbal reports and their validity, and also make their own experiments. The studies have a major theme that the test participants are "*sometimes (a) unaware of the existence of a stimulus that importantly influenced a response; (b) unaware of the existence of the response; and (c) unaware that the stimulus has affected the response*". Indeed, their own experiments show that although a specific stimuli has specific responses, the participants' verbal reports do not include the stimuli. Instead, the verbal reports include implicit causal theories and plausible, but non-influential factors that cause the response in the participants' opinion. Even when the experimenters present their theories and hypotheses, and also bring out the fact that the participants have behaved according to the hypotheses, the participants insist that the stimulus has no effect on their behaviour, although the hypotheses could apply to someone else. (Nisbett & T.D.C. Wilson 1977)

This criticism provoked Ericsson and Simon (1980) to make a review of several studies of verbal reports to show that verbal reports are a reliable source of data. Although they admit that the verbal information is not always complete, they still consider it adequate to give plausible guidelines on the cognitive processes, if the method is appropriately applied. According to their review, verbalising information affects cognitive processes only if the participants are asked to verbalise information that would not normally be in their short-term memory, such as with automated actions, and with actions under a heavy cognitive load when the contents of the short-term memory is overflowing. (Ericsson & Simon 1980)

Russo, Johnson and Stephens (1989) try to detect possible conflicts in the thinking aloud model by Ericsson and Simon, and to find factors that could damage the validity of verbal protocols. They name two forms of invalidity: reactivity and nonveridicality. Reactivity means that the protocol changes the primary process, and nonveridicality means that the protocol does not accurately reflect the underlying primary process. If a protocol is reactive, veridicality is of little interest, so Russo *et al.* concentrate in reactivity in their studies, and compare four different verbalisation protocols: a silent performance; concurrent thinking aloud; a retrospective condition with a replay of eye fixations; and a retrospective thinking aloud based on the given problem or on the problem solutions right after each task. The results of the experiment show

significant reactivity in some tasks on the part of the accuracy of performance. For example, the concurrent thinking aloud significantly improves the accuracy of multiplying decimal numbers, and significantly decreases the accuracy of summing up three three-digit numbers. As potential causes for this reactivity, Russo *et al.* name increased workload due to verbalising; auditory feedback; enhanced learning due to repetition; and a motivational shift due to monitoring. They also suggest that even retrospective protocol may increase the memory load by pressuring the subjects to remember something worth explaining. (Russo *et al.* 1989)

Studies comparing eye-movement with thinking aloud data by Rhenius and Deffner (1990), on their part, support the verbal reports to describe the concurrent thoughts, making concurrent verbalisation "*a viable tool in the study of cognitive processes*". Their results show that thinking aloud slows down the participants' performance, but the accuracy is similar, as well as the solution strategies after a slightly slower start (Rhenius & Deffner 1990). Also the studies by Youmans *et al.* (2013) and Pike *et al.* (2014) support these findings, as the only effect by concurrent thinking aloud in their studies is to slow down the performance either marginally (Youmans *et al.* 2013) or significantly (Pike *et al.* 2013).

In the study by Pike *et al.* (2013), measurements with functional Near-Infrared Spectroscopy and NASA-TLX questionnaire indicate that thinking aloud that includes active interventions from the moderator significantly prolong the task times. However, the performance accuracy remains the same in all the test conditions including silent performance; irrelevant verbalisation; classic thinking aloud without moderator interventions; and an invasive thinking aloud that allows the moderator to ask questions during the task performance. The results also show that verbalising irrelevant words, such as "*blah-blah*", significantly increases the mental effort, mental demands and even physical demands. The thinking aloud, on its part, slightly improves the users' performance: the mathematically high performing users are more accurate when using the classic thinking aloud, and the invasive thinking aloud suits better to the mathematically low performing users. (Pike *et al.* 2014)

3.2.2 Thinking aloud in usability testing

As usability testing started to emerge in the early 1980's, thinking aloud method soon became an essential part of usability evaluation method repertoire. Studies by C.H. Lewis and Mack (1982) and Mack, C.H. Lewis and Carroll (1983) were among the first ones in the field of human-computer interaction to utilise the thinking aloud method. The method was used to get insight into the users' mental processes as they learned to use new text processing systems. Studies, such as those by Jørgensen (1990) and P.C. Wright and Monk (1991a), thereafter, have shown that thinking-aloud is an effective method in user interface design in detecting usability problems, especially if the designers conduct the usability tests themselves and thereby get direct feedback from the users. The textbooks on usability evaluation in the early

1990's established thinking aloud method as a central part of the usability testing practice (*e.g.* Nielsen 1993; Dumas & Redish 1993; Rubin 1994).

The studies by Ericsson and Simon (1980, 1984) are sometimes cited as references for thinking aloud in usability testing (*e.g.* J. Nielsen 1993, p. 195; Preece *et al.* 1994, p. 621), but quite often, the method is introduced without any references (*e.g.* Tullis & Albert 2008, p. 57; Dumas & Loring 2008). Even in cases where Ericsson and Simon's work is cited, thinking aloud is rarely applied as instructed. For example, an explorative study by Nørgaard and Hornbæk (2006) shows that in practice, the test moderators ask more questions about hypothetical or expected problems instead of previously experienced problems, and the questions are sometimes abstract, leading or even impossible to answer. Still, the validity or reliability of the method has rarely been questioned, although studies already in early 1990's show remarkable effects of the thinking aloud in users' performance, such as in the study by R.B. Wright and Converse (1992). In this study, test users explaining their actions perform significantly better than the users performing silently if measured by task times and frequency of errors. Furthermore, the differences become even more significant as the test tasks become more complex. The studies by Trudel and Payne (1996), on their part, show that users are more effective in learning to use new systems if they stop to review their actions every now and then, indicating that even retrospective thinking aloud may bias the users' performance when done between the test tasks.

Despite the results of R.B. Wright and Converse (1992), concurrent thinking aloud soon became widely used, and studies of its effects remained quite rare until the article by Boren and Ramey (2000). These researchers from the field of technical communication criticise the ways usability practitioners apply thinking aloud method, and claim that if usability engineering wants to be a convincing, reliable and valid discipline, the practitioners need to be informed of the underlying theories, and proper ways to apply the methods. For example, they have observed quite miscellaneous instructions to think aloud, sometimes emphasising evaluation and explanation; too intrusive and directing prompts; and intervening questions when users are doing the test tasks. (Boren & Ramey 2000)

As the needs for thinking aloud in usability testing are quite different from the ones in cognitive psychology, Boren and Ramey (2000) suggest adjustments to the thinking aloud method so that it would better support the needs of product development in evaluating the product in hand instead of investigating the way of users' thinking. These adjustments fall into four categories: setting the stage; using the nature of speech to keep verbal reports natural; controlling interaction with test users; and eliciting additional information. Setting the stage includes instructions on the roles of the product, test user and test moderator. The product is the subject under evaluation; the test user is the work domain expert and primary speaker in the test; and the test moderator is the learner and primary listener. To keep the role of an active listener, and to keep up a natural atmosphere for talking, the moderator needs to use neutral and simple acknowledgement tokens, such as "*OK, yeah*" or "*mm hm*".

Even when the test user falls silent, simple *"Mm hm?"* or *"And now…?"* is recommended for the moderator to keep the role of a learner instead of a controller or a commander. To support the test user's role as a domain expert and a primary speaker, probes for more information should be left till the user has completed the task or even after all the tasks. (Boren & Ramey 2000)

Consequently, several studies of the effects of thinking aloud have been made in usability testing. These studies include comparisons between classic concurrent thinking aloud, retrospective thinking aloud, working silently and interactive thinking aloud. Also eye-tracking data has been used in these studies to validate concurrent thinking aloud reports (*e.g.* Cooke 2010), or to stimulate retrospective verbal reports (*e.g.* Eger *et al.* 2007). The study by Lynne Cooke (2010), for example, shows that users' eye movements verify their concurrent verbal reports in 80% of their task performance time, but as much as 77% of their verbalisations are about easily observable actions, and only 5 % are explanations for the rationale or motivation for the actions. Thus, concurrent thinking aloud appears valid, but as a drawback, gives only limited supplemental information (Cooke 2010). However, in a replication study by Elling *et al.* (2012) using more complex tasks, only 40% of the verbalisations were easily observable. In addition, the verbalisations gave information that would have been hard or impossible to get even with eye-tracking, such as reasons for fixations and potential problems. (Elling *et al.* 2012)

3.2.3 Concurrent thinking aloud

Concurrent thinking aloud is the most popular approach for thinking aloud, as 89% of the survey respondents select it as their most frequently used approach over retrospective thinking aloud and paired-user testing (McDonald *et al.* 2012). Reasons for its popularity include its fastness, efficiency and ease for users to relate to. Users' immediate responses and the added interest for clients observing the sessions are also mentioned as benefits of the concurrent thinking aloud. As many as 68% of the 207 respondents use the general instruction by Ericsson and Simon (1984, p. 78) to simply verbalise their thoughts, but 28% instruct the users to focus on user experience. Only 16% of the respondents always use a practice session before the test tasks, and only 33% avoid interventions during thinking aloud. (McDonald *et al.* 2012)

On the part of the training sessions, Ericsson and Simon (1984, p. 82) point out that an "*essential merit of T[hinking]A[loud] as compared with introspection, however, is that the former is a normal mode of processing that does not call for the extensive training needed for the latter*". Thereby, even a small warm-up problem should be enough to conform to the model by Ericsson and Simon. However, the instructions to explain their feelings and experiences as well as using questions instead of neutral reminders may change the users' behaviour by promoting self observation or emphasising the role of an outside listener instead of talking to oneself (Ericsson & Simon 1984, p.83).

The studies of concurrent thinking aloud primarily fall into two categories: comparisons between concurrent thinking aloud and silent performance; and comparisons between various types of concurrent thinking aloud protocols.

The comparisons with silent performance are addressed first in this subchapter, and only after that, the comparisons between different thinking aloud protocols.

Hertzum *et al.* (2009), for example, compare classic and relaxed thinking aloud to silent performance. Relaxed thinking aloud includes level 3 verbalisations as test moderators ask for explanations and comments during the tasks, and it is the type of thinking aloud that is generally used in usability tests in practice (Boren & Ramey 2000; Nørgaard and Hornbæk 2006; Shi 2008). Also in the survey by McDonald *et al.* (2012), 28% of the respondents use instructions focusing on user experience instead of the general instructions for the users to just verbalise their thoughts. The results of the study by Hertzum *et al.* (2009) show that the classic thinking aloud has very little effect on task performance except for slowing it down, as the review by Ericsson and Simon (1980) suggests. The effect of prolonging is stronger in tasks requiring assessment than in tasks searching for facts with both versions of thinking aloud. The accuracy of the solutions, however, is not affected by either version. Even so, the relaxed thinking aloud changes the participants' behaviour to browse and navigate more within and between the web pages. The perceived mental workload is also significantly higher in both thinking aloud conditions compared to silent performance. Thereby, the results confirm that classic thinking aloud obtains valid data about the use of the evaluated systems as long as the interaction between the test user and the test moderator is kept minimal. Relaxed thinking aloud, on the other hand, may not be a valid method for gathering data about users' performance, as the method itself may be reactive and change the participants' behaviour. (Hertzum *et al.* 2009)

Olmsted-Hawala *et al.* (2010a, 2010b) then, compare three different thinking aloud protocols and use a silent condition as a control condition. The thinking aloud protocols include a classis thinking aloud; a speech communication related protocol; and a coaching protocol having similarities with the active intervention presented by Dumas and Redish (1993, p.31), and the relaxed thinking aloud used by Hertzum *et al.* (2009). The study is a between-subject study with 80 participants, and 4 outside moderators that each conduct 1 test condition without knowledge of the true goals of the study. After a practicing task, the participants are left alone in the test room, and communications are carried out via microphone and speakers from the control room. In this study, the number of usability problems are not counted due to their subjective nature and potential evaluator effects. Instead, the accuracy in terms of success or failure on the tasks, the efficiency as a task completion time, and the subjective satisfaction scores are analysed (2010a), as well as the numbers of verbalised and non-verbalised counts of frustration and positive comments (2010b). The results of the first analyses (2010a) show that the accuracy is significantly better in the coaching protocol, as 60% of the tasks are completed accurately in this condition compared to the 30-40% in the other conditions. The coaching protocol also affects the satisfaction scores, as participants give more positive scores in this condition compared to the other thinking aloud protocols. However, when all the four conditions are compared, there are no

significant differences in the satisfaction scores. Also the task completion times have no significant differences among the test conditions even when compared to the silent performance. (Olmsted-Hawala *et al.* 2010a) The second analyses reveal no statistical differences in the three thinking aloud protocols concerning verbalised or non-verbalised counts of frustration or positive comments as the verbalised utterances are analysed and counted. (Olmsted-Hawala *et al.* 2010b).

McDonald and Petrie (2013) repeat a very similar study comparing classic thinking aloud and explicit thinking aloud to working in silence. Their instructions in the explicit thinking aloud emphasise the desire of getting information about the things that the users' like or dislike, or find confusing. During the sessions, "Keep talking" is the only reminder used if the users fall silent. Again, the classic version has no impact on task performance, but the explicit instruction increase scrolling and navigation within and between pages, similarly to the study by Hertzum *et al.* (2009). As the users assess their workload, both thinking aloud conditions raise their assessments on effort and frustration, and the explicit thinking aloud also increases the workload on mental demand and performance. (McDonald & Petrie 2013)

Also the meta-analysis by Fox *et al.* (2011) points out that mere verbalisation of thoughts on levels 1 and 2 is nonreactive, and only slows down the performance. However, enforcing participants to explain their thoughts and actions may change their performance. Therefore, all interactions between participants and moderators, and other differences between the verbalising condition and the silent performance should be reported to give the reader a possibility to assess the validity and reliability of the test results. (Fox *et al.* (2011)

The experiment by Hertzum and Holmegaard (2013) supports the finding that thinking aloud in levels 1 and 2 does not affect the task solution rate or the task completion times. However, when interruptions are integrated with thinking aloud, they have a significant interaction that affects both the task solution rate and the task times. For example, visual interruptions with thinking aloud significantly improve the task solution rates compared to the silent control group. (Hertzum & Holmegaard 2013)

Comparisons between various versions of thinking aloud then, have been conducted for example in studies by Krahmer and Ummelen (2004); Zhao and McDonald (2010); McDonald, McGarry and Willis (2013a); and Zhao, McDonald and Edwards (2014). The study by Krahmer and Ummelen (2004) compares classic thinking aloud and a speech communication related interactive version with brief acknowledgements, as proposed by Boren and Ramey (2000). The results of the study show several differences between the thinking aloud conditions. For example, participants using classic thinking aloud are significantly more lost in the web service visiting more nodes, and they also have much more difficulties in completing the tasks compared to the participants with the more interactive thinking aloud. However, neither the number of uttered words differs significantly between the test conditions, nor the number of detected usability problems. Based on these results, the more interactive thinking aloud may cause validity problems in usability testing, espe-

cially if task performance and feeling of being lost are of interest, so the classic thinking aloud is a more reliable source of data. (Krahmer & Ummelen 2004)

The study by McDonald *et al.* (2013a) supports these findings with more difficult tasks, but with simple tasks, their study finds no differences between the two thinking aloud versions. In this study, the classic concurrent thinking aloud is compared with thinking aloud in which test users are explicitly instructed to explain their choices of navigation when performing the test tasks. In addition to these two thinking aloud versions, the study includes two types of tasks: simple tasks requiring the use of menu elements with fully descriptive labels; and more difficult tasks requiring the use of menu elements having ambiguous labels or offering several likely alternatives. Only the more difficult tasks cause differences in the results, as the performance times are longer; task accuracy is improved; and less link traversals are made when using the explicit thinking aloud compared to the classic thinking aloud. Although the explicit instructions also yield more explanations potentially helping in identifying and fixing usability problems, the risk of reactivity with this explicit thinking aloud is considered too considerable compared to the potential advantages. (McDonald *et al.* 2013a)

In the study by Zhao and McDonald (2010), the classic thinking aloud is compared with a more relaxed thinking aloud, similarly to the study by Hertzum *et al.* (2009). The relaxed thinking aloud allows several intervention types and triggers for the moderator to get more information about the users' thoughts and experiences. The results show that the relaxed thinking aloud leads to more utterances that help in analysing the causes of the usability problems. However, the proportion of relevant utterances containing information about the problems or their causes is very low for both versions. Still, most of the test users (17 out of 20) prefer the more interactive thinking aloud because it makes them feel more natural and relaxed. Furthermore, the users provide also level 3 verbalisations and comments that are clearly aimed to someone else but themselves also in the classic condition regardless of the order of the test conditions. The researchers suggest that it is the social context of usability testing that overrides the instructions as the test participants want to be understood and want to help in improving the system. Therefore, they recommend to search for ways of maximising the number of useful utterances without moderator interventions, as the relaxed thinking aloud distracts some of the users' performance. (Zhao & McDonald 2010)

3.2.4 Retrospective thinking aloud

The effect of the time for verbal reports in the context of usability testing has also been studied in several experiments. For example, Ohnemus and Biers (1993) compare the test participants' performance and subjective ratings in three test conditions: concurrent thinking aloud; retrospective reports right after the test; and retrospective reports on the following day. The results show no significant difference between the groups in the task completion times, task success or subjective ratings of the system. Even the total time spent on verbalising is statistically same for all the test conditions. However, retrospective

groups make fewer verbalisations with longer durations, and these verbalisations have more value to the designers than the more frequent utterances in the concurrent group. As a possible reason for the lower value of the utterances in the concurrent group, the researchers suggest that the workload needed to perform the tasks and to think aloud simultaneously prevents the participants from utilising long-term memory and producing high value verbal reports. As the participants in the retrospective groups can concentrate on one task at a time, they are able to give evaluative comments and express uncertainty about their actions more frequently than those in concurrent conditions. Between the two versions of retrospective thinking aloud then, the quality or quantity of the verbalisations does not differ. As the retrospective thinking aloud requires almost twice as much time from a single test user, the usefulness of these retrospective verbalisations was also compared with the verbalisations of a pair reported by Hackman and Biers (1992). In this comparison, the retrospective thinking aloud has more utility than paired-user testing, and gives "*more valuable information to the designer despite the added time cost*". (Ohnemus & Biers 1993)

Van den Haak, de Jong and Schellens (2003) conduct similar studies 10 years later comparing concurrent thinking aloud and retrospective thinking aloud immediately after the test tasks. The results show no significant difference in the total number of problems found, but the problems are detected differently: the retrospective condition reveals more problems through verbalisation, whereas concurrent thinking aloud reveals more problems through observation. Additionally, the participants in concurrent condition make more errors in the tasks and are less successful in completing the test tasks. Even so, there is no significant difference in the types of problems; overall task completion times; times per task; ways of experiencing the use of the thinking aloud method; or estimations of own behaviour compared to normal situations. However, participants in the retrospective condition find the test situation to be more disturbing than the participants in concurrent condition, although all the measures regarding the participants' experiences are neutral or even positive in both conditions. Based on these results, "*concurrent and retrospective think-aloud protocols can be regarded as equivalent, but clearly different evaluation methods*". (van den Haak *et al.* 2003)

Also Guan *et al.* (2006) study the validity and reliability of retrospective thinking aloud. Instead of comparing the method to concurrent thinking aloud, they compare the verbal reports with eye movement data in a within-subject study. The validity of the retrospective reports is assessed as the extent to which the verbal reports on the tasks overlap with the objects and their order in the eye movement data. The results show that even 88% of the verbalisations correspond with the eye movement data. In correspondence to the silence or meaningless utterances in concurrent thinking aloud when facing difficulties, also the users' retrospective reports on these occasions are abstract and unclear. Overall, the results show the retrospective thinking aloud to be a valid and reliable method for gathering users' performance information in usability testing. (Guan *et al.* 2006)

The results of the study by Eger *et al.* (2007) are quite contradictory to the ones by Guan *et al.* (2007) on the part of the clarity and coherence of retrospective verbal reports. Eger *et al.* (2007) compare concurrent and retrospective thinking aloud as a within-subject study, and divide the retrospective setting into a between-subject study by the cue for the retrospective verbal reports. The cue is either a playback of the dynamic screen events or the eye-tracking data. The study includes only one test task searching for information with two online search engines, and it has 24 test participants. The independent variables include task completion time and rate; quantity and type of usability problems found; and subjective measures from post-test questionnaire including assessments on the working conditions in the test settings. The results show that the retrospective reports cued by eye-tracking data generate more usability problems than the concurrent thinking aloud method, specifically with the problems related to feedback and comprehension of the site. The screen events then, detect more problems related to the layout of the site when compared to the concurrent thinking aloud. The means of task completion times are slower in the concurrent thinking aloud condition, but no significant differences are detected. The completion rates, however, do have significant differences: all the participants in the retrospective setting with screen event cues complete all the tasks, and 60% with the eye-tracking data, but only 42 % in the concurrent thinking aloud condition. The users' assessments on the working conditions also reveal that the setting with concurrent thinking aloud is perceived as more unpleasant than the retrospective settings. The concurrent thinking aloud also slows down the users' performance in their opinion. In addition, the presence of a test moderator is perceived as more unpleasant and unnatural when the users are thinking aloud concurrently compared to the retrospective settings. This may be explained by the need of the moderator to probe the users when they fall silent in concurrent thinking aloud. As one more advantage for the retrospective setting, the verbal reports in the retrospective settings contain more coherent sentences than those in the concurrent settings. (Eger *et al.* 2007)

Although retrospective thinking aloud has several advantages compared to concurrent thinking aloud, it does have some problems as well, such as reliance on memory, and vulnerability to post-hoc rationalisations, bias and fabrications (Eger *et al.* 2007). To minimise these disadvantages and threats, Freeman (2011) suggests to use eye-tracking data as a cue for retrospective thinking aloud, and to instruct the users to work silently or to explain their reactions according to their own preferences, so that the moderator probes for further information only when something interesting pops up from the online eye-tracking data.

3.2.5 Effect of thinking aloud instructions

In addition to the time of thinking aloud, its level and type of probing, also the impact of the instructions given to the test users for the thinking aloud has been studied (*e.g.* Zhao, McDonald & Edwards 2014; McDonald & Petrie 2013). The study by Zhao *et al.* (2014) compares two instructions: the classic

instructions recommend the users to act as they were alone in the room talking to themselves, whereas the explicit instructions ask the users to verbalise their expectations, surprises, delights, confusions, frustrations and causes for these effects. Otherwise, the test conditions are kept similar, and the only interaction between the moderator and the test users during the tasks is minimised to the reminder: "*Keep talking*", when necessary. To avoid possible transfer effects, the study was made between the subjects. The dependent variables include task performance data; mental workload measured with NASA Task Load Index; test users' perceptions on their own behaviour; utterance data; and data on usability problems including their number, type, severity and source. (Zhao *et al.* 2014)

The results show that most of the dependent variables have no significant differences between the conditions. For example, the performance measures show no significant differences and neither does the number of utterances. However, the contents of the utterances differ, as the explicit instructions yielded significantly more utterances related to the users' expectations, explanations and positive comments compared to the participants with classic instructions. These utterances helped in finding more problems overall, but most of these extra findings were only of low severity. The test participants with explicit instructions reported significantly higher mental workload, and they also assessed their own behaviour as being significantly more focused on finding problems than the participants with classic instructions. Thereby, the researchers conclude that although the explicit instructions do not cause reactivity in task performance, they "*may lead users to be hypersensitive to interface issues in order to comply with the instruction*". (Zhao *et al.* 2014)

Unlike the study by Zhao *et al.* (2014), a similar study by McDonald and Petrie (2013) show some reactivity in the users' behaviour due to different thinking aloud instructions. In this study, the effects of classic and explicit instructions were compared with silent performance. The results show that the classic instructions do not change the users' behaviour, but do increase the users' ratings for effort and frustration. Explicit instructions, however, increase the users' within-page and between-page navigation and scrolling activity, and also yield higher ratings in the mental workload in the subscales of effort, frustration, mental demand and performance compared to silent performance.

3.3 Modifications of usability testing

Due to different goals and contexts, several alternative methods have been developed to get into the test users' thoughts and experiences in using the assessed systems. This subchapter presents some usability testing methods that have been developed to overcome various challenges with the traditional usability testing in laboratory settings.

3.3.1 Question asking protocol

Kato (1986) presents a question-asking protocol to offer a more natural way of behaving and reporting than thinking aloud in a usability test. A question asking session involves three persons: a user, a tutor, and an experimenter. The test user is encouraged to ask questions from the tutor sitting next to the user, and the test experimenter prompts both the test user and the tutor if needed. The list of these questions gives valuable data to the development team on parts to be clarified, and on instructions to be included in a manual. The protocol is well suited for testing new systems with novice users. The test tasks should be so challenging that the users most probably needed help from the tutor to finish the tasks. As weaknesses, the method is not appropriate in assessing manuals or in making quantitative analyses. (Kato 1986)

Also gradual disclosure builds on hints and prompts that the moderator gives. It is used to test the intuitiveness of the system in such an early phase of development when no support material for the system is yet available. (Rubin & Chisnell 2008, p. 304)

3.3.2 Cooperative evaluation

Cooperative evaluation is a method presented by P.C. Wright and Monk (1991b) to meet challenges of quickly proceeding product development. The method has three central characteristics:

1. The user performs predefined tasks focusing on features of interest and parts that the prototype supports,
2. The user thinks aloud while doing the tasks, and
3. The designer acts as a moderator in the test as she knows how the prototype works and its capabilities.

The participants think aloud while doing given tasks, and the moderator asks questions about the problems the participant faces. To equalise the situation, also the participant may ask questions from the moderator. (P.C. Wright & Monk 1991) In this respect, the method reminds the question asking protocol by Kato (1986).

In cooperative evaluation, the moderator is the designer of the system, because the method is applied in such an early stage of development that the design evolves rapidly, and it is not cost-effective to train an outsider to control the system. In addition, the designer gets instant feedback straight from the users for further development. Thereby, the method can be classified as a discount method suitable for formative evaluation. (P.C. Wright & Monk 1991)

The designers do not need to be experts in usability or human factors, but to follow brief instructions on what kind of questions to ask from the test users. The experiments revealed that the designers are better in detecting usability problems than their colleagues who have not been involved in the design. However, the designers are quite poor in predicting the problems in their own designs, so usability tests are needed to detect the real problems. (P.C. Wright & Monk 1991)

Also Høegh and Jensen (2008) studied the extent to which system developers and other persons participating in the development of a system could anticipate the usability problems of that specific system. The developers and other participants were asked to list all the problems they considered the system to have, and these lists were then compared to the problems revealed in usability tests. Although, some of the participants could anticipate almost all of the problems, only about third of the problems were correctly anticipated in average. The supporter and educator who teach the clients to use the evaluated system were better in anticipating the most severe problems than the developers. However, all the participants tended to underestimate the severity of the problems. (Høegh & Jensen 2008)

Skov and Stage (2012) and Bruun and Stage (2014), on their part, studied the possibilities to train system developers to plan, conduct, analyse and report usability tests themselves. Instead of recruiting actual system developers, 234 students on software development acted as test participants in the first study. The results of this study show that with the background of only one introductory course on usability issues the students could "*conduct usability evaluations and produce usability reports that were of a reasonable quality and with relevant results*". (Skov & Stage 2012)

In the latter study, Bruun and Stage (2014) gave eight software development practitioners a two-day course on usability evaluation, after which the developers planned, conducted, and analysed a usability test on their own. The results show that the developers were able to find as many problems as HCI specialists, but they assessed the severity of the problems lower than the HCI specialists. Even, the developers were able to make good relations to the test users in the test sessions, although they had problems in making the users to think aloud. (Bruun & Stage 2014)

3.3.3 Cooperative Usability Testing

Frøkjær and Hornbæk (2005) present a usability testing method with somewhat similar name as cooperative evaluation but with a different content. Their method, called Cooperative Usability Testing (CUT), consists of two parts. The first part is an interaction session in which the test users interact with the system in a similar way as in Contextual Inquiry or in a thinking aloud test. The second part then, consists of an interpretation session in which the test users and the evaluators cooperatively discuss on the problems that the users have faced in the first part. The recordings are used to help the users remember the incidents. Two evaluators are recommended in the sessions: in the first part, one takes notes as the other instructs the user, and in the second part, the evaluators switch their roles, so that the former note taker leads the conversation based on her notes. (Frøkjær & Hornbæk 2005)

In their studies, Frøkjær and Hornbæk (2005) found out that the test users rather reflect and comment their actions even in extensive debriefings than participate a traditional thinking aloud test. Also the evaluators value the interpretation sessions, as the sessions help in clarifying and understanding the most important usability problems. On the part of the analysis, the method

includes a risk of introducing new and potentially problematic interpretations, since the interpretations need to be done quickly in the first part, and the interpretation discussions are limited to 45 minutes. (Frøkjær & Hornbæk 2005)

Later on, Følstad and Hornbæk (2010) studied the use of Cooperative Usability Testing in the development of two work-domain specific systems. They modified the method by including an interpretation session after each task to get the users' immediate interpretations of the tasks, and to minimise the risk of users rationalising their behaviour. In addition, they used task-scenario walkthroughs as the basis for discussions, meaning that they just reminded the users about the route they had used to get through the task, instead of looking at the video recordings. This modification brought more flexibility to the discussions, and enabled comments on parts that were not used in the test. These interpretation sessions also offered a good opportunity to broaden the scope of the discussions to cover the users' needs and requirements as well as new design proposals along with the interaction in the tasks. (Følstad & Hornbæk 2010)

A very similar approach to use the tasks as prompts in retrospective reports is also used in Dual Verbal Elicitation by McDonald, Zhao and Edwards (2013b). McDonald *et al.* had ten participants performing four different kind of tasks with an intranet website while thinking aloud. After all the tasks, the users had access to the interface and the test tasks, and were asked to give a retrospective report on the test tasks. The moderator did not ask any specific questions, but only used acknowledgement tokens, such as "*mm-hmm*", to avoid bias in users' reports. The two thinking aloud reports complemented each other, as the concurrent reports indicated when users deviated from the correct task solutions, and the retrospective reports gave insight to the users' experiences, causes of problems and the context in which they normally use the system. As the retrospective thinking aloud did not add much extra time to the testing session when based on the test tasks instead of playing the test videos, the authors recommend to use both of these methods in a single test to get more causal explanations, problem indications and comments on user experience. (McDonald *et al.* 2013b)

3.3.4 Critical incidents and backtracking analysis

To enable reporting of problems in real use situations, Hartson and Castillo (1998) developed an instrumented method in which users report critical incidents themselves. They integrated a critical incident reporting tool to the system to be evaluated, and instructed the users how to use it. The tool includes two parts: a part for textual and structured report of the incident; and a part for automatically recording the screen activities just before the reporting. In this set up, the users were competent in identifying critical incidents, but too often, initiated the reporting with such a delay that the automated recording did not include the critical activities causing the problem. As a major advantage to the method, though, it could be operated remotely, asynchronously and independently from the users. (Hartson & Castillo 1998)

As a complement to these self-reporting tools, Akers *et al.* (2009, 2012) developed a method based on automatically logged data and discussions on the data. This backtracking analysis gathers several users at the same time to do predefined tasks in the same location. During the task performance, the logging is triggered as the test participant uses the backtracking operations undo or erase. After the users have finished the tasks, they are paired up for retrospective discussions on these logged incidents. First, the other participant describes her logged incidents, and the other one asks the questions prompted by the analysis tool. The answers are audio recorded, and integrated to the logged data for further analysis. After the first one has gone through all her incidents, the participants switch roles. This way, the amount of material for further analysis is substantially smaller compared to traditional usability tests, as only the critical incidents are recorded and explained. (Akers *et al.* 2012)

The comparisons between self-reporting and automated logging show that both methods find practically the same amount of problems, but the problems are different, so it is not recommended to use backtracking analysis as the only method for evaluation (Akers *et al.* 2009, 2012). Furthermore, any usability evaluation method relying on retrospective discussions faces the risk of loosing the original emotions: the situations that caused apparent frustration in concurrent thinking aloud tests were easily just laughed at in the retrospective discussions, thereby fading the original reactions (Akers *et al.* 2012).

3.3.5 Experience Clip

The popularity of mobile applications has made it necessary to develop evaluation methods that can be used in the fields in authentic use situations, *i.e.*, outdoors with real users and without disturbing observers. For example, Isomursu, Kuutti and Väinämö (2004) first tried to use a researcher as a shadow following the users on the streets and video recording the use. Soon, they found the setting inefficient, as the users became uneasy, and also very silent. Therefore, they modified the setting, and invited pairs of users from the passers by to participate in their study. This method is called Experience Clip. (Isomursu *et al.* 2004)

In Experience Clip, the other participant gets the evaluated application, and the other one gets a mobile phone with video shooting capability. The one getting the mobile phone is instructed to take video clips as the other participant uses the application. When returning the equipment, the participants briefly describe what they have done with the application, and how the application has worked. (Isomursu *et al.* 2004)

The video clips reveal typical use patterns, but also expressions of emotions and usability issues. The users seem to enjoy observing each other, so the usage situations in the video clips appear natural. However, some of the situations appear to be performances created solely for the designers to make the participants' point clear, and to suggest better solutions. Nevertheless, having participants of equal status, instead of a researcher and an observed, makes it more natural for the users to try out their own ideas with the system, and to comment it. (Isomursu *et al.* 2004)

The participants in the Experience Clip have quite a lot of independence and power. First, they choose what to shoot, and then they also choose what to describe when returning the material. In Cooperative Usability Testing, the researchers make these selections, and in backtracking analysis, the automated tool makes these decisions triggered by predefined user actions.

3.4 Usability inspection methods

Having real users evaluating a system is essential. However, users are sometimes hard to find, and it takes time to organise tests, and to recruit representative users. Therefore, methods that can be applied without users are needed to complement the iterative design and evaluation process. These inspection methods give means to find usability problems easily, quickly and very early in the development process, even before any prototypes are prepared (*e.g.* Desurvire, Lawrence & Atwood 1991).

The study by Molich and Dumas (2008) shows that "*expert reviews with highly experienced practitioners can be quite valuable*", and can produce comparable results to usability tests. Also Brooks (1994) finds inspection methods appropriate if the goal is to find some of the major usability problems in the system before user testing or implementation, or to compare design alternatives. However, she recommends to involve users whenever decisions are made on factors that are critical to the product success in the market.

Usability inspection methods rely on the evaluators' experience and knowledge when trying to predict the usability problems. Different evaluators find different problems, so a greater number of evaluators working for a shorter time has proven to be more effective than a smaller group of evaluators working for the same total time (Virzi 1997). Nielsen (1993, pp. 156-157) recommends to use about five inspectors, and to let them work alone to ensure independent and unbiased results. Sawyer *et al.* (1996), on the other hand, prefer two inspectors looking at the interface together to bring forth more problems and to generate better recommendations for improvements than inspectors working alone. When work-domain experts and usability experts have been compared as evaluators, the domain experts have found less problems, but they have had a greater impact in the development as the developers have given higher priority to these problems compared to those found by the usability experts (Følstad 2007).

Heuristic evaluation (*e.g.* J. Nielsen & Molich 1990; J. Nielsen 1993; J. Nielsen 1994b) and cognitive walkthrough (*e.g.* C.H. Lewis *et al.* 1990; Polson *et al.* 1992; Wharton *et al.* 1994; C.H. Lewis & Wharton 1997) are widely known and generally used inspection methods. Cognitive walkthrough is a task-based method concentrating on evaluating the ease of learning and learning by exploration, whereas heuristic evaluation is based on evaluators' expertise and a set of general usability guidelines called heuristics. The validity of inspection methods that take no account of the task context has been questioned for example by Carroll (1997). Therefore, methods, such as heuristic walkthrough (Sears 1997), have been developed to combine characteristics

from heuristic evaluation and task-based walkthroughs. Also Nielsen (1995) points out that heuristic evaluations can benefit from the use of detailed descriptions of the interaction between the system and the user especially with highly domain-dependent systems.

Also other guidelines and criteria are available for usability inspections, such as the ergonomic criteria by Bastien and Scapin (1995), and the metaphors of human thinking by Frøkjær and Hornbæk (2008). A new inspection method called utility inspection method is also available for evaluating the utility, usefulness and acceptability of a system. With this method, usability experts have been able to find more problems related to the use context. They have also considered these problems as more severe and complex than problems found through heuristic evaluations. (Johannessen & Hornbæk 2014)

3.5 Criteria for assessing usability evaluation methods

Validity, thoroughness and reliability are essential criteria for usability inspection methods (Bastien & Scapin 1995), and the same criteria can be extended to other usability evaluation methods, as well. For example, Sears (1997) defines validity in the context of usability evaluation as the power to evaluate the intended properties, and ability to detect real usability problems that have an impact on users. When formulated as a measurable value, validity is the ratio of real problems from the total amount of identified problems. Thoroughness, on its part, refers to the scope of the evaluation so that the interface is assessed as widely as possible to detect as many problems as possible. Reliability then, refers to getting similar results under similar conditions. (Sears 1997)

John and Marks (1997) call the validity criteria as a *predictive power* referring to the possibility that the problems predicted through usability evaluation methods may not occur to the users. An optimal way to measure this accuracy of predictions would be to compare the results of the evaluations to the observations in real use, but this is rarely possible. Therefore, usability testing has often served as a benchmark for comparisons, and as a control for confirming the predicted problems, especially for the usability inspection methods (*e.g.* Cuomo & Bowen 1994; John & Marks 1997).

Since validity, thoroughness and reliability are challenging measures requiring at least estimates from the real use, other more simple metrics, such as the number of problems found, are often used to compare usability evaluation methods. Jeffries *et al.* (1991) made one of the first comparisons on usability evaluation methods and, thereby, created a baseline for many comparisons. This study compares expert evaluations, software guidelines reviews, first version of the cognitive walkthrough, and usability testing; and uses the number of detected problems, problem severity, type of problems, and person-hours required for the evaluation as a criteria for comparisons. Their results show that usability expert evaluations relying on the evaluators' expertise are the most effective methods: they identify the greatest number of problems, are successful in identifying serious usability problems and are also lowest in cost. The cost-effectiveness of expert reviews is reduced because of the many low-

priority problems that the method also reveals. Usability testing, on the other hand, is an effective mean to identify serious and recurring problems, and avoids identifying low-priority problems. However, it is also the most expensive testing method requiring much more time for analysis than the other methods. (Jeffries *et al.* 1991)

The comparison by Jeffries *et al.* (1991) and several other comparisons on usability evaluation methods received heavy criticism from Gray and Salzman (1998) in their article "Damaged merchandise?". This article studies the comparisons made on usability evaluation methods as scientific experiments, and finds various faults in their validity. For example, a very basic element in the comparisons is an instance of a usability problem. Still, it is very challenging to determine whether a finding is a hit, a false alarm or a correct rejection, or if some problems are totally missed. As Gray and Salzman (1998) point out, there is no truth on the matter, as different evaluators and teams may classify and interpret the findings in very different ways. Consequently, John Karat points out in the commentary on "Damaged merchandise?" that it is not of interest in HCI field which usability evaluation method is the best, because the answer depends on so many factors, but it is more important to *"have knowledge and experience with a range of techniques*" to develop "*an understanding of when to use what in what proportions*" (Olson & Moran 1998, pp. 265-269).

3.6 Impact on development process

Some evaluation methods are focused solely on detecting usability problems and, consequently, give only little support on generating recommendations for improvements. Therefore, measures for the impact on the development process have been presented to enrich the tools for assessing and selecting suitable methods. For example, Sawyer *et al.* (1996) define a metric called *impact ratio* to indicate the proportion of the problems that the development team has committed to fix from all the problems found in the evaluation. John and Marks (1997) have a similar measure called *persuasive power*. They also have a measure called *design-change effectiveness* to indicate the effect of the change in the usability of the product. This effectiveness is assessed by testing the new versions, and by comparing the number of problems found with the previous results. (John & Marks 1997)

Sawyer *et al.* (1996) calculate their own impact ratio within ten usability inspections they have made, and end up with ratios ranging from 58% to even 96%. Based on their experiences, they identify several factors improving their impact ratio:

- *Developers' respect:* The usability group conducting most of the evaluations is an internal group in the company, so it has a long-term relationship with the development groups, and has already earned the developers' respect.
- *Multiple methods:* The evaluations involve several methods, thereby adding the reliability of the results.

- *Early involvement:* The group tries to work with the development teams as early in the development process as possible.
- *Client participation:* The group involves the development team in the evaluations by having a member of the team attend the evaluations.
- *Written reports:* The usability group gives the development teams written reports including descriptions of the problems as well as alternatives for fixing them.
- *Specific recommendations:* The group provides detailed and technically specific recommendations to fix the detected problems.
- *Severity level:* The group rates each problem by its severity.
- *Written response:* The group requires a written response from the clients to report their commitment to fix the problems.
- *Easy response process:* The group sets up meetings with their clients to go over the report, and to discuss on their plans to fix the problems.

Also Uldall-Espersen, Frøkjær and Hornbæk (2008) find the use of multiple evaluation methods to improve the impact on development process, as well as having people with complementing areas of expertise involved to get a deeper understanding of the problems, and to find "*more comprehensive and profound redesign solutions*". To be able to suggest effective modifications that actually support the product goals, Sawyer *et al.* (1996) recommend that the evaluators become familiar with the product, the product family and the design goals, so that the recommendations do not represent inconsistencies within the product family. Similarly, the studies by Hornbæk and Frøkjær (2008a) show that keeping the business goals in mind, and relating the found usability problems to these goals improve the impact of the results: developers rated the utility of these business-oriented reports even 30-42% higher than reports without explicit connections to the business goals.

Hertzum (2006), on his part, emphasises the role of the early involvement through evaluations focusing on finding and fixing the most severe problems at the beginning, and widening the scope to less severe problems only in the later evaluations. The experiences of Ehrlich *et al.* (1994) support the effectiveness of having developers witness the users' performance either in live situation or through videotapes. As Høegh *et al.* (2006) put it: "*observations of user tests facilitated a rich understanding of usability problems and created empathy with the users and their work*".

All in all, most of the factors affecting to the impact of the evaluations are not attributes of the used methods, but attributes of the overall process and especially of reporting (Hartson *et al.* 2003). For example, the resources available to implement the suggested changes are a relevant factor in the developers' motivation (John & Marks 1997). Therefore, Sawyer *et al.* (1996) suggest to find out the possibilities and resources for further development even before committing to the evaluations.

3.7 Experiments on contextual factors of usability testing

Contradictions in usability measures, especially between performance and preferences, have been a source for many studies and experiments. Although the test users' performance and ratings in the post-test questionnaires usually correlate, there are also contradictions. For example, in the meta-analysis by Nielsen and Levy (1994) covering 113 comparisons between alternative interfaces, 71% of the results show a strong positive correlation between users' performance and their preferences, but several users prefer the system that has been worse for them if assessed with objective measures. Furthermore, in a similar meta-analysis on data from 73 usability studies, Hornbæk and Law (2007) find only low or medium correlations between usability measures, and the analyses by Hertzum and Frøkjær (1996), and Frøkjær, Hertzum and Hornbæk (2000) find no correlations at all. Then in a more recent analysis on 90 summative usability studies made in practice, Sauro and J.R. Lewis (2009) find especially strong correlation between common usability metrics, when analysed at the task-level, but also the test-level shows at least medium correlations.

This subchapter outlines studies made on contextual factors that potentially affect the results of a usability test, and may also cause contradictions between various measures. The list of factors is not assumed to be all-inclusive. For example, the effects of culture is not addressed here. Even so, the list gives a good basis for considering the consequences of various decisions made when planning a usability test. It also gives ingredients for designing new experiments on the effects of various contextual factors in usability testing.

Factors of usability testing have been studied already in 1990's by David Biers and his colleagues as they made studies of "*factors which can potentially affect the results of such [usability] tests*" (Barker & Biers 1994). These factors include evaluator intervention (Held & Biers 1992); testing in pairs and the presence of an observer (Hackman & Biers 1992); retrospective and concurrent thinking aloud (Ohnemus & Biers 1993); user's self-consciousness and laboratory environment (Barker & Biers 1994); and prototype fidelity (Catani & Biers 1998). More recently, Juergen Sauer and Andreas Sonderegger have done similar studies of the effects of laboratory set-up and facilitator presence (Sonderegger & Sauer 2009); prototype fidelity and aesthetics (Sauer & Sonderegger 2009); user expertise and prototype fidelity (Sauer *et al.* 2010); testing environment and task scenario (Sauer & Sonderegger 2011a); product aesthetics and user state (Sauer & Sonderegger 2011b); design aesthetics (Sonderegger & Sauer 2010); product aesthetics and usability over time (Sonderegger *et al.* 2012); and socio-cultural background and product value (Sonderegger & Sauer 2013). As a basis for their studies, Sauer *et al.* (2010) present a four-factor framework of contextual fidelity building on users, tasks, prototypes and testing environment (See Figure 1 in Chapter 1.1 on p. 3). Compared to the factors addressed in this thesis, the framework leaves out the characteristics of the selected evaluation methods, such as the use of the thinking aloud method.

Based on the literature study and my own experiences, the factors to be addressed in this chapter include: test users' level of expertise; their expectations

of the assessed system; the test location and movement; the moderator presence; the fidelity of a prototype; and product aesthetics. The effects of test tasks, the amount of test tasks and their coverage and relevancy were already discussed in Chapter 3.8.1 “Sampling users and test tasks”, and studies of thinking aloud were addressed in Chapter 3.2 “Thinking aloud”. Outside the framework by Sauer *et al.* (2010), the diversity of evaluators and the use of the thinking aloud method are also discussed, as well as the effect of being a participant in an experiment in general.

3.7.1 Participating an experiment

Usability tests, especially summative tests, can be considered as experiments both from the evaluators’ and the participants’ point of view. The effects of this kind of experiments and of being a subject in an experiment have been studied in social psychology for a long time. For example, Orne (1962) describes studies in which participants were asked to do monotonous and meaningless tasks, and still, the participants kept on performing these tasks for several hours until the experimenter finished the test. The participants in scientific experiments usually think that the experiments are important, and are therefore eager to validate the hypotheses they assume the experiment to have. They try to infer the hypotheses from very different sources, such as: their former experiences with experiments; location of the experiment; and gestures and tones of the experimenters. This applies even if the experiment or tasks have no specific purpose. If the hypotheses are very clear and obvious, bias to the opposite direction is also possible. Therefore, Orne recommends to systematically study the participants’ perceptions of the experimental hypotheses, and to study if their behaviour correlates with these perceptions more than with the experimental variables. (Orne 1962)

Although usability tests are not generally considered as scientific or psychological experiments, many similar phenomenon as in Orne’s (1962) studies can be identified also in usability tests. For example, the test users are ready to ”*perform a very wide range of actions on request without inquiring as to their purpose, and frequently without inquiring as to their durations*”, and the tasks are carried out with care even for a long time (Orne 1962). It is unlikely that the test users would perform with similar care and persistence in contexts outside the tests. Although the test users are told that they are not the subjects in the test, they easily feel obliged to find as many problems as possible, or pressure to perform as efficiently as possible. Therefore, it is important to identify various contextual factors that may influence the test results, and thereby be able to assess their effect, and especially the extent to which the test results can be generalised to contexts outside the test settings.

3.7.2 Test users’ expertise

The studies of the effects of test users’ expertise have ended up to very diverse and even contradicting results. For example, Nielsen and Levy (1994) find in their meta-analysis that expert users use remarkably wider scale in their pre-

ferences than the novice users, although the variety of their performance does not differ more than the novice users' performance. In contrast, Barker and Biers (1994) find experienced users to make less errors than novice users, but found no significant differences between their subjective evaluations. In the studies by Mugge and Schoormans (2012), the expected usability depends on the looks of the system: the experts rate the expected usability slightly lower than novice users if the product looks traditional, but their expectations raise a bit if the product seems novel, whereas novice users' expectations decline significantly with novel looking products.

On the part of the number of usability problems found and their severity, the results of Kjeldskov, Skov and Stage (2005b) show that novice users find significantly more problems than experts, and the novices assess the problems as more severe than the experts. However, the study by Sauer, Seibel and Rüttinger (2010) shows that the expert users find more usability problems, but assess them as less severe. The differences in the results can partly be explained by the differences in the test settings: Sauer *et al.* (2010) have a traditional between-subject study with expert and novice users, whereas Kjeldskov *et al.* (2005b) have a within-subjects study by using the same seven users as test participants with a break of 15 months, so that the test participants have time to evolve from novice users into experts. Sauer *et al.* assess a floor scrubber, and Kjeldskov *et al.* a large commercial electronic patient record system. The severity assessments in the studies are also different in the respect that usability and human factors experts make the severity assessments in the study by Sauer *et al.*, whereas the users make the assessments themselves in the study by Kjeldskov *et al.*

The performance measures by Sauer *et al.* (2010) do not show significant differences between expert users and novice users, although the evaluators point out that the expert users perceive the cleaning work as a whole, and are more capable to predict potential usability problems in real use even outside the situations covered by the test tasks. Therefore, Sauer *et al.* 2010 recommend to include a possibility for the expert users to consider future usage in usability tests, and also a possibility to self-report potential problems.

3.7.3 Test users' expectations

Test users have various expectations on the evaluated product based on their prior experiences, state of mind, relation to the brand, and many other factors. The effect of these expectations is hard to assess even if the tendency of the expectations is known. Therefore, various experiments have been made to study the effect of expectations from different points of view: Raita and Oulasvirta (2011) study the effect of positive or negative priming; Sauer and Sonderegger (2011b) assess the effect of negative usage event; Mugge and Schoormans (2012) study the effects of novelty; and Sonderegger and Sauer (2013) study the effect of the price of a product.

Raita and Oulasvirta (2011) use two different versions of a product review to prime the users before the test: one appraising the product, and the other one criticizing it. Also the difficulty of the test tasks is manipulated between

the test users to see, if the subjective usability ratings change according to the actual performance in the test. A comparison group does not receive any review before the test. The results of the study show a very strong effect of positive priming on subjective post-test ratings regardless of the actual performance in the test: test users who have read a positive review of the system rate the system more positively than the ones with negative review or no review at all. However, no significant effect is found in the post-task questionnaires immediately after each task when the workload and emotions are assessed. The effect of the priming is significant also in the task success with the more difficult test tasks: the test users with negative priming complete significantly more tasks than the positively primed. When assessing the workload, the effect of priming is no longer significant, and neither is the interaction between the priming and the task difficulty. (Raita & Oulasvirta 2011)

Sauer and Sonderegger (2011b) study somewhat similar phenomenon by introducing impossible test tasks into the test situation, and thereby causing a negative usage event and possible negative user state. They assess the effect of this negative usage event on the ratings of perceived usability of the specific task and the subsequent tasks. These ratings decrease for the impossible task, but recover in the ratings for the next tasks. This result gives more freedom to the order of the test tasks, as the performance with the previous tasks does not affect the ratings of the next ones. (Sauer & Sonderegger 2011b)

Mugge and Schoormans (2012) study the effects of novelty on the expected usability by changing the appearance of the products to be traditional or novel. For example, the colour of washing machines is either traditional white or novel black in the comparisons. The test users do not use the product, but just see a picture of it, and read through its technical specifications. The results of the tests show a negative effect of the novel appearance on the expected usability. For the novice users, the effect is stronger than for the experts. The results indicate that people associate novelty with technological advancement, and therefore expect novel products to be less usable, especially for novice users. (Mugge & Schoormans 2012)

Sonderegger and Sauer (2013) then, study the effect of product value in two different socio-cultural regions. They manipulate the price of the evaluated coffee machine, and ask users in Switzerland and East Germany to assess its usability after brief test tasks. The results show that Swiss users assess the more expensive products to be more usable, whereas East German users evaluate them less usable. Thereby, these results also indicate that usability test results cannot be generalised across cultures. (Sonderegger & Sauer 2013)

3.7.4 Test environment

The studies on the effects of the testing environment have included at least: comparisons between laboratory and field settings (Kaikkonen *et al.* 2005; Kjeldskov *et al.* 2005a; Duh, Tan & Chen 2006; C.M. Nielsen *et al.* 2006; Sauer & Sonderegger 2011a); between laboratory settings and letting test users do the tasks on their own in the environment they choose themselves (Schulte-Mecklenbeck & Huber 2003); between clear and covered laboratory settings

(Barker & Biers 1994); and in laboratory settings between various movements (Kjeldskov & Stage 2004). In addition to the setting of the study, also the results of the studies have been very divergent.

The study by Barker and Biers (1994) is among the first ones to study the effects of laboratory settings on the results of usability testing. In this study, one of the three independent variables is the visibility of cameras and a two-way mirror in the test laboratory, so the test room itself is the same for all the test conditions. The results of this study show no effect of the visibility of the testing equipment. (Barker & Biers 1994)

The study by Schulte-Mecklenbeck and Huber (2003) then, has very clear difference between laboratory and other settings, but includes also many other differences, such as having an observer in the laboratory settings but not in the uncontrolled settings. Therefore, the effects of the location are left unclear in the results. Still, the results show that participants doing the test tasks on their own search for less information to make their decisions when compared to the participants in the laboratory settings. The risk of quitting is also considerably higher in the uncontrolled settings. (Schulte-Mecklenbeck & Huber 2003)

Also the comparisons between the laboratory settings and the tests in the field have reached very diverse results. For example, Kaikkonen *et al.* (2005) find no significant differences in the number of detected problems; their type or severity; performance times; or overall user performance between the test conditions. The only difference in the settings is that the users tend to act more casual, and comment more freely on the application in the field settings (Kaikkonen *et al.* 2005). Duh *et al.* (2006), on the other hand, have quite contrary results, as they find significantly more problems in the field, and the difference with the critical problems is especially clear. The performance times in the field are also significantly longer. Furthermore, the test participants behave more negatively in the field than in the laboratory where they have more relaxed expressions in the study by Duh *et al.* (2006). Unlike the test by Kaikkonen *et al.* (2005), the test of Duh *et al.* (2006) includes tasks requiring phone calls and interaction with other persons through the phone, so both the social and physical environment in moving trains are emphasised in this test settings. Indeed, Duh *et al.* (2006) list as possible reasons for the differences in the test results several factors from the field settings, such as: the noise level in the trains; need for moving; lack of privacy; additional effort needed to perform tasks in a moving train; and additional stress in public settings.

Social discomfort is brought forth also in the study by Kjeldskov *et al.* 2005a. In their study, social discomfort is reported only in the field settings, and the use in real context reveals also other problems in the validity and precision of the data in the mobile system that were not noticed in the laboratory settings (Kjeldskov *et al.* 2005a). Also C.M. Nielsen *et al.* (2006) find more critical problems in the field settings than in a laboratory, and discover problems related to cognitive load and interaction style only in the field settings. However, the overall satisfaction or workload has no significant differences. Even so, when the components of the overall workload are studied separately, a very significant difference is found in mental demands, and a significant dif-

ference also in frustration level, as both these aspects are higher in the laboratory settings. Therefore, *"it is worthwhile conducting user-based usability evaluations in the field, even though it is more complex and time-consuming"*. (C.M. Nielsen *et al.* 2006)

Sauer and Sonderegger (2011a) continue the comparisons between laboratory and field settings, but add two more independent variables to the study. They have one or two chores for the users to focus on when making tea with a kettle, and have a product information urging to save energy visible or not. They assess the effect of this product information by measuring the unused water, number of manual switch-offs, perception rate of the product information, self-reported rate of compliance with the product information, and perceived temporal demands. The results show that the effect of the product information in the amount of unused water is the greatest in the laboratory settings with a single task scenario. The participants also report to have noticed the product information significantly more often in this condition, as well as to have complied with the information. As possible reasons for the better compliance in laboratory settings, Sauer and Sonderegger name unfamiliar environment in which the users naturally observe their environment with more detail, and, thereby, more easily respond to environmental cues. The single task scenario, on the other hand, leaves more resources for controlling the kettle and switching it manually off regardless of the location of the test, and even regardless of the presence or absence of the product information. (Sauer & Sonderegger 2011a)

Also Kjeldskov and Stage (2004) adjust the level of concentration into the test task in two experiments. In the first experiment, the users sit still or walk in constant or varying speed in a laboratory, and they either need conscious attention in navigation or not. In the second experiment, half of the users play a computer game on a dance mat. The results when walking in a pedestrian street are used as a reference. The results of the first study show that all the combinations requiring moving reveal substantially the same amount of usability problems even when compared by their severity. Only the condition of sitting still by a table help to reveal significantly more problems, but most of them are cosmetic. The performance times, the number of completed tasks and the number of false button presses are similar in different conditions, but the perceived workload has clear differences especially in the mental demand, effort and overall workload. Statistically, sitting and walking at a constant speed are not experienced as very different when considering the workload, whereas walking at varied speed, walking on a changing course or walking on a street cause significantly greater perceived workload. In the second experiment with the dance mat, the results show that the users playing the game do not notice all the usability problems that are revealed while walking on the street. Furthermore, the participants walking on the street soon start to perform quicker than in an initial trial task, but the participants playing the computer game continue to perform slower through the whole test. (Kjeldskov & Stage 2004)

3.7.5 Moderator presence

The effects of the presence of other people on the test participants' performance have been studied in several experiments, but only in a few studies in the context of usability testing, such as: Held and Biers (1992); Hackman and Biers (1992); Shi (2008); and Sonderegger and Sauer (2009). Also Eger *et al.* (2007) address this issue when comparing concurrent and retrospective thinking aloud. As a background, Davis *et al.* (1968) summarise some of the studies outside usability testing: with routine tasks, the presence of another has been shown to facilitate the performance, whereas with tasks requiring abstract thinking, the effect has been inhibiting. However, in their own experiment requiring problem solving, Davis *et al.* find the presence of a test experimenter to have an insignificant or even a facilitating effect. As the authors conclude, the "*finding concerning experimenter presence has perhaps raised more questions than it has answered*". (Davis *et al.* 1968)

In the context of usability testing, Held and Biers (1992) are among the first ones to evaluate the effect of evaluator intervention. In their study, the intervention is restricted to the evaluator asking clarifying questions if users experience some problems. Held and Biers expected the intervention to "*force the user to dwell on the software problems and this would lead him/her to form a more negative impression on the product than ordinarily would be the case*". The results support their hypothesis on the part of the expert users, but with novice users, the effect is slightly opposite. (Held & Biers 1992) In another study by Hackman and Biers (1992), the test users' performance and preferences are compared when the users either perform alone in a test room; with an observer in the same room; or as a pair, but the results show no significant differences between these three conditions.

The studies of Schulte-Mecklenbeck and Huber (2003) do not focus on the effects of a moderator presence, but on the effects of the location of a test comparing laboratory settings and uncontrolled settings in locations selected by the users. Instead of revealing significant effects of the location, the results indicate that the mere presence of an experimenter may affect the results. For example, the results show that participants doing the tasks on their own use less information for their decisions. The difference in the number of dropouts is also considerable, as all of the participants in the laboratory settings complete the experiment, whereas 20% of the participants in the informal settings drop out. However, as the comparison settings have quite many variables, the effect of an experimenter presence cannot be analysed alone. (Schulte-Mecklenbeck & Huber 2003)

Also the study by Eger *et al.* (2007) focus on other issues, namely comparing concurrent and retrospective thinking aloud with two online search engines, but the post-test questionnaires address the effects of the moderator presence as well. Their results show that the presence of a test moderator has a negative effect on the test users when they are using concurrent thinking aloud compared to the retrospective setting in which they do not need to think aloud during the task performance, or to be reminded of that (Eger *et al.* 2007).

The field study by Shi (2008) focuses on the relationship and communication between Chinese moderators and test users. According to the studies by Yeo (2000), it was expected that the Chinese users would consider the moderator's feelings, and hesitate to give negative comments on the evaluated system. However, the results show that most users focus on the test tasks and take the role of helping to find potential usability problems instead of being overly polite to the moderator (Shi 2008).

The study by Sonderegger and Sauer (2009), on its part, focuses on the effects of having observers in the test room. Their study compares three conditions: no person present; a moderator present; and a moderator and two non-interactive observers present in the test room. The moderator does not help the users if problems arise, so there is minimal interaction between the users and the moderator. The non-interactive observers are presented as designers of the evaluated system, and they stay silent during the test. The comparisons are made based on performance data, subjective measures and physiological parameters including heart rate variability. The results show that the non-interactive observers have a significant effect in several measures. For example, the stress level increases, and the task times get longer, when non-interactive observers are present. The presence of a moderator or other observers also affect the emotional measures, as the users working alone rate their emotions as more positive than the others. However, the perceived usability or the aesthetic appeal are not affected by the test conditions. Although the users performing alone gave more positive ratings of their emotions, the researchers also noticed that a moderator who is able to set up a good rapport with the test users may also enhance their performance. (Sonderegger & Sauer 2009)

3.7.6 Prototype level

On the part of the prototypes, the effect of the fidelity of the prototype has been in focus in several studies. These studies have included comparisons between paper prototypes and interactive software simulations (Virzi, Sokolov & Karis 1996; Catani & Biers 1998; Boothe, Strawderman & Hosea 2013), as well as comparisons between prototypes and the real physical products (Archer & Yuan 1995; Sauer & Sonderegger 2009; Sauer *et al.* 2010). For example, the study by Archer and Yuan (1995), shows that software simulations of a keyboard are valid tools in evaluating the system in iterative development as the use of simulations affects substantially the same number of key presses as the physical keyboard, and also the user preferences are much alike. The study by Säde, M.H.T. Nieminen and Riihiaho (1998), on the other hand, shows that usability tests conducted even with inexpensive and quickly made three dimensional paper prototypes can produce convincing results for the decision making during development when considering between two different alternatives. Also the effect of having several design alternatives represented in the prototypes (Tohidi *et al.* 2006a), and the possibility for users to sketch their own redesign proposals (Tohidi *et al.* 2006b) have been studied.

Prototype fidelity has several dimensions, such as: the breadth of features; depth of functionality for these features; similarity of interaction with the final product; and aesthetic refinement (Virzi *et al.* 1996). For example, Rudd, Stern and Isensee (1996) classify prototypes into low-fidelity prototypes with limited functions, and demonstrating the general look of the interface instead of its operation; and into high-fidelity prototypes, usually including complete functionality and allowing users to explore the system as if it was the final product. Low-fidelity prototypes are valuable especially in the early phases of product development when gathering the product requirements (Rudd *et al.* 1996), whereas high-fidelity prototypes help in getting estimates of performance measures (Virzi *et al.* 1996). Usability tests can be conducted both with low- and high-fidelity prototypes as well as with finished products.

The first studies comparing the use of low- and high-fidelity prototypes show that both prototypes reveal substantially the same sets of usability problems. For example, in the studies by Virzi *et al.* (1996), paper cards presenting the screen and keyboard in various actions, and the moderator simulating a voice response system are used as low-fidelity prototypes to be compared with high-fidelity prototypes. The results show that the prototypes reveal similar sets of usability problems, and even the proportions of test users detecting particular problems are significantly the same with different prototypes (Virzi *et al.* 1996).

The results of Catani and Biers (1998); Sauer and Sonderegger (2009); and Sauer *et al.* (2010) with three different levels of prototype support these findings, as the performance results with different prototypes are similar, as well as the subjective evaluations. In the study by Sauer and Sonderegger (2009), also the aesthetics of the design is manipulated at two levels. Quite expectedly, the aesthetics affect the perceived attractiveness, but also the prototype fidelity has an effect in this assessment, as the fully functional system with lower aesthetics gets significantly lower ratings for attractiveness compared to all the other conditions. As a potential reason for this effect, Sauer and Sonderegger suggest that for the low- and medium-fidelity prototypes, the test users may have made their own implications on the looks of the final product, and make their assessments based on these implications instead of the prototype in hand. (Sauer & Sonderegger 2009)

The study by Boothe *et al.* (2013) continue the experiments on the fidelity of the prototypes, but focuses on the medium of the prototypes. The experiment substantially uses the same user interface either as printed hard copies or as a slideshow operated by a moderator on a computer. Similarly to the studies by Virzi *et al.* (1996), the results show that the medium does not affect the probability of finding a usability problem. However, in this study, the medium affects the probability of finding severe problems, as the computer medium proves to be more effective in this respect. The subjective ratings of the system usability, however, remain the same regardless of the prototype medium. (Boothe *et al.* 2013)

In addition to the fidelity of the prototype, also the effect of having several prototype alternatives has been studied. For example, Tohidi *et al.* (2006a)

compare the user ratings, amount of positive and negative comments, as well as the number of redesign proposals between settings in which users get either one prototype or three prototypes. All the design alternatives are paper prototypes with their own distinctive stylistic approach. The results of the study show that users give significantly higher ratings when they assess only one design compared to assessing three alternatives. The number of positive comments on a design is significantly lower when it is presented with alternative designs, and accordingly, the number of negative comments for the worst design is significantly higher in this setting. Some test participants even comment that they would never choose the worst design, when they see the alternatives, but no one comments anything similar in the single design condition. In the number of redesign proposals, however, there is no significant difference between the two conditions, but the number of suggestions focusing in quite irrelevant details is higher in the single design condition. As possible reasons for the results, Tohidi *et al.* suggest that having multiple design alternatives gives the test users more base of experience, and also clearly signals that the design is still under consideration. (Tohidi *et al.* 2006a)

To get more redesign proposals, Tohidi *et al.* (2006b) continue the study with a new approach in which they ask the test participants to sketch their ideal product design on a sheet of paper after the experiment described above. The participants in multiple design condition have seen more design alternatives, so also their sketches have more variations from the original versions, and include even versions with no obvious similarity to the original ones. As advantages of sketching, Tohidi *et al.* name the possibility for the test participants to reflect and discover new ideas, and for the evaluators to get reflective feedback and design proposals from the users instead of only reactive feedback and comments. (Tohidi *et al.* 2006b)

3.7.7 System aesthetics

Several studies have been made to assess the interactions of the aesthetics of a product and its perceived usability. The studies of Tractinsky, Katz and Ikar (2000) have inspired many others to similar studies, since the results indicate that the aesthetics affect both on perceived aesthetics and perceived usability in the post-test ratings, whereas actual usability has no effect on these assessments. Most of the follow-up studies find correlations between aesthetics and perceived usability, so that highly aesthetic products are valued as more usable regardless of the actual performance with the product. The objective performance measures have varied between the studies from more effective task performance with attractive products (Moshagen, Musch & Göritz 2009; Quinn & Tran 2010; Sonderegger & Sauer 2010) to worse performance (Ben-Bassat, Meyer & Tractinsky 2006; Sauer & Sonderegger 2011b). The differences have been explained partly by increased motivation with appealing products if they are for utilitarian use, and prolongation of joyful experience in leisure context (Sonderegger & Sauer 2010; Sauer & Sonderegger 2011b). Also in the study by Sauer and Sonderegger (2009) in which the effects of prototype fidelity and

aesthetics of design are assessed, the influence of aesthetics is highly significant on the perceived usability, whereas the prototype fidelity has no effect.

Also Hassenzahl (2004) studies this issue by analysing the interaction between perceived usability, hedonic attributes, goodness, and beauty. The results show that the assessed overall goodness of the product depends both on the perceived usability and the hedonic attributes, and the use experience strengthens the effect of the perceived usability in goodness. However, the hedonic attributes and beauty remain stable regardless of the use experience. (Hassenzahl 2004)

Hassenzahl's studies received criticism, for example from Monk (2004), on the narrow sample of products used in the studies, and on averaging only over users, and not over the products as well. Therefore, Hassenzahl and Monk (2010) made new studies to re-examine the relations. The results suggest that pragmatic quality (*i.e.* perceived usability) is only indirectly related to beauty, so that beauty affects to perceived usability through goodness, whereas beauty and hedonic qualities are directly related (Hassenzahl & Monk 2010).

Also the study by Ben-Bassat, Meyer and Tractinsky (2006) focus on the interplay of usability and aesthetics. Their experiment manipulates the usability and aesthetics of the system independently, and includes a condition in which the reward given to the users is based on their actual performance with the system, and another condition in which the users make bids in simulated auctions on the systems to be used in the final tasks. The results show that even in the test condition with the monetary incentive, the users value the more aesthetic systems over the more usable ones when assessing the system usability in questionnaires. However, the bids made in successive auctions correlate clearly with the actual performance with the system, so the researchers recommend to use multiple serial auctions. For example, they used five repeated auctions in their own study. Nevertheless, even after the auctions, the users assess the high aesthetic systems slightly more usable in the questionnaires contrary to their former performance and bids. (Ben-Bassat *et al.* 2006)

Hartmann, Sutcliffe and De Angeli (2008), on their part, introduce a method in which the test users are asked to select the ideal product from the ones they have tested for different target populations and contexts. Although the users prefer the more engaging and aesthetic version over the more conservative and usable one for their own future use, they recommend the more usable system to others if the use context is expected to be more serious than playful (Hartmann *et al.* 2008).

To study the stability of the positive effect of aesthetics, Sonderegger *et al.* (2012) measure the test users' performance, perceived usability, perceived aesthetics and emotion in a period of two weeks. The results show that in one-off usability tests, the positive effect of aesthetics is considerable, but this influence wanes over time. Therefore, they recommend to use longitudinal multiple session tests to minimise the effects of aesthetics on perceived usability and emotions. (Sonderegger *et al.* 2012)

3.7.8 Evaluator effect

The differences in the results of various evaluators have been studied with different approaches, such as: by using the same video recordings from the same usability test sessions (Jacobsen, Hertzum & John 1998; Vermeeren, van Kesteren & Bekker 2003; Hertzum, Molich & Jacobsen 2014); and by comparing the results of various usability groups evaluating the same system with the same goals and instructions (Molich *et al.* 2004; Molich & Dumas 2008). This subchapter presents these studies along with some recommendations to overcome or to minimise the evaluator effect.

The study by Jacobsen *et al.* (1998) is one of the first ones comparing the usability problems found from the same video recordings by different evaluators. In this study, all the four evaluators have experience both in HCI and usability testing, but two of them are quite experienced, and the other two are still novices in analysing usability tests. The evaluators are asked to list and describe all the usability problems that they detect from the video recordings, and to pick the ten most important problems to be fixed in the next release of the system. The results of the comparisons show that each evaluator finds from 39% to 63% of the total number of problems; only 20% of the problems are detected by all the evaluators; and as much as 46% are detected by only one evaluator. For the severe problems chosen to the top-10 lists, the evaluator effect is somewhat smaller, but still, not a single problem is selected by all the evaluators to the top-10 list. (Jacobsen *et al.* 1998)

The studies by Vermeeren *et al.* (2003) with 2 evaluators in 3 different cases, and by Hertzum *et al.* (2014) with 19 evaluators analysing 1 case also support these findings of very diverse evaluation results. Hertzum and Jacobsen (2001) study this evaluator effect also in the heuristic evaluation and the cognitive walkthrough in addition to the thinking aloud test, and find the evaluator effect with all the methods. As a way to minimise this effect, all the studies recommend to have more than one evaluator analysing the sessions, and to make the evaluators compare and discuss on their results.

If individual evaluators get such diverse results from the same video recordings, or with a strict procedure, such as cognitive walkthrough, the results of different usability teams conducting their own usability evaluations could be even more varied. Indeed, in the studies by Molich *et al.* (2004) and Molich and Dumas (2008) the results of the various teams overlap very little. In the first study, only usability testing is used as the method for evaluation (Molich *et al.* 2004), but in the latter study, the teams are able to choose the methods they apply (Molich & Dumas 2008). In the first study, 9 teams reveal a total of 310 usability problems, but none of the problems is reported by all the teams, and as much as 75% of the problems are reported by only one team (Molich *et al.* 2004). Also in the latter study with 17 teams, even 60% of the issues are reported by only one group (Molich & Dumas 2008). As conclusions on the findings, Molich *et al.* (2004) recommend to use an appropriate mix of evaluation methods instead of relying on one method, and to concentrate on the most important features instead of trying to cover the whole system in one evaluation.

The estimates on evaluators' agreement are usually based on lists in which the issues of individual evaluators are matched and integrated as one list containing unique usability problem tokens. Thereby, the process of matching the problems has a substantial effect on the estimates of agreement and overlapping of results. The study by Hornbæk and Frøkjær (2008b) focuses on this matching process, its criteria, and differences when made individually or in teams. In addition to the 50 novice evaluators, also 10 HCI experts outside the testing match the problems in this study. The process has three phases: first, all the novice evaluators conduct a usability test on their own; then, they try to individually match the problems found by all their five team members, and finally, they do the matching as a team. The usability experts who have not taken part in the usability tests or in planning the tests make the matching only individually. The results of the study show an agreement of about 40% in the matching, and the results of the experts do not differ significantly from those of the novices. Matching in teams seems to improve the agreement, and especially improve the evaluators' satisfaction with the matching, as the evaluators are able to discuss on the problems with the ones who have originally detected and reported the problems. (Hornbæk & Frøkjær 2008b)

3.8 Challenges in usability testing

This subchapter reviews articles that discuss the challenges and problems that usability testing faces. An overview of the status of these challenges is presented in the conclusions of this thesis, contrasting the problems to the methods and studies presented in this thesis.

Some central problems in usability testing are discussed in an article by Patrick Holleran already in 1991. In this article, Holleran argues that good usability testing is similar to good empirical research. A good usability test is, thereby, valid and reliable giving accurate measures of the attributes of interest, utilising several evaluation methods, and making statistical analysis on the quantitative measures. Contrasting to these aims, Holleran finds several pitfalls in usability testing that he categorises into three groups: sampling problems mainly when planning the tests; methodological problems when conducting the test sessions; and problems in interpreting the results. Sampling problems refer to the selection of subjects and test tasks, and also to the poor reporting of user characteristics and their potential effects to the test results. Methodological problems, on their part, cover issues related to having humans as test participants, such as: intensified motivation in test situations; potential moderator bias; and the possible effects of the use of the thinking aloud method. Problems in interpretation, finally, refer to emphasis on subjective interpretations of qualitative data instead of statistical analyses with content analyses and observational coding procedures. (Holleran 1991)

3.8.1 Sampling users and test tasks

As Holleran (1991) points out, usability testing is closely related to experiments in which test participant and test task sampling may considerably affect

the results. Indeed, sampling, biased test user selections, and misleading generalisations with small sample sizes are problems identified also in other studies of usability testing. For example, Barkhuus and Rode (2007) analyse the articles in the proceedings of CHI conference in five year intervals from 1983 to 2006, and find a trend of having less and less test users in empirical usability evaluations. Also the diversity of the test users is weak due to the high proportion of students as test participants (Barkhuus & Rode 2007).

The required number of test participants in usability tests has been discussed and studied extensively (*e.g.* Virzi 1992; J.R. Lewis 1994, 2001a; J. Nielsen 1994a), but studies of the effect of the test tasks have been more rare. Therefore, Lindgaard and Chattratichart (2007) recommend to shift the focus from the number of test users to the number of test tasks, as they find a significant correlation between the number of test tasks and the number of detected usability problems. In their study, Lindgaard and Chattratichart analyse the results of usability tests that nine separate evaluation teams make on the same web service. The number of reported usability problems is restricted to maximum of 50, but the number of test users and test tasks is not restricted[25]. Therefore, Lindgaard and Chattratichart set the number of test users and number of test tasks as independent variables in their study, and analyse if these variables correlate with the proportion of problems found. Their results show no significant correlation between the number of test users and problems found, but instead, show a significant correlations between the number of test tasks and problems found. (Lindgaard & Chattratichart 2007)

Another important factor affecting the quality of the test results that Lindgaard and Chattratichart (2007) find in their analyses is the recruitment procedure of the test participants: the web service evaluated in the study is intended for a variety of users with varying skills and background knowledge, so it is important to recruit a heterogeneous sample of users to cover as many problems as possible. Based on their results, Lindgaard and Chattratichart (2007) recommend to invest resources in a wide task coverage and careful participant recruitment instead of a large number of test users.

Also the quality and relevance of the test tasks have shown to significantly affect the number of problems found (Skov & Stage 2012). Still, most of the usability measures used in practice focus on micro level tasks emphasising perceptual and motor skills, instead of macro level tasks requiring also cognitive problem solving (Hornbæk 2006), and the test tasks are divided into such small activities that the test users find it difficult to assess the overall usability or utility (Dicks 2002). Furthermore, the test tasks usually include only tasks that can be done with the studied system motivating the users to try out more than when they doubt the existence of the required functions (J.R. Lewis 2001a). Lists of goals or forcing users to focus on one part of the system at a time, have also been shown to affect the users' learning (Trudel & Payne 1995), so predefined test tasks may bias the users' performance by the given focus.

[25] The possible effect of limiting the number of reported problems into 50 was not discussed in the article by Lindgaard and Chattratichart (2007), although the total number of problems from all the reported tests was as high as 176.

In addition to these challenges in sampling the test users and the test tasks, there is also the issue of an evaluator effect as showed in the studies by Hertzum and Jacobsen (2001); Molich *et al.* (2004); Molich and Dumas (2008); and Hornbæk and Frøkjær (2008b). These studies have shown that even the same methods applied by different evaluators may lead to very diverse findings and interpretations, causing a problem in the reliability of the evaluations and their results.[26]

3.8.2 Context of use, use over time and utility assessment

The studies by Thomas and Macredie (2002) indicate that the current usability evaluation methods poorly support the study of evolving use habits of new technologies and new applications. Ambience and attention are examples of features that are challenging or even impossible to evaluate in traditional usability test settings, because the effect of use context and user's other chores in real use situations is given too little attention (Thomas & Macredie 2002). Also Cockton's (2006) conference paper on worth-centred design and Hornbæk's (2006) review of usability measures discuss on similar issues concerning usability evolving over time. Cockton (2006) states that studying the lasting impacts beyond interaction, and the motives for using the system, purchasing it, and even recommending it, require self-reporting methods used in the real world. Sengers and Gaver (2006), on their part, emphasise that both the design and evaluation of new systems should take into account the various interpretations that users make on new technologies over time and through communication with various social groups. As appropriate methods for studying the actual use, they recommend a "*variety of assessments from a diverse population of interpreters*" including dynamic feedback in which the users are asked to give their reasoning for the results (Sengers & Gaver 2006).

Also Greenberg and Buxton (2008) point out that the controlled studies that HCI field favours give little or no room for the users to assess the usefulness of the design, or to come up with new ways of utilising the design. This kind of cultural adoption is hard to foresee but is one of the key factors in making new successful innovations (Greenberg & Buxton 2008). To study these various phenomena, Isbister and Höök (2009) call for new methods that take into account new contexts of use including social and environmental contexts, new input mechanisms with bodily interactions, new types of tasks aiming for fun and enjoyment, and evolving usability. This requires for "*contextually appropriate situations, getting nonverbal, nuanced feedback, [and] probing users in new ways*" (Isbister & Höök 2009).

3.8.3 Misuses of usability testing

The term "*usability testing*" has been used with different meanings in different contexts. Dicks (2002) discusses on the varied use of the term, and the loss of its value, if it is also used to cover inspections, surveys and other ways of gath-

[26] Studies of evaluator effect are further discussed in Chapter 3.7.8

ering usability data in addition to its original meaning. Although these discount usability methods can be beneficial in improving products during their development, they do not provide valid and reliable results on usability, thereby weakening the conception of usability testing if the terms are not clearly separated (Dicks 2002).

Even when the terms of usability testing and usability evaluation are used correctly, the methods can be misused. For example, Greenberg and Buxton (2008) list in their paper a number of problems related to usability testing, and situations in which usability evaluations can be even harmful *"if naively done 'by rule' rather than 'by thought'"*. The first problem they describe is the heavy push for usability evaluation in the HCI practice, research and education. Usability evaluation after each iteration cycle is considered a compulsory part of the development process if user-centred design is favoured, and top conferences hardly accept papers that present novel designs without usability evaluations. Greenberg and Buxton also raise up the problem of usability evaluation being quite a weak science, as it lacks several conventions from the experimental sciences. For example, research questions are easily chosen to meet the methods favoured by the review committees, and situations studied in the evaluations are selected to be favourable to the new designs to get positive results. Furthermore, replications of the evaluations are seldom done, and quantitative empirical evaluations are emphasised, although they do not give much room for the users' arguments or intuitions of the evaluated design. (Greenberg & Buxton 2008)

3.8.4 Dogmas in assessment of methods

Assessments and comparisons of usability evaluation methods have received heavy criticism as already discussed in Chapter 3.5 "Criteria for assessing usability evaluation methods". Consequently, Hornbæk (2010) has gathered a list of assumptions that he calls dogmas. These dogmas seem to hold in most of these studies of usability evaluation methods, and thereby, foster the academic debate on the validity and reliability of these studies. The assumptions address the following issues (Hornbæk 2010):

1. The number of problems found is used to rank and compare the methods despite the generality, type, clarity or the validity of the problems.
2. The matching of various problem descriptions is straightforward, although it may generalise the problems too much, or be on too detailed level.
3. Usability evaluation methods are so well instructed that finding usability problems is straightforward, and the effect of evaluators' expertise can be neglected.
4. An individual usability problem is used as the unit of analysis.
5. The evaluations are assessed in isolation from design, without assessing their impact to the development.
6. A single best usability evaluation method exists, and this single method can be recommended over the others ignoring the context and goals of the evaluation, and the parallel use of multiple methods in practice.

7. Usability problems are real if they are found in evaluations.

Also J.R. Lewis (2001a) raises the question whether HCI researchers and practitioners understand the differences between common usability evaluation methods, and the possibilities to compare them. Woolrych, Hornbæk, Frøkjær and Cockton (2011) continue with these themes and make an analogy between usability evaluation methods and recipes. They state that just as recipes alone do not ensure similar results every time they are used, neither do usability evaluation methods. Therefore, they recommend to think usability evaluation methods more as approaches or mixes of ingredients than recipes, and suggest two alternatives for the research on usability evaluation methods: either going above to meals instead of recipes and dishes, or going below from recipes to ingredients. (Woolrych *et al.* 2011)

3.9 Effects of literature review on this thesis

From the two alternatives for the research on usability evaluation methods proposed by Woolrych *et al.* (2011), this thesis has the latter approach of going into the ingredients of usability testing, *i.e.*, the factors of usability testing that potentially affect the outcome of the tests. The previous subchapters have presented several experiments studying the effects of various factors in the usability test results.

Also the experimental part of this thesis addresses these ingredients, namely the use of the thinking aloud method, and the presence of a test moderator. I selected these factors as the focus of my experiment, since they had been taken for granted in most of the textbooks on usability testing, and also in our own teaching. It was also somewhat surprising to notice that there has been very little studies on the effects of the moderator presence in the context of usability testing, and also the studies on the use of the thinking aloud have been quite rare before 2010 – the year I made the experiment in this thesis.

The variety of experiments on the factors of usability testing also show that there is still a need for both new and repetitive tests, as there are so many things affecting their results, and some of the results so far are quite contradictory. Also the request by Wixon (2003) to report more case studies with more precise descriptions of the product and evaluation methods inspired me to update our experiences with our modifications of usability testing.

4 Modifications of usability testing in Aalto University

This chapter presents modifications that we have made to the traditional usability testing in Aalto University, and some experiences in using these methods. The modifications aim to give ideas on how the usability testing settings can be modified to suit various goals and constraints, and also to shed some light on the effects that the modifications may have on the results.

The modifications in this chapter are not all unique. For example, paired-user testing is a method developed well before our studies, so this chapter presents merely its basics and some of our experiences with the method. Pluralistic usability walkthrough, however, was modified considerably for our purposes, so the modifications are presented in quite a detailed level in this chapter. Informal walkthrough and contextual walkthrough, on the other hand, are modifications that we have made to serve the goals of our evaluations in specific situations without explicit examples from literature. However, somewhat similar approaches have been found from the literature later on, and these are presented along with our modifications.

The methods used in our research group, our experiences with them and the way we have taught these methods have been presented in several publications already before this thesis:

- Koivunen, M-R., Nieminen, M.H.T. and Riihiaho, S. (1995) Launching the Usability Approach: Experience at Helsinki University of Technology. In *SIGCHI Bulletin*, Vol. 27, No. 2, pp. 54-60. [An article about teaching usability evaluation at the Helsinki University of Technology]
- Nieminen, M.H.T. and Koivunen, M-R. (1995) Visual Walkthrough. In G. Allen, J. Wilkinson & P.C. Wright (Eds.) *HCI'95, People and Computers, Adjunct Proceedings*. The School of Computing & Mathematics, University of Huddersfield, UK, pp. 86-89.
- Säde, S., Nieminen, M.H.T. and Riihiaho, S. (1998) Testing usability with 3D paper prototypes - Case Halton System. *Applied Ergonomics* Vol. 29, No. 1, pp. 67-73.
- Riihiaho, S. (2000) *Experiences with usability evaluation methods*. Licentiate's thesis. Helsinki University of Technology.
- Riihiaho, S. (2002) The pluralistic usability walk-through method. *Ergonomics in Design: The Quarterly of Human Factors Applications*, Vol. 10, No. 3, pp. 23-27.

- Riihiaho, S. (2009) User testing when test tasks are not appropriate. In *European Conference on Cognitive Ergonomics: Designing beyond the Product – Understanding Activity and User Experience in Ubiquitous Environments* (ECCE '09), L. Norros, H. Koskinen, L. Salo & P. Savioja (Eds.). VTT Technical Research Centre of Finland, Espoo, Finland, Article 21, pp. 228-235. [A conference paper on informal and contextual walkthrough methods]
- Juurmaa, K., Pitkänen, J., Riihiaho, S., Kantola, T. and Mäkelä, J. (2013) Visual walkthrough as a tool for utility assessment in a usability test. In *Proceedings of the 27th International BCS Human Computer Interaction Conference* (HCI 2013). Electronic Workshops in Computing (eWiC), pp. 1-6.

4.1 Paired-user testing

Methods involving two users together trying to solve a problem or exploring a system have many names in usability testing, such as: constructive interaction (O'Malley, Draper & Riley 1984), co-discovery learning (Kennedy 1989), team usability testing (Hackman & Biers 1992), paired-user testing (Wildman 1995), and co-participation (C.E. Wilson 1998). In this thesis, the term paired-user testing is used to describe these methods, but also the other methods are explained if they have some unique characteristics.

Constructive interaction is a technique for studying how people solve shared problems. O'Malley *et al.* (1984) brought this method into the studies of human-computer interaction in the mid 1980's. In this method, two subjects with comparable expertise explain their ideas and rationale behind their hypotheses to their partner. The subjects are encouraged to experiment with the studied system without predefined test tasks, and they are disturbed or interrupted only if the discussion ends. O'Malley *et al.* recommend this method to be used when emphasis is on understanding or developing system concepts instead of learning procedures. (O'Malley *et al.* 1984)

Co-discovery learning shares many principles with constructive interaction, but in addition, has a list of specific tasks, and also includes a reflection on the task difficulty after each task. Sue Kennedy (1989) and her colleagues used this method in evaluating various communication systems by making two users cooperate and discover how to use a new system by trial and error. In co-discovery learning, the experimenter gives a list of tasks to the users, and leaves them performing the tasks on their own. After the users have completed a task, the experimenter asks them to jointly assess the difficulty of the task. According to the experiences of Kennedy, both the discovery and the jointly made assessments give valuable information on the usability problems of the system. (Kennedy 1989)

Paired-user testing has been considered as one way of making thinking aloud more natural for the users (Wildman 1995). As another benefit of paired-user testing, Wildman mentions the reduction of the moderator effect as the moderator is moved further from the test users. The studies by van den

Haak and de Jong (2005) support this remark, as their results show that the test users working as a pair pay less attention to the test moderator, and behave more securely than the users alone in a thinking aloud test. On the part of the efficiency of the paired-user testing, the studies by Hackman and Biers (1992) and Kallio and Kekäläinen (2004) ended up with quite contradictory results: Hackman and Biers (1992) detect the same amount of valuable statements in half the time in one paired-user test compared to two single user tests, whereas Kallio and Kekäläinen (2004) find less problems with paired users if the problems are counted by the number of users. The experiment by Shrimpton-Smith, Zaman and Geerts (2008) support the latter finding, as in their study couples find slightly less problems per user but significantly more problems per test session compared to single users.

Peer tutoring is also one possible way to make use of the natural interactions between two users. For example, Höysniemi, Hämäläinen and Turkki (2003) use peer tutoring to evaluate an interactive computer game with children. They have either a pair of children or one child at a time first learning to use the game, and then teaching it to another child. This way, the interaction between an adult and a child is minimised, and the ease of learning and teaching to use the game can be assessed in quite a natural way. Especially the part where the children act as tutors give a lot of information about the usability of the game, whereas the role of the tutee keeps the children pretty quiet and gives only little new information. (Höysniemi *et al.* 2003) Also in the study by van Kesteren *et al.* (2003), peer tutoring produces more verbal information from the 6-7 years old children compared to the co-discovery method that keeps the children pretty quiet and sometimes even leads to competitions between the test participants. However, peer-tutoring is applicable in assessing new systems as children are learning to use them, whereas co-discovery and paired-user testing can be used also with systems already in use.

4.1.1 Experiences with paired-user testing

We have applied paired-user testing with quite different types of systems including consumer electronics (Studies 21 and 61); mobile and office telephones (Studies 39 and 68); software systems for professional use (Studies 73, 100, 120, 127, 128, 131 and 132); a payment terminal (Study 107); and gaming slot machines (Studies 12, 65, 70, 76 and 77). We have tried to select pairs of users who already know each other to make the conversation and working together as natural as possible. The interaction between the participants is most fruitful if they are on the same level of knowledge and experience with similar systems. Therefore, we do not combine novices and experts as pairs, but try to find pairs with similar background knowledge.

In the test of an office telephone system (Study 39), we had pairs of secretaries using the telephone. This worked very well, because the secretaries knew each other very well and were used to working together. In addition, they had similar background knowledge of telephones, because their work required extensive use of telephones and their functionalities. The evaluated telephone was quite good, so it did not have any fundamental problems. Even so, it had

some flaws that slowed down the secretaries' work. These flaws came out very clearly in the conversations during the test, as well as some practical new features that the participants wanted to have.

Gaming slot machines have also been very natural objects for paired-user testing (Studies 12, 65, 70, 76 and 77), as people like to play them in pairs or even in larger groups. It is easy to find suitable users for these kinds of tests that do not require specific background knowledge, so we have usually combined both single user tests and paired-user tests to cover different kinds of social contexts.

However, paired-user testing does not work in all situations. For example, when we were evaluating a videotape recorder (Study 21), the tests were conducted in the apartment of a member of the evaluating team. The test users were his friends, so the environment was quite natural for the users, but the setting was too challenging for paired-user testing. Because the recorder was mainly operated through a small remote control, the second user did not have much opportunities to participate in the tasks, especially as the use of the manual was avoided to assess how intuitive the recorder was. Additionally, it was too tempting for this second user to quit the test when problems occurred, and move on to other chores in the apartment. However, considering the positive experiences with couples testing an interactive digital television in the experiment by Shrimpton-Smith *et al.* (2008), the reason for negative experiences in this study may lie more in having a too familiar moderator in the test than in having couples as test users.

Usually, the conversations between a pair of users in our tests have been more relaxed and natural than with single users supporting the findings by van den Haak and de Jong (2005). Also the moderator's role is considerably smaller in paired-user testing, since the participants usually work out the tasks together without the help of a moderator.

4.1.2 Experiences with peer tutoring

We have used peer tutoring as a way to evaluate the learnability of a system by studying how well the functionality is conveyed to the user, and what parts of the system have been left unclear. A typical setting for peer tutoring has included a third party entering the test room after the test user has finished the essential tasks alone (Studies 97 and 106). When studying a work related system, this third party has acted as a trainee who has just started her work (Study 106), and with a recreational system, the role has been of a relative or a friend asking for an advice in using the system (Study 97).

As test users have started to explain how to use the system, they have articulated more clearly their doubts and uncertainties in using the system than when performing alone. Comments like: "*I did not quite catch this, but…*" or "*I think you can do it here or there…*" are quite typical in peer tutoring indicating different levels of uncertainty in controlling the system.

4.2 Pluralistic usability walkthrough

This subchapter presents both the original version of the pluralistic usability walkthrough method by Bias (1994), and our modified version. The modifications have been presented also in my article in Ergonomics in Design (Riihiaho 2002) and my licentiate's thesis (2000).

4.2.1 Original pluralistic usability walkthrough

Pluralistic usability walkthrough (Bias 1994) is a usability evaluation method bringing representative users, system designers and usability experts together to evaluate and discuss on new design ideas. The discussion is based on tasks that the participants try to perform with the help of hard copy panels of the system. The participants get a set of hard copies of the dialogues that they need to perform the given tasks. Documentation or help functions are rarely available at this point, so the system designers usually serve as "living documentation", and answer questions that users indicate they would try to solve with the help of the system documentation. In this way, the users are able to carry on with their tasks, and the designers get valuable hints for their documentation. (Bias 1994)

Bias (1994) gives five defining characteristics of the method:

1. The method includes three types of participants in the same walkthrough session: users, system designers and usability experts.
2. The system is presented with hard copy panels, and these panels are presented in the same order as they would appear in the system.
3. All participants take the role of a user.
4. The participants write down the actions they would take to perform the given tasks.
5. The group discusses the solutions to which they have reached. The moderator first presents a correct answer. Then the users describe their solutions, and only after that, do the designers and usability experts offer their opinions.

The users in a pluralistic usability walkthrough session are representative users matching the system audience descriptions, the system designers may be architects, coders or writers, and a usability expert moderates the session. The moderator's role is to make sure the designers' attitude to the users' comments remains positive, because if the system designers try to explain away any problems, the users' willingness to give comments soon vanishes. (Bias 1994)

According to Bias, the pluralistic usability walkthrough method provides reliable data on a particular user interface panel in much the same way as traditional usability testing. The efficiency of the system, the user interface flow and navigation throughout the interface, on the other hand, cannot be reliably evaluated with this method. Compared to usability testing, pluralistic usability walkthrough is better in revealing uncertain decisions. In usability tests, these "lucky guesses" might go unnoticed, but in the pluralistic usability walkthrough sessions, the users can easily report that although they had the right solution, they were not sure about it. (Bias 1994)

4.2.2 Modified pluralistic usability walkthrough

We have made some major and minor modifications to the pluralistic usability walkthrough method in our studies in Aalto University. The most important change has been to keep the user testing and inspections separate from each other. In the procedure that Bias (1994) presents, the usability experts conduct usability inspections in the sessions while trying to perform the given tasks.

We want to conduct inspections and user testing separately for two main reasons. Firstly, if users are present in an evaluation session, we want them to feel important, and to let them be the focus of the session. Secondly, both the users' and the usability experts' time is saved if the inspections are conducted separately. The pluralistic usability walkthrough sessions usually last longer than a purely inspection session, so the sessions are less effective for the experts. If the experts comment on the system in the sessions very much, the users may feel left out, and feel their time is being wasted when they listen to comments with special terminology.

The minor changes we have made are mostly related to the separation of inspections and user testing. In our version, the sessions are very similar to usability testing in the sense that users and their comments are the focus of the session. All the other participants are already familiar with the system, and do not need to search for a solution for the tasks, but only to remind themselves of the solution. The usability experts' role is to moderate and observe the session, and the designers' role is to provide enough information about the system for the users to be able to comment on it.

Usually, we have only one moderator in a test session, but in a pluralistic usability walkthrough session, we have two moderators with slightly different roles: the main moderator concentrates on running the session with the predefined test tasks and questions, and the other one focuses on the discussions and ad hoc questions. Table 18 summarises our modifications.

Table 18: Modifications to the original pluralistic usability walkthrough method.

Original version (Bias 1994)	Our modification
User testing and usability inspections are combined in the same session.	Inspections are conducted separately, so user testing is the sole issue in the session.
Most of the participants are new to the system. They all take the users' role and try to perform the tasks.	Only the users are new to the system, so other participants just go through the tasks as a reminder of the necessary steps.
One moderator conducts the session.	A second moderator supports the main moderator by asking further questions.
Only one path through the tasks is available in the hard copy panels.	Several paths through the tasks are available so that the users can select a suitable one.
The moderator announces the right answer for the task before conversation takes place.	The users present their solutions from scratch, and finally the designers tell which solutions the system supports.

Bias (1994) names three limitations with the pluralistic usability walkthrough method:

1. The walkthrough must progress as slowly as the slowest participant on each panel,
2. Only one linear path is available in the paper prototypes, and
3. All the participants must conform to the selected path although their own solution might have been viable.

Our modifications to the method have helped us to diminish some of these problems. In our sessions, the users are the only participants really performing the tasks, and they usually solve the tasks quite quickly. If a user is especially slow, the moderator may help with small hints. In addition, the session is shorter, because the usability experts' comments are separated from the session. The problems 2 and 3 we have fixed by offering several alternative paths in the paper prototypes. Still, some paths may not be covered, if there are numerous alternatives.

4.2.3 Pluralistic usability walkthrough sessions

In our pluralistic usability walkthrough sessions, there are two or three users, one or two system designers and two usability experts present. One usability expert moderates the session and the other supports her by asking further questions, and by observing the users' reactions. In our studies, the system designers and the usability experts are already familiar with the system, so only the users represent novice users. Even so, we ask all the participants to take the role of a user to ensure a user-centred attitude in the session. This also encourages the designers to remind themselves of the steps necessary to perform the tasks.

We record the sessions with one or two video cameras. If we have a working prototype or demonstrations available in the session, the moderator may use those to give an overview of the general interface style. We try to place the computers out of the designers' reach, so that the designers are not tempted to demonstrate new features every now and then. Figure 13 presents a typical setting for our pluralistic walkthrough session.

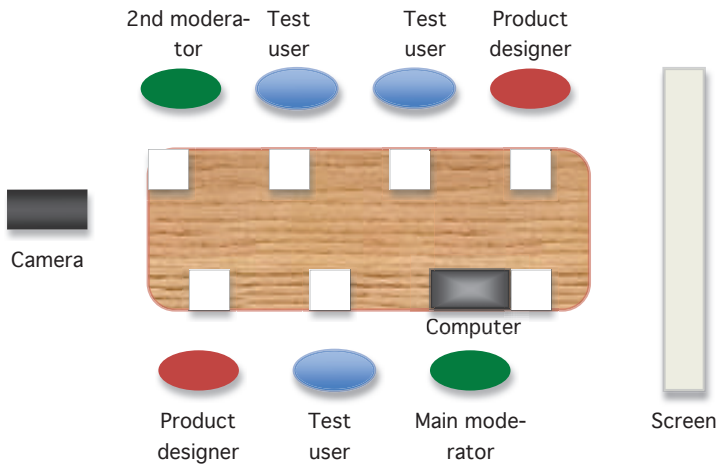


Figure 13: An example of a setting in a pluralistic usability walkthrough session.

In the beginning of the session, we explain the goal and the procedure of the session, and describe the context of use. Then we give the first task and the

related hard copy panels. As in usability testing, we start the session with an easy task to make the participants familiar with the procedure, and to make them relax. We ask the participants to write down on the panels all the actions they would take. For example, they may circle the buttons they would push or the menus they would select, and explain their actions. If the dialogues to be studied include selection lists, we present the lists in the same sheet of paper as the dialogue. The system menu is usually kept handy on a separate paper or on a slide that is visible during the whole session.

In the procedure that Bias (1994) presents, only one linear path through the task is presented in the hard copy panels. In our studies, we try to offer various paths, and let the users select a suitable one for them. This means that the users sometimes get quite a big pile of hard copy panels in a binder, and the moderators must help them to find the right ones according to their actions.

After all the participants have written down their solutions, the main moderator starts the conversation. Unlike in the procedure that Bias (1994) presents, our moderator does not present any "right" answer in the beginning. Instead, we want the users to present their solutions from scratch. If there is a silent user or a novice user in the group, the moderator usually starts the conversation with that user to ensure that the expert users or very talkative users do not rule the conversation. Only after all the users have commented on the task, are the system designers allowed to say which solutions the system supports. The designers usually suggest some new ideas for the system based on the users' comments, and all the participants are welcome to comment these ideas, and to generate new ones.

4.2.4 Applying pluralistic usability walkthrough

We conducted a pluralistic usability walkthrough session for the first time when we were evaluating the usability of an elevator monitoring and command system (Study 27). The system was designed to monitor and control the operation of an elevator and escalator bank for example in a department store. The system was aimed at caretakers working in the department stores, and at elevator service and maintenance men.

We were planning to conduct a traditional usability test for the system, and had already thought through the test tasks, and selected two test users. The day before the tests, it became clear that the system was not ready enough to be transferred in time for the tests. Therefore, we decided to conduct a walkthrough with paper prototypes. Since we had not been able to prepare ourselves for the use of the system, we needed the system designer to serve as a living documentation in the walkthrough. To encourage conversation, we decided to invite both the test users to the same session, although one of them was an expert user representing the elevator service and maintenance men, and the other one was a novice user representing the caretakers.

We wanted to get users' comments about three alternative design ideas. Two ideas could be presented via small computer-based demonstrations, but the newest design was only at the level of paper sketches. With each task, we first showed the related demonstrations and asked the users' opinion of them.

Then we gave the users the paper sketches of the newest design idea, and asked them to write down the actions they would take to perform the task.

Both the demonstrations and the paper sketches inspired vivid conversation: the toolbar presented in the paper prototypes proved to be effective, and the users suggested some new commands to be added there. The designer got lots of new ideas, and was very pleased with the effectiveness of the session. Although he seemed very motivated with his work already before the session, interaction with the users seemed to inspire him even more.

After some positive experience with pluralistic usability walkthrough, we were asked to evaluate the usability of a student register at our university. The project team developing the system had no extra resources for usability evaluation, but still, needed the results quickly. Therefore, we decided to conduct a pluralistic usability walkthrough session (Study 47). This way, we were able to get quick results without separate analysis and reporting, since the system designers attended the session and could see the problems for themselves.

In the session, there were two users, two system designers and two usability experts present. The users represented both ends of the use of the system: one user typed the results into the system, and the other used the system to transfer the results into the official register of our university. The system designers had already discussed with both of these users before, but not at the same time. Having representatives from both ends of the use of the system helped to reveal a function that was never needed. The function required a heavy database search, and was difficult to implement, so it saved a lot of time and effort for the system designers to find out that it was unnecessary in the current use. Since both ends of the use were represented, it was possible to confirm that the function was not needed in any part of the work, and could be left out.

4.2.5 Benefits of pluralistic usability walkthrough

It is difficult to find the moment in a system development process when the users are already able to comment the design ideas, and the system designers are still willing and able to make changes that reflect those comments. All too often, users are asked for their comments only when the system is almost ready. At this point, modifications are hard and expensive to make. Therefore, usability evaluations with the users are recommended already in the early phases of system development. Still, system designers often seem reluctant to make paper prototypes, and to hand the system over for user testing before it is finished. A method is therefore needed that motivates designers to evaluate their designs early enough.

It is easy, fast and inexpensive to make paper prototypes of the user interface. The prototypes make the system descriptions understandable to the users, so the users are able to comment on the system, and even to modify the design. Prototypes also help to uncover unarticulated aspects of users' work as users get a chance to "use" and envision the system (Bødker & Grønbæk 1991). Pluralistic usability walkthrough, thereby, gives the users an opportunity to evaluate how well the system supports their work activities, and to participate in the design of an improved system.

In our experience, pluralistic usability walkthrough is a very effective and useful method: with very little effort, the system designers get performance and satisfaction data straight from the users even before any functional prototype is available. As the system designers meet the users face-to-face, they easily put themselves in the users' place, and become very interested in the users' comments. The users sense this attitude, and are therefore, especially motivated to make comments on the system, and to suggest changes in the design. Furthermore, the sessions can be conducted in such an early phase of system development that it is still easy and inexpensive to make even major changes in the design.

4.2.6 Similar methods for multiple participants

Pluralistic usability walkthrough has features both from traditional usability testing and focus groups, as it brings several users together to discuss on the evaluated system. *Usability roundtables* (M.B. Butler 1996) also give a chance to invite a few users at the same time to present their work to the developers that are participating the same session. However, usability roundtables focus more on collecting user data for new ideas and improvements than evaluating a new design with the users.

Downey (2007), on the other hand, presents a study in which they invite 13 and 20 users at the same time to perform individually 2 basic tasks with the evaluated system. The tasks are given as written instructions, and the participants have a chance to ask for help from the observers walking around in the test room. Similarly, the observers can ask some brief questions from the users to get further information on specific issues. As the users are not thinking aloud during the task performance, and there is no room for thorough discussions with the users when they are doing the tasks, an additional hour is reserved for group discussions after the test. These group usability tests make it possible to collect information from several users in a short time with quite a few observers, and the group discussions after the tests allow more specific discussions in focus groups or workshops. (Downey 2007)

4.3 Visual walkthrough

Visual walkthrough is a user testing method that has been developed in our research group to get information about users' perceptions and interpretations of the evaluated user interface and its components (M.H.T. Nieminen & Koivunen 1995). The method can be used to complement a usability test or as a separate method. It is very similar to the *picture analysis* of screen images described by Dehlholm (1992), as they both go through the screen several times getting deeper into details after the first impressions.

During a visual walkthrough, the users are not allowed to explore the user interface, but to concentrate on one view at a time. At first, the users are asked to tell what they see and notice on the screen. After this general overview, the users are asked to describe what kind of elements, groups and details they notice. The next step is to ask the users to explain what they think the terms and

symbols mean, and what kind of functionalities the elements provide. After that, a task may be presented and the users are asked to state their intentions before any actions. (M.H.T. Nieminen & Koivunen 1995)

The method gives valuable information about the visibility of user interface elements and their understandability. However, focusing on one screen at a time, and limiting the use of the system may affect the users' learning (Trudel & Payne 1995), so this potential effect should be taken into account when using the method and interpreting its results.

4.3.1 Modification of visual walkthrough

We have used the visual walkthrough method also in a modified version that evaluates the utility of a system in addition to its usability. Several studies have brought out the problem that only few usability evaluation methods take utility into account (Johannessen & Hornbæk 2014), and leave only little room for assessing usefulness, value and evolving use of systems (Sengers & Gaver 2006; Greenberg & Buxton 2008). Also Cockton (2006) craves for methods assessing user experiences beyond interaction, and suggests using self-reporting methods for this purpose. The modification we have made to the visual walkthrough tries to tackle these challenges to some extent.

In the modified version, we ask tests users to mark in the given hard copies with different colours the parts that they need and use the most; parts that are rather important; and parts that they never use or that they find useless (Juurmaa *et al.* 2013). We used this method recently when we were evaluating an information system for healthcare professionals (Study 137). The preliminary interviews and usability inspections indicated that the system suffered from considerable information overload, so one of our goals in the study was to filter the most relevant information from the service. We also needed an easily comprehensible and convincing way to present the results to our customer.

To reach these goals, we complemented our usability test with this modification of the visual walkthrough method. For the presentation of the results, we integrated the results into coloured maps. As Choros and Muskala (2009), we combined some user interface elements into blocks and coloured the whole blocks either with one colour if the users' colourings were convergent, or with stripes if the colourings varied a lot. These aggregated block maps show the most relevant parts as green, whereas the red colour indicates blocks that the participants have used only rarely or have never used. Figure 14 shows an example of a hard copy with a user's markings, and an aggregated map showing the combined ratings for the various parts of the user interface.

In the walkthrough session, the hard copies were presented one by one to the users. Since the main page was already familiar to all of the test users, it could be walked through before the actual test tasks as a warm-up task to the test. The hard copies from the other parts of the service, however, were addressed only after the test tasks to avoid possible changes in the users' behaviour if they were forced to focus on the screens in detail before the test tasks.

Figure 14: On the left, a hard copy of the homepage with one test user's markings, and on the right, an integrated block map from the results of several test users (Study 137).

4.3.2 Requirements and benefits of visual walkthrough

In our study with the modified version of visual walkthrough, all the test users had some experience with medical issues, and most of them were already very familiar with the studied service. Therefore, it seemed rather effortless for the users to mark the relevant parts of the service, and to rule out the parts they considered useless both for themselves and their colleagues. On that account, it is recommended that the test users in visual walkthrough with utility assessment are experts on the domain, and familiar with the tasks that the service is intended for. With walk-up-and-use systems or other systems that are intended for anyone, the requirements are naturally less strict.

The method of colouring the elements in the hard copies proved to be simple and inexpensive, as it is applicable even with paper and pens. It is also fast, since the analysis of results can be done right after the tests in quite a straightforward procedure as long as a few basic criteria for combining the responses has been formed. In the walkthrough, the users are not given much time to study the system, so this method relies on the expertise that the users have on the domain area and their work. In a way, making the markings gives the users a chance to self-report their experiences and evolved use with the system, as Cockton (2006) suggests. We also found the coloured block maps to be valuable tools in communicating the results in a convincing and easily understandable format to our customers.

4.4 Informal walkthrough

Informal walkthrough is a mixture of usability testing, observation and interview. The method gives one tool to fix the bias of "*I know it can be done or you wouldn't have asked me to do it*" that Cordes (2001) among others has dem-

onstrated in his studies. To avoid this bias, informal walkthrough does not use predefined or specific test tasks. Instead, the test moderator has a list of features that the users are supposed to walk through in their own pace and order, as well as some scenarios that can be used in case the users do not spontaneously use the features in focus. The test users are asked to explore the system in the way they would if they were alone, and usually also to think aloud and comment the system while using it. Thereby, the method has similarities with the explorative approach in the constructive interaction technique by O'Malley *et al.* (1984). Indeed, the method can be applied both with single users or multiple users. The test moderator may interrupt the users for questions, but mainly just observes the session.

4.4.1 Applying informal walkthrough

We developed the informal walkthrough method when we were evaluating the usability of a news-on-demand service running on a television (Study 43) as a part of Mika P. Nieminen's Master's thesis (1996). At that time, the service was on the level of a functional prototype, and it could be operated with a remote control. The main goal of our evaluation was to study the intuitiveness of the service, and also to assess how spontaneously the users used the new features. Therefore, we did not want to give the users any predefined test tasks to indicate that they are doable with this service, but wanted them to find the features themselves. To get comments about the new features from all the users, we made a table with the functionalities that the system supported, and encircled the ones that we wanted every user to visit (Figure 15). As the users explored the system, the moderator marked the visited features with an "*X*" if the users found the feature themselves; with an "*A*" if they found it by accident; and with a "–" if the users did not try out the feature. After the users had explored the system, the moderator checked, if there were some features left that needed comments, and used some predefined scenarios or questions to guide the users to these features if necessary.

Functionality	User found (X/A/–)	Correct use	Needed help
Newspaper concept			
Navigation in table of contents	X	X	
Navigation in news	X	X	
Page numbers	A		
Background info	–		X

Figure 15: A checklist of the system functionalities. The features in focus are encircled.

We have used the informal walkthrough method for example in evaluating gaming slot machines (Studies 76, 77, 96 and 107), because the concept of the machines is already familiar to the users, and specific test tasks easily feel unnatural. Since the slot machines are quite often used in pairs and in groups, we usually have paired-user sessions in addition to single user tests. To ensure that all the users try a set of specific functions, we have prepared some scen-

arios leading the users into situations where these functions are needed. For example, the main scenario in Study 107 placed the users in a bus station waiting for a bus. The bus was told to be delayed, so the users had some spare time while waiting. After the users had played with the game for a while, they were told that the bus finally arrived. This meant that they had to finish their games, collect their wins, and possibly cash the money they did not have time to spend in gaming, but these subgoals were left for the users to decide and find.

4.4.2 Requirements and benefits of informal walkthrough

The informal walkthrough method requires a finished product or a functional prototype that the users can try out. The product concept must be familiar to the users so that they are able, willing and confident to explore the product. If the product is evaluated in such an early phase that there is only a partly functional prototype available, we recommend that the system developers participate or even moderate the tests as in Cooperative evaluation by P.C. Wright and Monk (1991b). As the method is informal, has qualitative focus, and can be conducted with less preparations than traditional usability testing, it also suits to agile development processes that according to Lárusdóttir, Cajander and Gulliksen (2013) emphasise informal methods offering qualitative data.

Since the users do not get any specific tasks, each session is unique. All the users have their own ways of learning and exploring new systems, so the test moderator needs to let all the users behave in their own way, and proceed in their own pace and order. This way, it is possible to get valuable information about how easy the system is to learn; which features the users find easily; which ones they try out first; and what features they still long for. Since the users are usually asked to think aloud and comment the system as much as possible, the method is not recommended for summative testing, but for assessing intuitiveness, ease of learning and ease of navigation.

Informal walkthrough method can be used both in the laboratory settings and in the fields. For example, researchers in Tampere have used informal walkthrough method in real use context involving participants with cognitive disabilities (Lepistö & Ovaska 2004). If the evaluated system is still on the level of low-fidelity prototypes and needs complicated settings, it is better to keep the testing environment controlled, and to focus on finding the most severe usability problems as well as the most enjoyable and rewarding features.

4.5 Contextual walkthrough

Contextual walkthrough is a method to evaluate the use of a system in a real use context. It includes elements from traditional usability testing and *Contextual Inquiry*, and shares the four main principles of Contextual Inquiry, namely: context, partnership, interpretation and focus. However, the goals of the methods differ, as Contextual Inquiry aims to *"get data about the structure of work practice*"; *"make unarticulated knowledge about work explicit*"; and *"get at the low-level details of work that have become habitual and invisible*" (Beyer & Holtzblatt 1998, pp. 37-38). The goal of the contextual walk-

through, in turn, is to reveal usability problems in the evaluated system, and to generate ideas for improving the system. Usability experts conduct these walkthroughs instead of designers, and the analysis of the results follow the conventions of usability testing rather than those of Contextual Inquiry.

Despite the differences in the goals of the methods, Contextual Inquiry provides a suitable structure for contextual walkthrough. Beyer and Holtzblatt name four phases in Contextual Inquiry (1998, pp. 64-66):

1. *The conventional interview* at the beginning to give room for the users to get used to the interviewer, and to explain the principles of the interview. This phase should take less than 15 minutes.
2. *The transition* to make it explicit that the contextual interview starts, and the user starts to be the master, and the interviewer the apprentice. This phase takes less than a minute, but is a necessary part to separate the contextual interview from the conventional interview.
3. *The contextual interview* in which the users do their daily work tasks and explain their doings.
4. *The wrap-up* ends the session, and includes a briefing of the interviewer's interpretations of the findings, so that the users can correct possible misunderstandings. This phase may take about 15 minutes.

Although the master-apprentice model used in the Contextual Inquiry (Beyer & Holtzblatt 1995) is not a compulsory part in contextual walkthrough, it helps in creating an appreciative rapport for open discussion and constructive criticism. It also helps in deciding, whether the user is appropriate for contextual walkthrough: if the user can take the role of a master, contextual walkthrough is a noteworthy alternative for a testing method.

4.5.1 Applying contextual walkthrough

We used the contextual walkthrough method for the first time when we were evaluating a new application to be used in a call centre (Study 59). The customer wanted us to study the efficiency of this application, and to measure the operators' performance times in the low-level actions in a call centre. However, as we went to the call centre and observed the operators' work, we soon realised that the performance times of the low-level actions in the application were irrelevant compared to the other tasks that the operators typically needed to do during the calls and especially after the calls. Therefore, we decided to observe the operators as they worked similarly to the Contextual Inquiry instead of measuring their exact performance times.

Clearly, the operators' work could not be interrupted as they talked with their customers on the phone. Therefore, we could not apply Contextual Inquiry as such. Instead, we observed the operators during the calls, made notes, and waited for a moment when the operator could explain the call and the actions required to finish the task, as Beyer and Holtzblatt recommend in similar situations (1998, p. 74-75). We recorded the sessions on videotape, but did not review the tapes with the operators as our notes were enough to support the discussions. At the end of the day, our team collected all the notes and found the remarks so unanimous that there was no need for additional, detailed an-

alysis or transliteration of the videotapes. The need for several applications and their physical locations in the call centre came as quite a surprise to our customer, and for its part, demonstrated the importance of contextual methods in developing and evaluating new systems.

We have used the contextual walkthrough method also in a student assignment evaluating a web service for applying building permits (Study 115). Contextual walkthrough was used to evaluate the service with expert users, whereas the novice users' performance was studied with traditional usability tests and predefined test tasks. The contextual walkthroughs were conducted at the expert users' workplaces, and the users were asked to use the web service in their own projects. Doing their normal work instead of faked projects made it easier for the users to assess the utility of the system, and to compare it with the traditional way of filling in paper forms, and delivering them to the officials in office hours. For the evaluators, these contextual walkthroughs helped them to understand the use requirements in practice, and thereby, helped in making realistic test tasks for the novice users, as well as in generating credible redesign proposals to improve the service.

4.5.2 Requirements and benefits of contextual walkthrough

When considering the use of contextual walkthrough, a few things have to be checked: the type of the system; users' expertise with similar systems; level of prototype; and access to the real use context. The systems to be evaluated with contextual walkthrough must be systems that the users are already using or will start using if the system is taken into use in their company. The users need something to try out with their own material, so the system must be on a level of a functional prototype or a running system. The users also need to be experienced enough to take the role of a master showing how their work is done.

As always when going into a company and observing users doing their work, it is necessary to get a permission to visit the users' workplaces, and often challenging to find suitable times for the visits. In many companies, the information that the users deal with is confidential, so it is important to agree on what sort of data can be used in the walkthrough; what may be recorded; and how it may be reported. On the other hand, contextual walkthrough gives an excellent opportunity to assess both the usability and utility of a system in a real use context.

4.5.3 Similar methods considering context

Similar methods taking the real use context into account have been developed also by Åborg *et al.* (2003); McDonald *et al.* (2006); Rosenbaum and Kantner (2007); and Savioja, Norros and Salo (2008). The method by Åborg *et al.* (2003) is called *ADA* as an acronym from Swedish term "användbara datorsystem" (usable computer systems). The main users of the method are experienced occupational health experts who visit companies and try to assess their work environments. The method is based on observations, interviews and questionnaires that are instructed in an interview guide. The occupational

health experts need a two-day tutorial to be able to use the method. (Åborg *et al.* 2003) The ADA approach evaluates the work environment and its computer systems as a whole focusing on ergonomics and cognitive requirements, whereas contextual walkthrough concentrates on assessing the usability of only one system in its real use context,

The *Rapid Contextual Evaluation* by McDonald *et al.* (2006) is very similar to the contextual walkthrough on the part of its process. The process is modified from the Rapid Contextual Design by Holtzblatt, Wendell and Wood (2005), but the phase of problem extraction and reporting is systematised by utilising a structured problem report format. This report format includes descriptions of the problems; the specific difficulties the problems will likely cause; the context in which the problems take place; and the assumed causes for the problems. The method helps to reveal a range of problems "*far beyond those that would be uncovered in laboratory testing*". As one more advantage of the method, McDonald *et al.* point out the possibility to iterate the understanding of the use context and intended value to the users when evaluating the system in the fields, thereby potentially helping in making applicable redesign proposals to fix the problems found. (McDonald *et al.* 2006)

The *field usability testing method* by Rosenbaum and Kantner (2007) builds upon Contextual Inquiry and traditional usability testing, similarly to the contextual walkthrough method. In this method, the tasks are designed to address the participants' own goals with the users' own files, bookmarks and databases, whereas in the contextual walkthrough, the tasks emerge from the context and real events. The field usability testing method has two approaches: the ethnographic model tries to get an insight into how people use a product when given high-level tasks; and the structured model "*more resembles a traditional usability test*", and is used when a realistic emulation in laboratory settings is not possible, or the users cannot or will not come to the laboratory (Rosenbaum & Kantner 2007). From these approaches, the ethnographic model is closer to the contextual walkthrough method.

Savioja *et al.* (2008), for their part, have developed an evaluation method called *Contextual Assessment of Systems Usability* (CASU) that aims to evaluate both the information content and the presentation of the information in new systems. They use real contexts in evaluating complex systems in nuclear power plant main control rooms similarly to contextual walkthrough, and they also leave certain degree of freedom to the definition of the tasks. However, for safety reasons, they use simulated process scenarios instead of authentic ones. (Savioja *et al.* 2008)

4.6 Informal and contextual walkthrough *vs.* other methods

Usability methods can be classified in numerous ways, such as: according to the goals, objects or locations of the studies; number of participants in a session; or the type and origin of the tasks to be studied. Ethnographic methods, for example, can be used to explore the patterns of life and culture to acquire better understanding of people's behaviour. These methods require longitudi-

nal studies, and usually imply only a few participants due to their profound approach to the subjects. Contextual Inquiry is originated from these ethnographic methods, but concentrates on the work in which the designed systems are to be used instead of concentrating on the employees' lives.

As Contextual Inquiry goes one level further in the specificity of the object of observation compared to ethnographic studies, contextual and informal walkthrough methods go even more specific: they concentrate on the usability of the system to be evaluated. In contextual walkthrough, the evaluation is made in the real context giving the opportunity to assess the role of the system in the daily work or tasks at hand. Informal walkthroughs, on the other hand, can be conducted both in laboratory settings and in the fields. Table 19 summarises these differences between various methods.

Table 19: Some methods for usability studies classified according to the goal and object of the study.

Goal of the study	Understanding	Requirements gathering		Usability Evaluation	
Object of the study	Life, culture	Work	Leisure time	Professional systems	Leisure time systems
Suitable methods	Ethnographic studies: observations, interviews	Contextual Inquiry	Probes, use diaries	Usability tests	
				Contextual walkthrough	Informal walkthrough

The usability evaluation methods also vary in the extent of resembling the real use context, and the duration of the studies. Observations and ethnographical studies, for example, are usually conducted as longitudinal studies observing the participants for a longer period in real use context. Contextual inquiry, however, lasts only a working day or less, so "boring routines" may be left out of that specific day. Contextual walkthrough and informal walkthrough in a real use context share this same constraint of seeing only small samples of the users' routines. Moreover, they focus only on the use of the evaluated system, instead of a work or a hobby as a whole. Finally, evaluations conducted in laboratory settings or with fixed test tasks miss many issues that real context and authentic tasks can reveal. Figure 16 shows some methods according to their analogy to the real use context and authenticity, placing observations and controlled usability tests in laboratories as the opposite extremes.

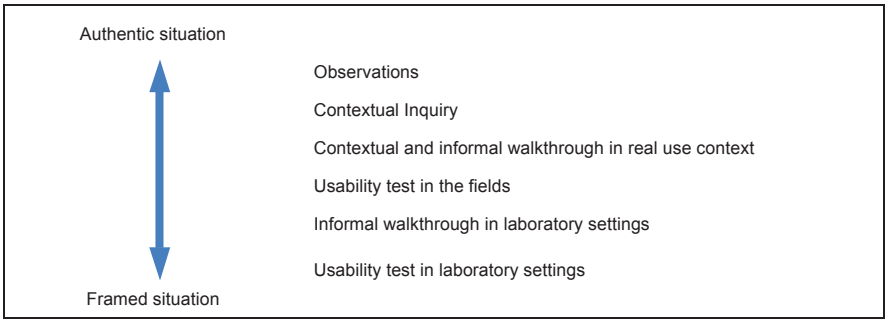


Figure 16: Some user research and usability evaluation methods according to their authenticity.

5 Experiment on thinking aloud and moderator presence

Methodological problems in usability testing have been brought forth already in early 1990's as Holleran (1991) lists several pitfalls in usability testing including a possible bias of the presence of a test moderator, and effects of thinking aloud. Even so, the effect of thinking aloud has been very little studied in the context of usability testing until Boren and Ramey (2000) approach the method from the professional communication point of view. Thereafter, there have been more and more studies of the effects of thinking aloud in its different versions, but the results of these studies are still somewhat conflicting leaving room for new studies. On the part of the moderator presence, the issue is very little studied in usability testing, and the studies by Held and Biers (1992) and Sonderegger and Sauer (2009) are among the very few ones to focus in this issue before the experiment in this thesis.

Using the thinking aloud method and having a moderator next to the user are common procedures both in our own studies at Aalto University, as the data of the 143 studies shows, and also in the Finnish usability and user experience companies, such as Adage and Cresense. The version of thinking aloud in these studies has been the concurrent relaxed thinking aloud emphasising the reasons for the users' actions. The studies by Boren and Ramey (2000) and Nørgaard and Hornbæk (2006), as well as the survey by McDonald *et al.* (2012) show that the relaxed version of concurrent thinking aloud is very common also in other countries. In the survey by McDonald *et al.* (2012), as much as 71% of the usability practitioner oriented respondents use a flexible approach, or always intervene when gathering thinking aloud data. Furthermore, the portion of those using a flexible approach was the highest in the group of the most experienced respondents (McDonald *et al.* 2012).

Based on the literature study and its conflicting results on the part of thinking aloud, and scarce results on the part of the moderator presence, my experiment focuses on the effects of concurrent relaxed thinking aloud and moderator presence, making them the independent variables in the experiment. The concurrent thinking aloud used in the experiment is quite interactive and relaxed allowing the moderator to ask clarifying questions on interesting issues between the test tasks. Still, interventions during the tasks are minimised. Thereby, the setting is very similar to the relaxed thinking aloud condition in the study by Hertzum *et al.* (2009) in which they compared the classic and relaxed versions of thinking aloud to performing in silence.

This chapter presents the experiment on the effects of the use of concurrent relaxed thinking aloud in usability testing as well as the effects of the presence of a test moderator near the test users. The overall process of this experiment has been very slow as the experiment took place in February 2010, the video analyses were made in November 2010, and finally, the statistical analyses were made in summer 2013.

5.1 Dependent variables and hypotheses

The dependent variables in this experiment include time on tasks, number of usability problem instances, user preferences, and the evaluators' certainty in their judgements. A usability problem in this study is defined as something that causes the user to distract from the optimal or workable sequence of actions; something considerably slowing down the performance; or something causing notable frustration. Thereby, cosmetic problems are not reported. Instead of the number of individual problems in the system, we are interested in the number of times a test user faces any usability problem. Therefore, the problem instances are not matched. In addition to these variables, the participants freeform comments about the test conditions are analysed.

To lay ground for the hypotheses, a brief summary of the effects found in similar studies including both concurrent thinking aloud and silent performance is presented in Table 20. The effects of moderator presence then, are presented in Table 21. On the part of the thinking aloud, the results are somewhat contradictory, and for the moderator presence, there are quite little results.

Table 20: Summary of the studied effects of using concurrent thinking aloud compared to silent performance in usability test settings.

Article	Effect of thinking aloud *vs.* silence	Specific
R.B. Wright & Converse 1992	Speeds up performance; Improves accuracy	Users requested to state reasons for every step
Eger *et al.* 2007	No effect on performance times; Negative effect on completion rates; Experienced as more unpleasant	
Hertzum *et al.* 2009	Slows down performance; No effect on accuracy	Both classic and relaxed thinking aloud compared to silence
McDonald & Petrie 2013	Classic instructions have no effect on performance; Explicit instructions increase navigation and scrolling	Compared classic and explicit instructions for thinking aloud with performing in silence
Hertzum & Holmegaard 2013	Time constraints have no effect; Has no effect on task times, but with visual interruptions improves task solution rate	Compared level 1 and 2 thinking aloud with silent performance using interruptions and time constraints

Table 21: Summary of studied effects of having a moderator present in a usability test setting.

Article	Effect of moderator presence	Specific
Held & Biers 1992	More negative feedback	Expert users
Eger *et al.* 2007	Thinking aloud feels more unpleasant and unnatural than working in silence, and the moderator presence feels more unpleasant when thinking aloud is required	Concurrent and retrospective thinking-aloud compared.
Sonderegger & Sauer 2009	Users alone rate emotions more positively; No effect on performance, perceived usability or attractiveness; Good rapport may enhance performance	Easy and more difficult tasks included

5.1.1 Hypotheses of task times and number of problems

According to the studies by R.B. Wright and Converse (1992), the concurrent thinking aloud should improve the test users' performance both by speeding up the performance and by making less errors. Also the study by McDonald *et al.* (2013a) supports an enhancing effect by explicit thinking aloud with difficult tasks, and the study by Hertzum and Holmegaard (2013) indicates that the task solution rates may improve if visual interruptions are used. However, many studies outside usability field suggest that thinking aloud slows down the performance (*e.g.* Gagné & Smith 1962; Ericsson & Simon 1980; Rhenius & Deffner 1990; Fox *et al.* 2011), and in the usability related studies, Hertzum *et al.* (2009) find both the classic and relaxed thinking aloud to slow down the performance, but not affecting the accuracy. Eger *et al.* (2009), on the other hand, find no effect on the performance times, but a negative effect on completion rates. In the study by McDonald and Petrie (2013) then, the explicit instructions for users to explain their likings, dislikings and confusions make them navigate and scroll more than in working silently. Therefore, the hypothesis about the effects of the concurrent relaxed thinking aloud with explicit instructions for users to explain their experiences is that the thinking aloud makes the users spend more time with the test tasks.

Although not from the field of usability testing, the results from the experiments by Davis *et al.* (1968), on the other hand, indicate that the presence of a test moderator could facilitate the performance (1968). Although the experiment by Sonderegger and Sauer (2009) show no significant differences in the test users' performance when working alone or with a moderator, also this study indicates a "*possibility that a facilitator who has established a good rapport with the test participant may represent a source of support with performance-enhancing effects*". Therefore, the hypotheses of task times are the following:

- H1: Time on tasks is longer in test conditions with thinking aloud (To and TM) compared to the silent conditions (oo and oM).
- H2: Time on tasks is shorter in test conditions with the moderator present (oM and TM) compared to working alone (oo and To).

On the part of the accuracy of the performance, the results of the studies are quite versatile, but still, quite many indicate improved accuracy both by the thinking aloud and the presence of a test moderator. Thereby, the hypotheses of the number of usability problems encountered are the following:

- H3: Users face less usability problems in using the system when thinking aloud is used (To and TM) compared to silent conditions (oo and oM).
- H4: Users face less usability problems in using the system when the moderator is present (oM and TM) compared to working alone (oo and To).

5.1.2 Hypotheses of test users' system preferences

According to Held and Biers (1992), expert users give more negative feedback on the system if they are required to explain the usability problems to a test moderator. With novice users, the effect is the opposite, but not significant. In the study by Sonderegger and Sauer (2009), on its part, the moderator presence has no effect on the users' preferences. The test users in the experiment in this thesis are closer to novices than experts as the system is new to everyone. Therefore, the hypothesis regarding the presence of a test moderator is that the effect is positive, meaning that users having the test moderator in the test room rate the system usability higher than the others.

On the part of the thinking aloud method, the literature presented in this thesis does not give corresponding results of the potential effect on the users' preferences. Therefore, a null hypothesis is made, stating that the thinking aloud does not affect the subjective ratings of the system.

H5: Novice users performing with the moderator present (0M and TM) rate the system preferences higher than those working alone (00 and T0).

H6: The use of the thinking aloud method does not affect the users' subjective ratings of the system (T0&TM *vs.* 00&0M).

5.1.3 Hypothesis of evaluators' certainty of their assessments

Even in normal test settings, the evaluators' judgements may vary substantially as the studies by Jacobsen *et al.* (1998); Molich *et al.* (2004); and Molich and Dumas (2008) show. Hertzum *et al.* (2014) study this evaluator effect also by comparing moderated and unmoderated test sessions, and find no differences in the detection rates or severity ratings between these two conditions, as they both have substantial disagreements between the evaluators.

In the experiment of this thesis, the evaluators have to base their judgements on different kind of information depending on the test condition, as half of the users explain their thoughts aloud, and half of the users mostly stay silent. Therefore, the evaluators are asked to assess also the certainty that they have on their judgments when analysing the tests. Their certainty is assessed on three judgments: if the user has any usability problem in a task; what kind of a problem it is; and what is the reason for the problem. The hypothesis in this experiment is that the evaluators' certainty of their assessments is lower when users perform silently than if they are thinking aloud, and the difference is most notable in the assessments regarding the cause of a problem.

H7: The evaluators' certainty of the causes of a usability problem are lower in conditions where users perform silently (00 and 0M) than in conditions with thinking aloud (T0 and TM).

5.1.4 Hypotheses of test users' feedback on test settings

Having a moderator next to the test user may make the thinking aloud more natural as there is someone listening to the user in the same room. Additionally, the setting gives better opportunities to ask clarifying questions related to

the tasks when the actions are still fresh in mind compared to discussing all the issues only after all the test tasks. Dumas and Loring (2008, p. 131) present very similar reasons for having the moderator in the same room instead of the setting commonly used in the early 1990's when the user and the moderator were separated to avoid moderator bias.

Thinking aloud is quite often used in Finnish usability tests, and also taught to the students as a relevant part of the testing. Therefore, it may be assumed that the test users expect the thinking aloud to be used also in this test. Therefore, performing in silence, especially when the moderator is beside the user, may feel even less natural than thinking aloud. Although Eger *et al.* (2007) find that test users consider the presence of a test moderator as more unpleasant and unnatural when they are thinking aloud concurrently compared to the retrospective settings, the relaxed version without systematic reminders to think aloud may affect differently. Thereby, the hypotheses of the relation of thinking aloud and moderator presence in this experiment are:

H8: Thinking aloud is rated as more natural in condition where the test moderator is in the same room (TM) compared to working alone (T0).

H9: Performing silently is rated more natural in condition where the test users are alone (00) compared to having the moderator present (0M).

5.2 Methodology

This subchapter presents the details of running the test sessions including the test participants and their pre-screening, the test tasks, and the test procedure.

5.2.1 Participants

The participants in the experiment were selected from the 107 students attending our bachelor level course on user-centred product development. Participating a usability test or some other experiment was a compulsory part of the course, so the students did not receive any other compensation for their participation. Nevertheless, five cinema tickets were drawn among the participants a week after the tests as a surprise.

The participants for the experiment were pre-screened with a background questionnaire according to their age; education; experience with smart phones; the manufacturer of their current mobile phone; and their attitude to new technologies and applications. The background questions are presented in English in Figure 17, and the original Finnish version in Appendix B.

The goal in pre-screening the participants was to get a homogeneous user group, and as similar groups for comparison as possible. For each test condition, 10 students was aimed for, so 42 students were selected to compensate potential losses. As 4 students missed the tests, the total number of test participants was 38 in the end. All the participants were students of engineering; had earned 180-220 credits in their studies; and were 20-30 years old (30 students of 20-25 years; and 8 students of 26-30 years). As in engineering stu-

dents, the participant group had more men (31) than women (7). Most of the participants currently had a mobile phone made by Nokia (32), and 6 participants had an Apple iPhone.

Study programme:
Computer Science / Information Technology / Communications Engineering /
Other engineering programme / Doctoral candidate / Other

Current amount of study credits:
<60 / 60-99 / 100-139 / 140-179 / 180-220 / >220

Gender: male / female

Age: <20 / 20-25 / 26-30 / >30

What kind of mobile phone do you currently have?
I do not have a mobile phone / Basic mobile phone without camera, e-mail or Internet/
Multimedia phone / Mobile phone for reading e-mails and web sites /
Smart phone in which I can load more applications

On what grounds did you select your current phone?

What is the brand and model of your current mobile phone? If you have more than one, list them all, and put the one that you use the most as the first one.

Which ones of the following activities do you do on a weekly basis with your mobile phone?
Talking
Short messages
Multimedia messages
Taking pictures
Listening to music
Watching videos
Calendar
Reading and sending e-mails
Web browsing
Other, what?

Which purchases have you paid with your mobile phone?
Bus, tram or subway ticket
Drinks, candies or other groceries from a vending machine
Parking ticket
Something else, what?

Assess your interests in new technologies using the scale 1-5, in which
1=totally disagree
5=totally agree

I am very interested in new technologies and applications utilising them.

I take new technological applications and products into use as soon as possible?

Are you ready to participate a usability test taking about an hour?
The test is video recorded for later analysis.

Figure 17: Pre-test questionnaire used to select the test participants.

Also the students' attitude and interests in new technology was studied in the background questionnaire to get homogeneous groups in all the test conditions. The students were asked to assess their interests in new technologies with a statement: "I am very interested in new technologies and applications utilising them", on a scale from 1 to 5, where 1= totally disagree, and 5=totally agree. All the selected participants gave ratings from 3 to 5 in this statement. Similarly, their eagerness to adapt new technologies was assessed with a statement: "I take new technological applications and products into use as soon as possible", and ratings between 2-5 were required to be selected. The requirements also included that the participants had prior experience with smart phones, and that they were ready to participate a usability test.

This rather homogeneous group had two distinct factors dividing them into two groups for both factors: gender and current phone. Since all the female participants had a Nokia phone, it was enough to divide them evenly into the test groups. Similarly, the three participants having exactly the same phone model as used in the experiment were divided into different test groups, as well as the six Apple iPhone owners. The selected participants also included two students whose mother tongue was not Finnish, but still spoke Finnish quite fluently. Since I prefer test users to be able to use their mother tongue when using thinking aloud, these participants were assigned to the groups working silently. Additionally, one participant who had attended several of our courses and who knew our staff better than the others, was assigned to the test condition of working alone to minimise the moderator effect. With these constraints, the subjects were randomly divided into the test condition groups.

5.2.2 Design

The study includes two independent variables, namely the use of concurrent relaxed thinking aloud, and the presence of a test moderator. Both of these variables have two variations, since the test users either think aloud while doing the tasks or perform silently, and the moderator is either present all the time or leaves the test room after the first warm-up task. Thereby, the test participants are divided into four groups presenting all the combinations of these variations. These groups are called T0 (thinking aloud, but no moderator); TM (thinking aloud, and moderator present); 0M (no thinking aloud, but moderator present); and 00 (no thinking aloud, and no moderator). Due to some students missing the tests, the groups 0M and TM have only 9 participants, and the groups 00 and T0 have 10 participants (Table 22).

Table 22: Test groups and conditions used in the experiment.

Test group	Thinking aloud	Moderator present	Number of participants	Description of condition
00	No	No	10	User silent and alone
0M	No	Yes	9	User silent with moderator present
T0	Yes	No	10	User thinking aloud alone
TM	Yes	Yes	9	User thinking aloud with moderator present

The experiment is made as a between-subject study to avoid learning effects and irreversible carry over of instructions from one test condition to another, especially in the case of having the silent performance after the relaxed thinking aloud. Furthermore, the real goal of the test is hidden from the users, so the usual goal of finding the most critical usability problems out of the assessed service is presented as the main objective to ensure that the participants act as they would in a normal formative usability test. Thereby, having the participants perform the same test tasks several times in different test settings as required for a within-subject study does not fit in this approach.

To control for individual factors, the participants were pre-screened, and a careful matching between groups was made to form as homogeneous and similar groups as possible. The group sizes in different conditions are also rather high compared to similar studies: 19 participants for thinking aloud *vs.* 19 per-

forming in silence; and 18 participants performing with the moderator next to them *vs.* 20 performing alone.

5.2.3 Evaluated system and recording

The tests were conducted in our usability laboratory in Aalto University. A very simple setting was selected to make it possible for the moderator to control the recording alone while conducting the tests. Therefore, only one camera was used, and it was placed next to the moderator. The screen of the mobile phone was duplicated to a monitor so that the moderator could see it both in the test room and from the observation room, and the monitor could be recorded with the video camera without distractions (Figure 18). Also the mobile phone and the user were included into the video image. The test moderator sat on the user's right hand side, slightly in front of the user, so that she could see the monitor. The user did not see the monitor to avoid focusing on wrong screen.

Figure 18: A separate monitor was used so that the moderator could see the user's actions with the phone, and get a good video recording even if the user moved the phone.

The evaluated system was the first prototype of a mobile ticketing service implemented in our research project by our cooperating company representatives. The system could be operated by a mobile phone, but the connections required for full functionality and money transfers were not yet implemented. The tests were conducted using a Nokia N95 mobile phone.

5.2.4 Scenario and test tasks

The prototype used in the evaluation was the first technical implementation of the system, so the user interface was not very considered. Hence, the scenario needed to explain various oddities, such as: why the test participants had to use a mobile phone that was not their own; why the phone already had some information that was not theirs; and why the participants had to use the sys-

tem on their own. The prototype was implemented in Tampere, so some local transportation companies were listed in the prototype, and they could not be removed. Therefore, Tampere was chosen as the use location in the scenario, although the tests were conducted in Espoo. The scenario is presented in English in Figure 19, and the original Finnish version is presented in Appendix D.

There is an interesting two day seminar on usability in Tampere. Your friend has given you a place to stay, so you have arrived to Tampere already a day before. You and your friend are trying to find out what is the best way to get to the seminar, and your friend recommends you to take a bus. She also recommends to use an eTicket as it is cheaper than the normal one-way ticket. She happens to have an extra phone with the application, and the phone is not much used because of its little strange keyboard, so you can borrow it. However, your friend wants you to use your own credit card information with the application. She is about to go to a party, but is still able to check that you can find the application before she leaves.

Figure 19: Scenario for the usability test of the eTicket service.

The test included six test tasks on separate papers. The first task mainly offered an opportunity to get familiar with the mobile phone and the test setting, including the thinking aloud method for half of the users. Thus, the recording started only after this first task. The test tasks were purposely selected to include both simple and more complicated tasks, but still to be done in about 20-30 min, so that the whole test would last less than an hour. To ensure that time does not run out, the last test task was designed so that it could be left out. The test tasks are presented in English in Figure 20, and the original Finnish tasks in Appendix D along with the test scenario.

1. The application needed is called eTicket. Try to find it from the mobile phone.
2. Give your own credit card information to the application so that you do not accidentally use your friend's money. Use the Visa Cardboard attached.
3. Buy a one-day ticket for the Tampere local bus so that you can use it tomorrow. After that, you need to prepare other things for the following day, so quit the application.
4. After a good night sleep, you are walking to the bus stop. The bus should arrive in a few minutes, so it is a good time to check that your ticket is ready to use and valid for travelling.
   Inside the bus, there is a scanner that reads the ticket from the phone as you hold it close enough. The application does not need to be open for scanning, but leave it open anyway for the trip.
5. The trip to the seminar takes a while, so you decide to buy a one-day ticket also for the following day. Therefore, you re-buy your ticket.
   Make sure that you received the new ticket.
6. The bus should arrive at half past. If you still have a few minutes left, remove your own credit card information from the application. If the bus is almost there, leave it to some other time.

Figure 20: Test tasks for the usability test of the eTicket service.

5.2.5 Post-test questionnaires

The post-test questionnaire had two parts: the first one focused on the usability of the evaluated system, and the second on the test setting. The first part consisted of 9 statements about the use of the system, such as: "*Overall, the use of eTicket was pleasant*", and the users rated their opinions on the scale from 1 to 5 (1= totally disagree, and 5= totally agree). There were also two open questions with freeform text fields at the end of the questionnaire. This first part was the same in all the test conditions. All the statements and questions are presented in Finnish with English translations in Appendix E.

The second questionnaire had nine claims in common to all the users, and some statements specific to the test setting. The common claims assessed how natural the setting felt; how pleasant it was to do the tasks; whether users tried

to perform the tasks as quickly as possible or as correctly as possible; and whether they did the tasks more patiently than usual. The original Finnish claims are in Appendix F, and Figure 21 shows the English translations.

- The test situation felt natural and I did not stress about my performance
- Doing the tasks was pleasant
- The scenario in the beginning of the test helped me to understand the tasks
- I understood easily what I was expected to do
- I tried to do the tasks as quickly as possible
- I tried to do the tasks as correctly as possible in the first try and to avoid mistakes
- I tried to get familiar with the whole system before starting the tasks
- I tried to find as many solutions as possible for each task before selecting my own solution
- I did the tasks more patiently than usual

Figure 21: Statements on the test setting common to all participants rated on scale 1-5 in which 1=totally disagree, and 5=totally agree.

After these shared claims, there were some statements specific to the test condition. For example, the users were asked to assess how natural it was to think aloud or to do the tasks in silence according to the test group. Similarly, there were statements related to the presence or absence of the test moderator. The statements specific to certain test conditions are presented in English in Figure 22, and the original Finnish versions are in Appendix F.

Thinking aloud
- Thinking aloud felt natural
- Thinking aloud affected to the extent to which I thought about the right way of doing the tasks

No thinking aloud
- Doing the tasks silently felt natural

Moderator present
- The presence of the test moderator disturbed my performance
- The presence of the test moderator encouraged me to finish all the tasks

Moderator in backroom
- I would have preferred to have the test moderator in the same room with me

Figure 22: Claims on test conditions (scale 1-5, 1=totally disagree, 5=totally agree).

5.2.6 Test procedure and given instructions

The experiment had two separate phases: the pre-test questionnaire used to select the participants, and the test session. The test session, on its part, had six phases: introduction, scenario, test tasks, brief interview, post-test questionnaire on eTicket, and a post-test questionnaire on the test (Table 23).

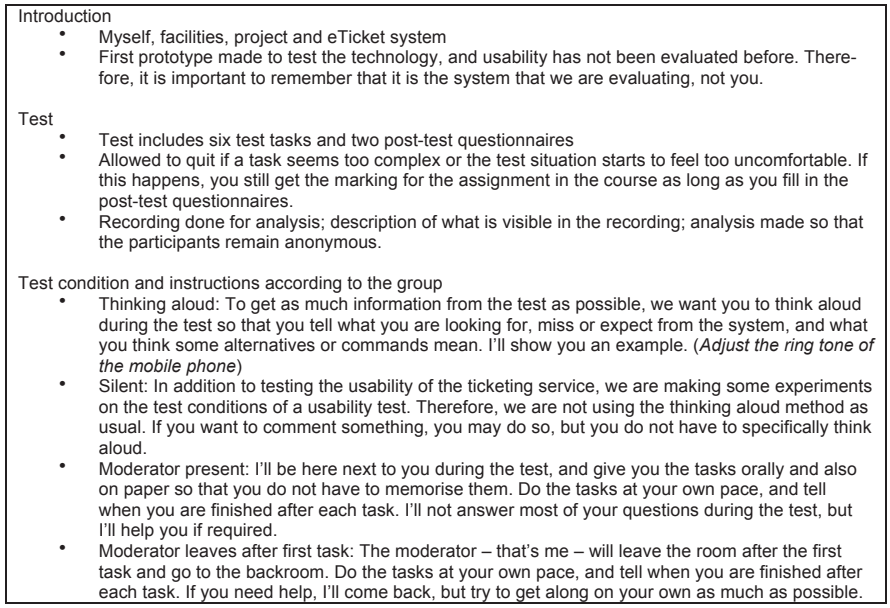
Table 23: Timetable for one test session.

Phase	Duration (min)
Introduction	10-15
Scenario	2
Test tasks	15-30
Interview	5
eTicket questionnaire	5-10
Test setting questionnaire	5-10
Total time	42-72

For each test session, an hour was reserved, but most of the participants completed the test in half the time, and only two participants finished the questionnaires on their own time. I moderated all the sessions to keep them as

similar as possible. As the test users arrived, I introduced the facilities and myself, and very briefly told about the research project in which the evaluated prototype had been constructed. Then, I told what the system was aimed for, and also told that the evaluated system was just the first prototype made to test the technology, so the usability had hardly been considered. Then, I reminded that it was the system that was evaluated, not the user.

After explaining the procedure of the test, I gave instructions about the test condition according to the test group. For those using the thinking aloud method, I showed an example of thinking aloud by adjusting the ring tone of the mobile phone used in the test. Thereby, the participant saw how to operate the phone if the model was not already familiar. The instructions for thinking aloud did not conform to the classic instructions, as users' expectations and interpretations were asked for instead of verbalising their thoughts as if they were alone. Thereby, the instructions were close to the explicit instructions in the studies by Zhao *et al.* (2014) and McDonald and Petrie (2013). The details of issues gone through with the test participants are listed in Figure 23. The original moderator's checklist is in Appendix C in Finnish.



Introduction

- Myself, facilities, project and eTicket system
- First prototype made to test the technology, and usability has not been evaluated before. Therefore, it is important to remember that it is the system that we are evaluating, not you.

Test

- Test includes six test tasks and two post-test questionnaires
- Allowed to quit if a task seems too complex or the test situation starts to feel too uncomfortable. If this happens, you still get the marking for the assignment in the course as long as you fill in the post-test questionnaires.
- Recording done for analysis; description of what is visible in the recording; analysis made so that the participants remain anonymous.

Test condition and instructions according to the group

- Thinking aloud: To get as much information from the test as possible, we want you to think aloud during the test so that you tell what you are looking for, miss or expect from the system, and what you think some alternatives or commands mean. I'll show you an example. (*Adjust the ring tone of the mobile phone*)
- Silent: In addition to testing the usability of the ticketing service, we are making some experiments on the test conditions of a usability test. Therefore, we are not using the thinking aloud method as usual. If you want to comment something, you may do so, but you do not have to specifically think aloud.
- Moderator present: I'll be here next to you during the test, and give you the tasks orally and also on paper so that you do not have to memorise them. Do the tasks at your own pace, and tell when you are finished after each task. I'll not answer most of your questions during the test, but I'll help you if required.
- Moderator leaves after first task: The moderator – that's me – will leave the room after the first task and go to the backroom. Do the tasks at your own pace, and tell when you are finished after each task. If you need help, I'll come back, but try to get along on your own as much as possible.

Figure 23: Moderator's checklist about the test instructions.

In psychological experiments, the real goal of the test is often hidden from the participants to avoid affecting the participants' behaviour. This was the initial plan also in this experiment, but as I had lectured in the course the students were attending about how to conduct usability tests, I needed to justify the differences between my lecture and the test on hand by revealing that there were also some other subgoals regarding the settings of a usability test. Even so, these subgoals were mentioned very briefly. However, for the group using the thinking aloud method and having the moderator present, *i.e.*, having the traditional usability test setting, these other subgoals were not mentioned.

As the moderator, I helped all the test users during the first task especially if the mobile phone was not already familiar. This procedure gave the participants some time to relax and to get used to the test situation, and also a chance to practice the thinking aloud method. This task was neither recorded nor analysed. After this first task, I reminded that the users should from there on try to solve the tasks on their own, and started the recording. In test conditions 00 and T0, I gave the rest of the test tasks to the user as a pile of small papers in the right order. The required credit card Visa Cardboard was attached to this pile of test tasks. Then I asked if there was something that the user still wanted to ask, and then, left the test room, and continued observing the session from the laboratory backroom. I could see and hear the test users and the monitor from the backroom, but had no way of giving audio feedback or probes to the users from the backroom. After the test users had finished all the test tasks, I returned to the test room, asked some clarifying questions on the user's actions, and instructed to fill in the post-test questionnaires.

In the sessions where I was in the test room all the time, I tried to keep my involvement minimal. I read the tasks aloud one at a time, and then placed the task description on a piece of paper on the table in front of the user. During the task performance, I acknowledged the users' comments by brief utterances, such as: "*OK*" or "*Mmm*", and gave some general hints. For example, I reminded that the PIN code needed in the system could be found from the backside of the credit card, if the user started to wonder about it. In the setting with silent performance, the users were allowed to comment and express their thoughts, so several users asked questions and expressed their doubts on proceeding with the action sequence. With silent setting, it was easier not to answer these questions, but in the traditional thinking aloud, it seemed more natural to answer some questions. For example, when one user asked very doubtfully if some information was really required to be able to complete a task, I told that it was not necessary, and marked the issue as a usability problem. However, I did not answer questions related to the user interface during the task performance, but got back to these questions after that specific task if these answers did not affect the following tasks. If the users ended up with a wrong solution, for example bought a train ticket instead of a bus ticket, or used the friend's credit card instead of their own, I asked questions, such as: "*Which ticket did you buy?"* or "*Whose credit card did you use?*" If the users could not find answers to these questions or were not concerned about those issues, we moved on with the task, and came back to these issues after all the test tasks. I did not remind the users about the thinking aloud if they fell silent, since quite often they had just explained what they were about to do.

I ran a pilot test a few days before the actual tests to check the clarity of the test tasks and questions in the questionnaires. A colleague from our research group acted as the pilot user. Although the use of colleagues is not recommended in pilot tests (Preece *et al.* 1994), the colleague was an outsider in the experiment, as he was not a member of the project in which the system was being developed, and had not taken part in the design of the experiment.

5.3 Analysis procedure

The material gathered in the experiment includes both quantitative and qualitative material, so various approaches were used to analyse the material and to determine the statistical significance of the results. This subchapter presents the principles and procedures in analysing the test data to study the dependent variables, and to test the hypotheses.

5.3.1 Analyses of video recordings and qualitative data

It was expected that there will be differences in the results of different evaluators even when they analyse the same video recordings, as in the studies by Jacobsen *et al.* (1998); Vermeeren *et al.* (2003); and Hertzum *et al.* (2014). Therefore, more than one evaluator was asked to analyse all the test sessions, but our limited resources allowed only two evaluators. Thereby, two colleagues from our usability research group analysed the video recordings of the experiment. Both of them had several years of experience in usability research, and they used a common Excel sheet that we designed together in a discussion before the video analyses. The evaluators made their analyses independently from each other in November 2010.

The free form answers in the post-test questionnaire were analysed only by me due to the small amount of these comments. Figure 24 summarises the issues analysed from the video recordings, and the scales used in rating the findings. After the evaluators had finished their video analyses, I merged the problem descriptions from their reports together with the problems from my own notes into a single list of usability problems.

- Time to start and finish a task
- Assessment, whether the user had problems with the task (yes/no)
- Severity of the problem (1= confuses or irritates the user / 2= user needs to correct the action, or the incident slows down the user's performance / 3= user cannot finish the task or ends up with a wrong result)
- Brief description of the problem (if there were several problems during one task, each of them was described and assessed separately)
- Cause for the problem (a brief description)
- Level of instruction or guidance from the moderator (0= none / 1= brief questions to raise thoughts / 2= clear instruction to overcome a problematic point / 3= instructed through the task)
- Other observations on the task
- Three assessments on the certainty that the evaluator has on the analysis related to the presence of the problem, the type of the problem, and its cause (1= not sure/ 2= fairly sure / 3= definite)

Figure 24: Issues analysed from the video recordings and the used ratings.

5.3.2 General principles in statistical analyses

Since the experiment included two independent variables, *i.e.*, thinking aloud and moderator presence, and it was a between-subjects study, a factorial analysis of variance (ANOVA) could be used to compare normally or approximately normally distributed data. ANOVA tests assume that the errors in data are independent of each other, and that these errors are identically and normally distributed, meaning that the variances in the compared data may not be very different, and the data may not be highly skewed. To analyse frequency counts, such as ratings on a Likert scale, a nonparametric chi-square test (χ^2)

could be used. The chi-square tests require the data points to be independent from each other, and the total sample size to be 20 or more. (Lazar *et al.* 2010, pp. 75-94)

The following subchapters present the selections for samples and analyses to test the hypotheses. On the part of the hypotheses, the probability of getting the effect by chance should be less than 5% to be able to call the results statistically significant (Lazar *et al.* 2010, p. 34). The statistical analyses were made with Microsoft Excel and IBM SPSS Statistics, version 21 in summer 2013.

5.3.3 Analyses of numerical measures

The experiment had two numerical measurements as dependent variables: time on task, and number of usability problems. Although the experiment included six test tasks, statistical analyses were made to only one of them due to very limited analysis resources, and the type of the test tasks. The first test task was a warm-up task; the second task was a rather complex task requiring the users to enter the user information into the system; and the rest of the tasks were somewhat simpler including buying new tickets and activating them. As the goal in the second task was clear to all the users, but still had some challenges to be solved, it was selected as the sample for analysing the task times and the number of usability problems that the users faced. The same task was also used in analysing the evaluators' certainty on their judgments.

The performance times for the analyses were taken from the notes I had made during the tests, and checked from the recordings. One user's performance in group T0 had to be left out from these analyses, because he entered the required information only when doing the next task. Thereby, the number of participants in groups 0M, T0 and TM was 9 in these analyses, and only group 00 had 10 participants. Although performance times are not normally distributed as they are all positive, they approximate to normal distribution well enough for the statistical analysis (Figure 25), so the factorial analysis of variance (ANOVA) could be used.

On the part of the number of usability problems faced by the test users, the two evaluators ended up with quite different results, supporting the findings of the evaluator effect by Jacobsen *et al.* (1998); Vermeeren *et al.* (2003); and Hertzum *et al.* (2014). For example, in the analysed test task, the 1st evaluator reported from 1 to 9 problems per participant, and 151 problem instances in total, whereas the 2nd evaluator reported from 1 to 5 problems per participant, and 84 instances of usability problems in total. Also the severity ratings differed substantially, as the 1st evaluator rated about 7 % of the problems as level 3 problems that lead the test user to a wrong task solution, whereas the 2nd evaluator rated only 1 problem, *i.e.*, about 1% of the problems into this level. Thereby, the number of problems found was analysed separately for both the evaluators, as well as the certainty of the causes. Similarly to the task times, the number of usability problems is a positive number, but also this data approximates to normal distribution, so factorial analysis could be made.

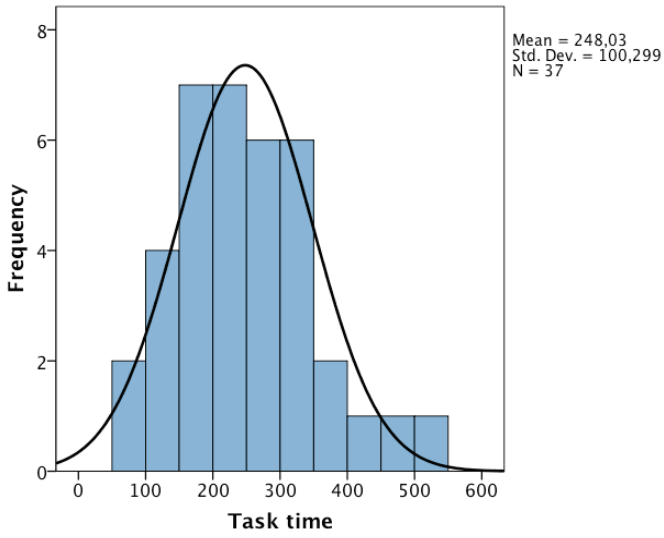


Figure 25: Distribution of time tasks as a histogram and an approximate to normal distribution.

5.3.4 Analyses of ratings and freeform comments

The dependent variables include subjective ratings of the evaluated system and the test setting, and the certainty the evaluators have on their judgements on the causes of the usability problems. The test users' assessments about the usability of the eTicket service were compared between the different test settings based on the value that they gave in the post-test questionnaire to the statement: "*Overall, the use of eTicket was pleasant*". In this assessment, a Likert scale from 1 to 5 (1= totally disagree, and 5= totally agree) was used. As the alternatives for answers were not linearly distributed but categorised ordinally according to the users' agreement, a nonparametric statistical chi-square test (χ^2) was used to analyse the data.

The analyses on the naturalness of thinking aloud and working in silence were done very similarly, but they had to be analysed within the groups using the thinking aloud method (To *vs.* TM), and within the groups performing silently (OO *vs.* OM). The test users' ratings from 1 to 5 (1= totally disagree, 5= totally agree) in the post-test questionnaire to statements: "*Thinking aloud felt natural*", and "*Doing the tasks silently felt natural*", were used as a basis for these comparisons. Although the total sample size in the experiment was 38 participants, these comparisons were made in smaller groups, so the expected count for various ratings became so low that a chi-square test alone did not necessarily give accurate results. Therefore, Fisher's exact test was used to confirm the results of the chi-square test in these analyses.

To give more information about the test users' feelings about the naturalness of the test setting in different test conditions, also the ratings given to the statement: "*The test situation felt natural, and I did not stress about my performance*", were analysed. This statement was common to all the groups, and

the scale was again from 1 to 5. All the ratings for the overall naturalness of the test situation were very positive and ranged between 3 and 5, so the degree of freedom was only 2 in this chi-square test.

The certainty of the evaluators' judgements was analysed from the ratings the evaluators gave when considering their certainty of the cause of a problem that a user had encountered. The scale used in this analysis was from 1 to 3, where 1=not sure, 2=fairly sure, and 3=definite. As in the analysis of users' performance, only the assessments made on the second test task were analysed. The two evaluators used the scale of certainty in very different ways, so their assessments were analysed separately. Again, the video analyses of one test user in group T0 had to be left out, because the user did some of the required actions only in the next task. Similarly to the test users' preferences, these ratings were ordinal but not linear, so the chi-square test was used.

To further analyse the test participants' feelings about the test setting, a qualitative analysis was made to the freeform comments given in the post-test questionnaire. Comments on thinking aloud and working is silence; the naturalness and pleasantness of the test setting; and the presence or absence of the moderator were gathered, and categorised into positive and negative comments. For example, a test user's comment: "*Thinking aloud felt unnatural at first, but was no longer a problem at the end*", was marked only as a negative comment, because the end situation is only neutral without positive emphasis. As another example, a comment about forgetting to think aloud was marked as a negative comment, because forgetting is a negative feeling, and because this user strongly disagreed with the statement: "*Thinking aloud felt natural*".

5.4 Results of experiment

This subchapter presents the results of the analyses by going through the hypotheses one by one. In addition to the results of the statistical analyses, the data is presented in several other ways to give an overview of the data.

5.4.1 Time on tasks

The performance times on the 2nd test task ranged from 86 s to 519 s. The two quickest times were made without thinking aloud but having the moderator present (0M), whereas the longest task time was made with thinking aloud but no moderator present (T0). Table 24 shows the average tasks times in the second test task, as well as the standard deviations and minimum and maximum values according to the use of thinking aloud and the moderator presence.

Table 24: Average task times, standard deviations and the range of task times in the second test task according to the test conditions.

Test condition	Average task time (s)	Standard deviation (s)	Range (min-max s)
No thinking aloud (N=19)	216	100	86-472
Thinking aloud (N=18)	282	92	140-519
No moderator (N=19)	257	104	114-519
Moderator present (N=18)	239	99	86-419

The results of the factorial ANOVA test for the task times in different test settings are presented in Table 25. The test shows that the thinking aloud has a significant effect in performance times ($F(1, 33)=4{,}16$, $p<0{,}05$) with a medium effect size of eta-squared (η^2) 0,112. The moderator's presence does not have a significant effect, and there is no significant interaction effect between the two independent variables. Thereby, the hypothesis H1 of thinking aloud slowing down the performance is supported by the results, but the hypothesis H2 of moderator presence speeding up the performance is not supported.

Table 25: The results of the factorial ANOVA test for the effects of thinking aloud and moderator presence on performance times in the 2nd task (N=37, $df_1=1$, $df_2=33$).

Source	Mean square	F	Significance
Thinking aloud (T)	40148,331	4,161	0,049
Moderator presence (M)	3588,433	0,372	0,546
T * M	812,741	0,084	0,773

5.4.2 Number of usability problems

The hypotheses on the number of usability problems concern the number of times users face usability problems when doing the tasks. Therefore, the instances of usability problems are first analysed, and only after that the number of distinct problems, and the evaluators' any-two agreement. The number of usability problem instances in different test conditions is shown in Table 26 on the part of the 1st evaluator and in Table 27 on the part of the 2nd evaluator. Table 28 and Table 29 show the means and standard deviations accordingly.

Table 26: Cross tabulation of usability problem instances detected by the 1st evaluator.

Test condition	No moderator	Moderator present	Total
No thinking aloud	36 (N=10)	36 (N=9)	72 (N=19)
Thinking aloud	38 (N=9)	41 (N=9)	79 (N=18)
Total	74 (N=19)	77 (N=18)	151 (N=37)

Table 27: Cross tabulation of usability problem instances detected by the 2nd evaluator.

Test condition	No moderator	Moderator present	Total
No thinking aloud	20 (N=10)	19 (N=9)	39 (N=19)
Thinking aloud	20 (N=9)	25 (N=9)	45 (N=18)
Total	40 (N=19)	44 (N=18)	84 (N=37)

Table 28: Average number of problem instances indicated by the 1st evaluator, standard deviations and the minimum and maximum values in the 2nd test task.

Test condition	Average number of problems	Standard deviation	Range (min-max)
No thinking aloud (N=19)	3,789	1,228	2-6
Thinking aloud (N=18)	4,389	1,944	2-9
No moderator (N=19)	3,895	1,560	2-8
Moderator present (N=18)	4,278	1,708	1-9

Table 29: Average number of problem instances indicated by the 2nd evaluator, standard deviations and the minimum and maximum values in the 2nd test task.

Test condition	Average number of problems	Standard deviation	Range (min-max)
No thinking aloud (N=19)	2,053	1,026	1-4
Thinking aloud (N=18)	2,500	0,985	1-5
No moderator (N=19)	2,105	0,994	1-4
Moderator present (N=18)	2,444	1,042	1-5

The results of the factorial ANOVA tests for the number of usability problem instances indicated by the 1st and 2nd evaluator in different test settings are presented in Table 30. Although the evaluators assessed the sessions quite differently, neither the thinking aloud nor the moderator presence had a significant effect on the number of problem instances, and there was no significant interaction effect between these variables, either. Thereby, the hypotheses H3 and H4 on thinking aloud and moderator presence improving the users' performance and lessening their usability problems are not supported.

Table 30: The results of the factorial ANOVA test for the number of problems faced in the 2nd task (N=37, df_1=1, df_2=33).

Source	F (1st)	Significance (1st)	F (2nd)	Significance (2nd)
Thinking aloud (T)	1,171	0,287	1,770	0,193
Moderator presence (M)	0,454	0,505	0,995	0,326
T * M	0,004	0,952	0,442	0,511

The total number of usability problems uncovered from the eTicket system in the 2nd test task was 17. The most severe and common problems were related to the unclear menu structure with poor terminology leading the users to wrong selections; defunct selection button; problems in typing with the keyboard; and confusions in the relations of the travel accounts and the credit cards. After these, the most common problems concerned the ambiguous prompt "Name:" when giving the credit card information; the need to open information before being able to edit it; too easy exit from the service that broke off many users' tasks; need to remove default values before entering new information; and poor visibility of the credit cards in the service. All the problems identified during the 2nd test task are listed in Figure 26.

1. Unclear menu structure with poor terminology
2. The selection button did not work
3. Difficulties in entering information into the system
4. Unclear relations of the credit cards and the travel accounts
5. Unclear which name should be given when the credit card information is entered
6. All the information need to be opened before being able to edit it
7. Users exit the system too easily by mistake
8. Default values need to be removed before entering new data
9. User's credit cards hard to find from the lists
10. Insecurity in entering PIN code especially without system hiding it
11. Uncertainty on which credit card is used
12. Removing text is difficult, and the button C did not delete text
13. No spaces were allowed between the numbers when giving the credit card number
14. Difficulties in browsing through the lists in the system
15. The cursor is hard to see when entering information
16. Uncertainty on how long the travel ticket is valid
17. Too little information visible on the list of travel tickets

Figure 26: List of problems uncovered through the 2nd test task.

For counting the any-two agreement of the evaluators, both the union and the intersection of the problems identified by the evaluators needs to be known. Table 31 shows this information, and also shows in which test settings the problems were identified by each evaluator. Evaluator A is the author, and these findings are based on the notes made during the test sessions; evaluator B is the 1st outside evaluator; and evaluator C is the 2nd outside evaluator. The

numbers after the letters indicate the test group (1=00, 2=0M, 3=T0 and 4=TM).

Table 31: Cross tabulation of the usability problems (UP) found by evaluators A (author), B and C in different test settings (groups 1, 2, 3 and 4), and the total number of problems they uncovered.

UP	A1	A2	A3	A4	A	B1	B2	B3	B4	B	C1	C2	C3	C4	C
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X	X	X	X	X	X		X	X
5	X	X	X	X	X	X	X	X		X		X	X		X
6	X		X	X	X	X	X	X	X	X	X		X		X
7	X	X	X	X	X	X	X	X	X	X					
8	X	X	X	X	X		X	X		X				X	X
9	X		X	X	X	X		X	X	X				X	X
10	X			X	X	X			X	X	X			X	X
11	X	X		X	X	X			X	X	X				X
12	X		X		X	X	X			X	X				X
13				X	X				X	X					
14							X		X	X					
15			X		X			X		X					
16				X	X										
17									X	X					
Total	12	8	11	13	15	11	10	10	12	16	8	5	5	7	11

The evaluator effect could be distinguished also in this analysis, as evaluators A and B found 15 and 16 problems, but the evaluator C only 11 problems. Out of the total of 17 problems, 3 problems were detected only by 2 evaluators, and 3 problems just by 1 evaluator. Still, the any-two agreements of the evaluators was as good as 82%, 73% and 69% making the average over all the pairs 75%. Furthermore, the problems were uncovered quite evenly among different test settings.

5.4.3 Test users' system preferences

The users rated their system preferences with a Likert scale from 1-5 (1=totally disagree, 5=totally agree) for the statement: *"Overall, the use of eTicket was pleasant*". The answers in the different test conditions distributed as shown in Table 32, and the results of the chi-square tests are presented in Table 33.

Table 32: Test users' overall ratings of the pleasantness of the evaluated system with a scale from 1 to 5 (1=totally disagree, 5=totally agree).

Group/rate	1	2	3	4	5
00 (N=10)	1	5	2	2	0
0M (N=9)	1	2	3	3	0
T0 (N=10)	0	7	2	1	0
TM (N=9)	2	0	4	2	1

Table 33: The results of the Pearson chi-square tests on the effects of thinking aloud and moderator presence on the users' subjective ratings of the system (N=38, df=4).

Independent variable	$\chi^2(4)$	Asymptotic significance (2-sided)
Thinking aloud (T)	1,591	0,810
Moderator presence (M)	10,385	0,034

The results indicate that the presence of a moderator has a significant effect on the users' system preferences ($\chi^2(4)$= 10,385, p<0,05) with a small effect size

of eta-squared (η^2) 0,061. Also Figure 27 shows that the users working alone mostly disagree with the statement: "*Overall, the use of eTicket was pleasant*", whereas users having the moderator in the test room are more neutral or agree with the statement. Thereby, the analyses support the hypothesis H5 claiming that users with a moderator present give better subjective ratings of the system. They also support the hypothesis H6 about thinking aloud having no effect on the subjective ratings.

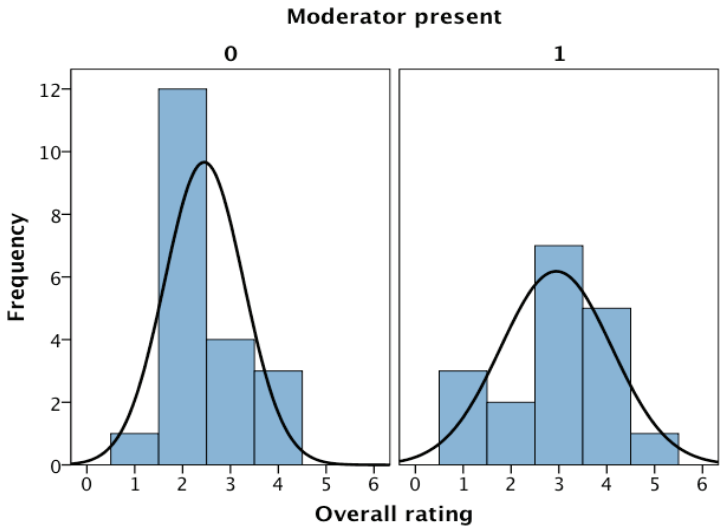


Figure 27: The distributions of the users' overall ratings of the evaluated system. On the left, users performing alone, and on the right, users with the moderator.

5.4.4 Evaluators' certainty of their assessments

The evaluators also assessed their certainty of the causes of the usability problems they found on a scale from 1 to 3 (1=not sure, 2=fairly sure, and 3=definite). Since the evaluators used the scale very differently, their assessments are analysed separately. The distributions of their assessments are presented in Table 34 (1st evaluator) and Table 35 (2nd evaluator).

Table 34: The frequency counts of the 1st evaluator's assessments on his certainty of the causes of the usability problems in different test groups.

Group/ Assessment	Not sure	Fairly sure	Definite	Total
00 (N=10)	0	4	16	20
0M (N=9)	0	9	11	20
T0 (N=9)	0	3	18	21
TM (N=9)	0	6	19	25
Total	0	22	64	86

Table 35: The frequency counts of the 2nd evaluator's assessments on his certainty of the causes of the usability problems in different test groups.

Group/ Assessment	Not sure	Fairly sure	Definite	Total
00 (N=10)	6	20	10	36
0M (N=9)	5	17	14	36
T0 (N=9)	7	18	13	38
TM (N=9)	6	25	10	41
Total	24	80	47	151

The Pearson chi-square test was used to analyse the distributions of the assessments and their relationship to the independent variables. The 1st evaluator used the scale only from 2 to 3, so the degree of freedom for the 1st evaluator is 1, and 2 for the 2nd evaluator. The results are presented in Table 36.

Table 36: The results of the Pearson chi-square tests of the evaluators' certainty of the causes of the detected usability problems.

Evaluator and test condition		Value	Asymptotic significance (2-sided)
1st evaluator	Thinking aloud	1,880	0,170
(N=86, df=1)	Moderator presence	2,979	0,084
2nd evaluator	Thinking aloud	0,314	0,855
(N=151, df=2)	Moderator presence	0,328	0,849

The results suggest that neither the thinking aloud nor the presence of the moderator has significant effect on the certainty that the evaluators have on the causes of the detected usability problems. Thereby, the hypothesis H7 presuming that the silent performance decreases the certainty of causes for the usability problems is not supported by the results.

5.4.5 Test users' feedback on test settings

The test users assessed various aspects about the test settings in the second post-test questionnaire. Participants using the thinking aloud method assessed how natural the thinking aloud felt, and the others assessed the naturalness of performing silently. The statements were: "*Thinking aloud felt natural*", and "*Doing the tasks silently felt natural*". The scale used was from 1 to 5 (1=totally disagree, 5=totally agree). Table 37 shows the cross tabulation of the ratings when thinking aloud, and Table 38 when performing in silence.

Table 37: Frequency counts of test users' assessments on the naturalness of thinking aloud in scale 1-5 (1=totally disagree, 5=totally agree).

Test condition / Rating	1	2	3	4	5	Total
No moderator (00)	1	1	3	3	2	10
Moderator present (0M)	0	2	3	4	0	9
Total	1	3	6	7	2	19

Table 38: Frequency counts of test users' assessments on the naturalness of performing in silence in scale 1-5 (1=totally disagree, 5=totally agree).

Test condition / Rating	1	2	3	4	5	Total
No moderator (00)	1	0	0	3	6	10
Moderator present (0M)	0	2	1	4	2	9
Total	1	2	1	7	8	19

The Pearson chi-square test was used to analyse the effects of the moderator presence on the naturalness of performing in silence, and the naturalness of using thinking aloud. These results are presented in Table 39.

Table 39: The results of the Pearson chi-square test on the effect of the moderator presence on the users' assessments of working silently or with thinking aloud (df=4).

Group	$\chi^2(4)$	Asymptotic significance (2-sided)
Working silently (N=19)	6,107	0,191
Thinking aloud (N=19)	3,433	0,488

Since the comparisons are made within groups smaller than 20 users, the Fisher's exact test was used to confirm the results (Table 40). Also these results suggest that the presence of the moderator has no significant effect on the users' assessments on the naturalness of working silently or with thinking aloud.

Table 40: The results of the Fisher's exact test on the effect of the moderator presence on the users' assessments of the naturalness of working silently or with thinking aloud (df=4).

Group	Value	Exact significance (2-sided)
Working silently (N=19)	5,523	0,138
Thinking aloud (N=19)	3,188	0,700

If the test users' estimations for the naturalness of working silently or with thinking aloud are treated as linear values from 1 to 5 to demonstrate the differences, the average ratings in groups working silently are 4,3 when performing alone, and 3,7 with the moderator present. For the thinking aloud, the average ratings are 3,4 when performing alone, and 3,2 with the moderator present. Although the difference in the users' assessments on the naturalness of working in silence is to the direction that the hypothesis H9 predicts, *i.e.*, silence is more natural when working alone, the difference is not statistically significant. On the part of the naturalness of the thinking aloud, even the direction of the difference is opposite to the hypothesis H8 predicting that thinking aloud would be more natural when the moderator is present.

The naturalness of the test setting was assessed also by analysing the users' ratings to the statement: "*The test situation felt natural, and I did not stress about my performance*". The frequency counts of these ratings are presented in Table 41 along with means for the ratings when simplifying them into numerical values for demonstration. The results of the Pearson chi-square tests show that neither the presence of a test moderator ($\chi^2(2)=1,629$, $p=0,443$) nor the thinking aloud ($\chi^2(2)=1,259$, $p=0,533$) has significant effect to the perceived naturalness of the test setting.

Table 41: Frequency counts of the test users' ratings on the naturalness of the test setting in a scale from 1 to 5 (1=totally disagree, 5=totally agree).

Group / Rating	1	2	3	4	5	Mean
00 (N=10)	0	0	0	5	5	4,5
0M (N=9)	0	0	0	3	6	4,7
T0 (N=10)	0	0	1	5	4	4,3
TM (N=9)	0	0	0	4	5	4,6
Total (N=38)	0	0	1	17	20	4,5

In addition to the numerical estimates in the post-test questionnaire, the test users' freeform comments about the test conditions were analysed. Comments about the thinking aloud; working in silence; effect of the moderator presence, and the naturalness and pleasantness of the test setting were searched for, and categorised into positive and negative comments. These freeform comments about the test settings are presented in Figure 28.

Thinking aloud
+ "*It was refreshing to think and justify aloud one's own behaviour*"
- "*Thinking aloud seemed to be left in the background…*"
- "*At first, the talking might have felt a bit strange, but as one is used to explain to colleagues what one is trying to do, talking to a camera was not a problem in the end*"
- "*Thinking aloud did not feel natural, as I do not normally talk to myself when I'm using an apparatus*"

No thinking aloud
+ "*I liked to work in silence*"
+ "*When there was no need to talk in the test situation, the usage felt natural – the way I would imagine the usage to be if I tested the system on my own*"
+ "*The test was considerably more comfortable than a traditional "thinking aloud" method*"
- "*In a usability test, I am used to speak out loud about things. Therefore, it felt a bit strange to do the tasks silently. On the other hand, maybe the settings tried to provide a natural use experience, and I believe that for most users, the thinking aloud would have been the strange thing.*"

Moderator present
+ "*The presence of the test moderator did not disturb at all, on the contrary, it would have felt as if something relevant was missing from the test if I were on my own doing the tasks*"

Moderator absent
+ "*Buying tickets went smoothly as I could work independently in my own thoughts as in normal use situation*"
+ "*The absence of a test moderator might have affected in a way that I could make mistakes and wonder about the system in peace*"
- "*It felt cold and clinical to be left alone in front of a video camera*"

Figure 28: Test users' freeform comments about the use of thinking aloud and performing in silence, as well as comments about the presence and absence of the test moderator. Positive comments are marked with a "+" and negative comments with a "-".

Freeform comments were optional in the post-test questionnaire, so some of the participants did not give any comments, or just briefly summarised: "*OK*". Therefore, the total number of comments about specific issues is low, and statistical analysis is not grounded.

The only negative comment about working in silence is from a user who has participated several usability tests before, and therefore is used to thinking aloud in such situations. Still, also this user comments that thinking aloud might feel strange for most of the users.

The only positive comment about thinking aloud, on the other hand, is from a user that considers thinking aloud as a refreshing experience, and an opportunity to learn about one's own behaviour. This user and three other users also report that the thinking aloud changed their normal behaviour by slowing them down, and giving them more time to think of various alternatives, and even by making them more patient. One quite neutral comment points out that the thinking aloud in this test was more like reporting actions than explaining thoughts.

On the part of the moderator presence, one user comments that there was more room to make mistakes, and to wonder about the system, when the moderator was in another room. However, another user commented that it felt cold

and clinical to be left alone with the recording equipment in the test room. The counts of positive and negative user comments in the post-test questionnaire about the test settings are summarised in Table 42.

Table 42: Frequency counts of positive and negative user comments about the test setting.

Group / Issue	Working silently		Thinking aloud		Test setting: natural and pleasant	
	Positive	Negative	Positive	Negative	Positive	Negative
00	2	0	-	-	2	0
0M	1	1	-	-	3	0
T0	-	-	1	2	3	0
TM	-	-	0	1	3	0
Total	3	1	1	3	11	0

5.5 Conclusions of experiment

Based on the results of the experiment, most of the hypothesis do not get statistical support. Table 43 shows a summary of the hypotheses and the results whether the experiment supports the hypotheses or not. After that, the results are reflected to the previous studies.

Table 43: The hypotheses for the experiment and the results that the analyses provide.

Hypothesis	Claim	Result
H1	Time on tasks is longer in test conditions with thinking aloud (T0 and TM) compared to the silent conditions (00 and 0M)	Supported
H2	Time on tasks is shorter in test conditions with the moderator present (0M and TM) compared to working alone (00 and T0)	Not supported
H3	Users face less usability problems in using the system when thinking aloud is used (T0 and TM) compared to silent conditions (00 and 0M)	Not supported
H4	Users face less usability problems in using the system when the moderator is present (0M and TM) compared to working alone (00 and T0)	Not supported
H5	Novice users performing with the moderator present (0M and TM) rate the system preferences higher than those working alone (00 and T0)	Supported
H6	The use of the thinking aloud method does not affect the users' perceived usability of the system (T0&TM *vs.* 00&0M)	Supported
H7	The evaluators' certainty of the causes of a usability problem are lower in conditions where users perform silently (00 and 0M) than in conditions with thinking aloud (T0 and TM)	Not supported
H8	Thinking aloud is rated as more natural in condition where the test moderator is in the same room (TM) compared to working alone (T0)	Not supported
H9	Performing silently is rated more natural in condition where the test users are alone (00) compared to having the moderator present (0M)	Not supported

5.5.1 Task times and number of problems

The effect of thinking aloud slowing down the performance is opposite to the results of R.B. Wright and Converse (1992), but similar to many other studies finding prolonging effects (*e.g.* Gagné & Smith 1962; Ericsson & Simon 1980; Rhenius & Deffner 1990; Hertzum *et al.* 2009; Fox *et al.* 2011). However, Eger *et al.* (2007); McDonald and Petrie (2013); and Hertzum and Holmegaard (2013) find no significant differences in task times between classic or explicit thinking aloud when compared to silent performance, although the task times are slightly longer in thinking aloud conditions.

In the study by R.B. Wright and Converse (1992), the test tasks require quite complex problem solving, whereas in this experiment, the tasks seemed to be quite simple for the users. Although the level of thinking aloud was not systematically analysed, the thinking aloud mostly stayed on levels 1 and 2, and only rarely included explanations for the actions. Indeed, one test user comments in the post-test questionnaire that although he is used to thinking aloud, in this test, he did not report his thoughts but merely his actions and observations. Thereby, the results of the studies by Ericsson and Simon (1980) saying that level 1 and 2 thinking aloud does not change the users' performance but may slow it down seem more appropriate for this experiment than the results of the tests with problem solving and strategy searching (*e.g.* R.B. Wright & Converse 1992; Hertzum & Holmegaard 2013).

The results by Davis *et al.* (1968) show insignificant or improving effect from the presence of the moderator, but also their study includes tasks requiring abstract thinking. However, the tasks in this experiment were quite simple, and there were no repetitive tasks, so there was not much room for learning or improving the performance with more concentration. Still, the results partly support the results of Davies *et al.* (1968), as the moderator presence did not have a significant effect on the task time or the number of errors.

The effect of thinking aloud in the number of errors is not significant in this experiment, either. However, the results of R.B. Wright and Converse (1992) suggest that the performance could be improved, and the study by McDonald *et al.* (2013a) support this view if explicit thinking aloud is used with difficult tasks. Even so, the results of Rhenius and Deffner (1990), and Hertzum *et al.* (2009) suggest that the use of the thinking aloud method does not affect the accuracy. Thereby, the results of this experiment give support to the latter findings.

5.5.2 Test users' system preferences

The effects of the moderator presence in the test users' feedback are one of the most important findings in this experiment. Held and Biers (1992) do not find significant effect of the moderator presence in the users' preferences, but identify a positive tendency with the novice users. Somewhat similarly, Sonderegger and Sauer (2009) find no effect on perceived usability. In this experiment, however, the positive effect is statistically significant ($p<0,05$) maybe due to the test users being novice users with the new service. The thinking aloud, on the other hand, has no significant effect, so it can be used without need to consider its effects on the validity of the test users' assessments.

There are several possible explanations for the positive effect of moderator presence in the users' preferences. As Barnum (2011, p. 208) reminds, test users are eager to please, and the studies by Reeves and Nass (1996) show that people are polite even to the computers. The moderator was present in the room in all the test conditions preparing the room for the next sessions when the users answered the post test questionnaires, but only the users having the moderator in the same room while doing the test tasks gave more positive feedback on the evaluated system. Thereby, the effect is probably based on the

relationship built up during the task performance, because the introductions, interviews, and conditions in post-test questionnaires were similar to all the test groups.

In the introduction, the users were told that our research group participates the project in which the system is developed, but it was also told that the system was implemented in a cooperative company, and our research groups was only studying its usability. Still, having a moderator next to the test user brings along a risk that the test users connect the moderator to the development team and are therefore polite when assessing the usability – or they are being polite in any case.

Having the moderator next to the users may also build a rapport of sharing the problems with someone, and thereby, making the problems less severe. To get some certainty on the reasons for the positive tendency, further studies are required, so that the relationship between the moderator and the development team is clearly ruled out. In the meanwhile, the possible positive tendency should be taken into account when interpreting the results of a usability test.

As another point of view, the more positive ratings from the users having a moderator in the same room could result from getting particularly negative ratings from the ones performing alone, as Dumas and Loring (2008, pp. 126-128) suggest. The overall ratings of the naturalness of the test setting, however, are very high for all the test conditions, and have no significant differences. Also the ratings for the statement concerning the absence of the moderator strongly suggest that the users do not regard the situation as unpleasant or disturbing, as only 1 user out of the 20 participants working alone slightly agree with the statement: "*I would have preferred to have the test moderator in the same room with me*". All the other users give either neutral rating (3 users), disagree slightly (9 users) or totally disagree with the statement (7 users). For these reasons, it is more likely that the users having the moderator next to them rate the system more positively than that the users performing alone rate the system especially negatively.

5.5.3 Evaluators' certainty of their assessments

Another interesting finding from the experiment is that the use of the thinking aloud does not make the evaluators significantly more certain about the causes of the usability problems. This point of view has not been studied in the previous experiments, so the result still needs support from further studies with a variety of systems including also more complex systems than the eTicket application in this experiment.

Although the identification of the cause of a problem is one of the first steps in improving the system, it does not alone help in finding redesign proposals. In discovering ideas for improvements, discussions with the users are most valuable, and they take place quite naturally in the tests with the moderator sitting next to the user, and another evaluator taking notes in the observation room. To enhance active discussions with the users, it is also worth considering to have the test users thinking aloud, since it is "*surprisingly hard to start*

talking after working in silence", as one of the test users comments in the post-test questionnaire after performing silently.

5.5.4 Test users' feedback on test settings

Working alone silently receive very high ratings in the test users' assessments of the test settings, thereby supporting the findings of Eger *et al.* (2007) as their test participants find the thinking aloud significantly more unpleasant than performing in silence in the retrospective comparison setting. If the ratings in this experiment are treated as linear values for demonstration, the mean for the statement: "*Doing the tasks silently felt natural*", is as high as 4,3. Even with the presence of the moderator, the silent performance gets higher ratings (3,7) than either of the groups thinking aloud. Those working alone rate the naturalness of thinking aloud as 3,4 in average, whereas the ones with the moderator present give the lowest ratings (3,2).

Nevertheless, the test users assessed the naturalness of the test situation very positively in all the test conditions without significant effect of the thinking aloud or the moderator presence. Therefore, the use of the thinking aloud method needs to be considered according to the goals of the test. If silent performance is required due to efficiency measures or similar, it is also recommended to consider letting the users work alone to make the silence more natural, although the effect of the moderator presence is only indicative in this experiment.

5.6 Reliability and validity of experiment results

Experiments should be replicable and produce similar results to be reliable (*e.g.* Lazar *et al.* 2010, p. 57). To facilitate repetition tests, the process of selecting the users; the methods used; the instructions given to the users; the test procedure and the test tasks; as well as the procedures in analysing the data are presented in quite detailed level in this thesis. The human subjects inevitably bring some random error to the results, as Lazar *et al.* (2010, p. 57) remind, but the group of test users was selected with care to enable as homogeneous test groups as possible and, thereby, to avoid systematic errors due to the test participants. When allocating the users into different test groups, their genders and current mobile phones were balanced between the groups, but otherwise the allocation was random from the prescreened group of students.

The participants in this experiment were representative users for evaluating the usability of the eTicket system as they all were young, used smart phones and were interested in new mobile applications indicating they were potential lead users for this type of mobile services. Although these test users fulfilled the aspirations of our cooperative company representatives, they were not the optimal user group for this experiment on the part of assessing the naturalness of various test conditions, because they were students in our usability course. Thereby, they were likely to have certain expectations of the test settings including a moderator sitting next to the user, and the test user thinking aloud. As these typical settings were not used in all the test conditions, it was re-

quired to briefly motivate the used test condition. These expectations also prevented me from totally leaving the users alone in the test conditions 00 and T0, because I had lectured to these same participants how to honour the test users, and be present to listen and support them while making the tasks. Therefore, I helped all the users through the first task, which may have smoothened the differences between working alone and with the moderator in the same room. Nevertheless, this is a similar procedure to the study by Davis *et al.* (1968) in which the evaluator reminded that he will continue monitoring the session from the backroom. In the study by Sonderegger and Sauer (2009), the setting was also the same with the exception that the monitoring was done without the participants' awareness after the evaluator left the room.

The experimental procedure was the same for all the test users except for the test conditions, so possible systematic errors in the procedure should be the same in all the test groups, and thereby have no effect in comparisons between the test groups. The moderator was the same in all the tests, the instructions given were similar to everyone including the same background information for all the users, and the test tasks were exactly the same. All the tests were conducted in our usability laboratory in which the environment could be controlled both for the physical and social part. The measurement instruments were very simple including a timer and two web questionnaires with statements to be rated.

The sample size in this experiment was 38 test users making the group size in each test condition 9 or 10, which is a little bigger than typically in similar studies. For example, Gagné and Smith (1962) had 28 participants in a study with 4 test conditions; Davis *et al.* (1968) had 48 participants with 8 test conditions; and R.B. Wright and Converse (1992) had 24 participants with 4 test conditions. All of these tests were between-subjects studies similar to this experiment. On the other hand, the split-plot study by Sonderegger and Sauer (2009) had 60 participants with 6 test conditions exposing as many as 20 users to each laboratory condition.

On the part of the video analyses, evaluator triangulation was used as two independent evaluators who had not participated in the design of the experiment made their analyses using a common format. Despite the common instructions and common format, the two evaluators used the reporting format and even the scale of certainty assessments quite differently. Therefore, their results were analysed separately both for the number of problems found and the certainty of the causes of these problems.

The experiment was planned and conducted in a similar way as a typical usability test with the exception of adding an extra questionnaire to the end of the test to study the users' experiences of the test setting. The setting TM with thinking aloud and moderator present represents the most common setting in formative usability evaluation in Finland, and the setting 0M with silent performance next to the moderator is commonly used in summative testing. Although rare, the test users are sometimes left alone with the test tasks, or the test is moderated remotely. Thereby, the experiment represented quite well a

usability test and could be used as a sample to study the effects of relaxed concurrent thinking aloud and the moderator presence.

However, as the studies of the contextual factors of usability testing have shown, the results may vary considerably according to the difficulty of the test tasks and the complexity of the evaluated system. The study by Sonderegger and Sauer (2009) also indicates that the moderator's ability to set up a good rapport with the test users may affect their performance. Although the results of that specific study give just the opposite results, as the users working alone rate their emotions as more positive than the others, it is possible that the moderator may affect some subjective results in addition to the users' performance through the improved rapport. Therefore, the results of this experiment can not be generalised to all situations, but still give a good indication that the presence of a test moderator may cause the users to give more positive feedback than without the moderator. In addition, the use of the thinking aloud may slow down the users' performance with rather simple tasks and simple systems.

5.7 Limitations of experiment

The experiment has several limitations in addition to covering only one quite simple system with rather simple test tasks. As mentioned before, I could not conduct the experiment with such strict differences as otherwise might have been possible, because I was the teacher who had told the participating students in previous lectures how to moderate usability tests and interact with the test users. Therefore, even after the initial instructions, I did not consider suitable to leave the participants alone without further explanations. Still, I wanted the same moderator to conduct all the sessions to keep the conditions as similar as possible, and considered myself as the most experienced moderator from the limited resources. Thereby, if I were to do a similar experiment again, I would try to recruit participants who have not acted as test users in a usability test before, and therefore, have no specific expectations concerning the tests. This would enable bigger differences between the test conditions as the users could be left alone for the whole session without further explanations.

Moreover, to hide the real goal of the experiment and to engage the test participants in evaluating the usability of the eTicket system instead of the test setting, I had to use a between-subject design. This design helped in keeping the test sessions brief enough for the students to stay focused, and also helped in avoiding any carry over between the test settings. However, the test groups became quite small and sensitive to personal differences, but I tried to minimise these personal differences by selecting the test participants carefully with the help of the pre-test questionnaire.

On the part of the methods used in the experiment, only one type of thinking aloud was used, since this relaxed concurrent thinking aloud focusing on the reasons of users' behaviour is the common way of applying thinking aloud in the Finnish usability practice. The post-test questionnaires were not stand-

ardised questionnaires, so following to the recommendations by Hornbæk and Law (2007), I would use in future tests the NASA-TLX to study the workload users perceive while doing the tasks, and the SUS questionnaire to give more standardised results for the comparisons of subjective ratings between the test groups.

On the part of analysing the data, only two evaluators analysed the video recordings, but they both were experienced evaluators and had not participated in the design of the experiment, making them outside coders. Due to our very limited resources, more evaluators were not available, and already these assessments showed that although the evaluators had different style in their evaluations, the differences between the test conditions were to the same directions for both the evaluators. Especially, on the part of the evaluators' certainty of the causes of the found usability problems, there was notable differences between the two evaluators. Even so, neither of them showed any significant difference in the ratings between the different test settings, including the thinking aloud that was expected to increase this certainty.

6 Discussion

This chapter discusses on the contributions of this thesis, its limitations, and future work.

6.1 Contributions of thesis

This thesis has several goals. As an academic thesis it aims to bring something new to the field of usability research. In this respect, the thesis has reached its goals, as it presents several testing methods developed in our research group along with some experiences giving more insight into these methods. Although the methods are to some extent presented in my Licentiate's thesis in 2000, there are new experiences to be shared, and also a new modification of the visual walkthrough. The variety of the methods and their modifications along with some examples of their use hopefully convince the readers that a set of suitable evaluation methods can always be found whether it includes usability testing or not, so that even the needs of strict summative testing in procurement procedures, or the more formative and iterative needs of agile software development can be met.

This thesis has three approaches, each of them giving its own contribution to the research area in usability testing. First, the literature based part gathers a wide set of studies in this area including studies related to the established process of usability testing, and studies focusing on certain contextual factors and their effects, giving special attention to the effects of thinking aloud and moderator presence. Secondly, the empirical part presents five modifications of the common usability testing. Four of them are developed in our research group with my considerable contribution, as the publications of these modifications indicate. These methods have been taken into use also outside our research group. For example, researchers in the University of Tampere have used the informal walkthrough method in their studies. The papers about the methods have also contributed some academic discussions, as my conference paper on informal and contextual walkthroughs (Riihiaho 2009) has been cited three times in international academic forums: Interacting with Computers, Lecture Notes in Computer Science, and Behaviour & Information Technology. Furthermore, my licentiate's thesis (Riihiaho 2000) has been cited 48 times[27] nationally and internationally. Thereby, the empirical part of

[27] Result of a search in Google Scholar on October 13th 2014

my studies has already given a contribution in this field, and hopefully the new updated experiences will give even more contribution.

Finally, the experimental part addresses two issues that have been taken for granted in usability testing in our university and also elsewhere based on the textbooks in the field of usability testing. Similar studies of the effects of thinking aloud in usability testing have been conducted at the same time as I have made my own, and more have been done after it. However, the results of these studies are still somewhat contradictory, so even more studies are needed, and each contribution is valuable. The other issue studied in my experiment was the effect of the moderator presence, and this issue, has been addressed only in few experimental studies in the field of usability testing before. Thereby, this contribution is important and substantial to this field.

Although many of the hypotheses in the experiment are left without confirmation, the experiment and its results nicely support the findings of similar previous studies, and complements the studies on some parts. For example, the positive effect that the moderator has on the test users' subjective ratings had not been clearly reported in the area of usability testing before. In the studies by Held and Biers (1992), the effect was not significant with novice users, and with expert users, the effect was negative. Also the result that the thinking aloud does not have significant effect on the users' ratings is important for the usability testing practice, as the use of the thinking aloud has become almost the standard policy in the field.

On the personal level, I aimed to get a holistic picture of the present state of the research on usability testing, and thereby, to find out, if there still is active research on it or not. I was pleased to find out that there indeed is active research, and new methods are still being developed. In quite recent proceedings of the Human Factors and Ergonomics Society Annual Meeting, Youmans *et al.* (2013) summarise some of the results of previous thinking aloud studies as follows: "*Given the widespread use of concurrent verbal protocols in applied domains, it may be surprising for analysts and researchers alike to learn that nobody fully understands how concurrent verbal protocols affect task performance.*" However, also they conclude that concurrent verbal protocols are "*an excellent method for finding potential problems with user interactions, developing training procedures, and for obtaining information about how users go about solving problems, but not for making reliable estimates about user performance*" (Youmans *et al.* 2013).

Discussions with usability practitioners confirmed my assumptions that concurrent thinking aloud still is an essential part of usability testing. It helps in identifying the usability problems and their causes, and furthermore, helps the customers in understanding the problems when they observe the sessions or watch the video clips. Thereby, the impact that thinking aloud can have on the development team usually overcomes the potential effects on the users' performance. Still, the effects should be better acknowledged.

The same applies to the effect of having a moderator present, the effects of which seem to be even less studied in the field of usability testing. Again, the possibility to discuss on task specific issues when they are still fresh in mind

usually overcomes the potential effect in subjective ratings. Acting as a moderator next to the test user also offers an opportunity to intensify the empathy towards the users, and to deepen the understanding and expertise in the application domain. Still, the effects of the moderator presence should be better known and taken into account in interpreting the test results.

6.2 Coverage and limitations of the study

My discussions with usability practitioners also brought forth the issue of evaluators' expertise in addition to the complexity of the evaluated system. The complexity of the system is considered to some extent in the experiment, as the evaluated system is quite simple, and the results cannot be generalised to more complicated systems without further studies. Indeed, the recent studies by McDonald *et al.* (2013a) indicate that explicit thinking aloud significantly improves test users' performance compared to classic thinking aloud when doing complicated tasks, whereas simple tasks do not have the same effect. However, their study does not include silent conditions.

The effects of the evaluators' expertise was not studied in my experiment, but both the evaluators analysing the test recordings had several years of experience in usability studies, although not specifically in usability testing. Thereby, the users' actions alone gave enough information about the problems for these experienced analysers with such a simple system. Still, the users' knowledge and experience is emphasised with more complicated and domain specific systems, and concurrent verbal reports provide very practical tools in getting into this information. However, if the usability experts have also the required knowledge in the domain area, even the expert reviews without user involvement can produce comparable results to usability tests, as the studies by Molich and Dumas (2008), and Frøkjær and Hornbæk (2008) indicate.

The limitations of the experiment are already discussed in Chapter 5 along with the reliability and validity of the experiment results, so the remainder of this subchapter focuses on the limitations of the literature review and the empirical part. The literature review focuses on the top HCI forums to limit the amount of material, but also to follow the recommendations by Levy and Ellis (2006) to use literature from leading, peer-reviewed journals as the major base of literature review. During the review, I studied about 530 articles, and selected about half of them, so that the total number of references in this thesis ended up to be 290. Thereby, the coverage should be wide enough, especially on the part of the top HCI forums, and also the depth should be deep enough on the part of usability testing and thinking aloud method.

Also the time frame of the references is quite extensive, the oldest being from the 1920's and 1930's, and the newest from 2014. It was fascinating to read the articles about the old experiments, and learn from their precise and systematic reporting. Also their results are interesting, although the connection to usability testing are not always obvious. For example, studies by Duncker (1945) and Maier (1931) show that participants do not report or even register in their working memory the very last hint or pointer that help them in

solving a problem. In usability testing this may mean that some key elements that actually help users in finding the right functions are not necessarily reported, but can be observed only through their actions. Studies by Orne (1962), on their part, show that people are ready and willing to do things in an experiment that they would not even consider otherwise. Therefore, the scarce dropouts in usability tests do not necessarily prove that the systems are easy and pleasant to use, but instead indicate that many people are very willing to help in improving the evaluated systems by trying them out and commenting them in the best way they can.

The empirical part of this thesis is limited to my experiences in our research projects and course assignments. Still, these experiences cover 143 different usability evaluations for various systems. Although most of the evaluations have been made as course assignments by our master or doctoral level students in our guidance, all the topics have come from real cases and have involved representatives from companies developing these systems. Thereby, the pool of these case studies is rich and multifaceted.

6.3 Future work

As so often in research, also this experiment and its results have perhaps raised more questions than they have answered, so there are many interesting issues still to be further analysed from the data gathered in this experiment, or to be studied in new experiments with more complex systems or with a variety of different thinking aloud methods. For example, there are several studies comparing the time on tasks and number of errors, but not many studies comparing the users' strategies in solving the tasks, analysing if they vary between different test settings. To study this, one simple task could be analysed from the present experiment data from each test user, and modelled as a sequence of actions that could be compared within and between the various test conditions as string comparisons, similar to the comparisons made in the study by Guan *et al.* (2006). Interesting questions include, if the presence of a moderator or use of thinking aloud affect how much users explore the system. The studies by Hertzum *et al.* (2009) and McDonald and Petrie (2013) address these questions on the part of thinking aloud, and show that the amount of navigation and scrolling increases in relaxed thinking aloud, or if users are explicitly instructed to report things they like, dislike or find confusing.

The contents of the test users' utterances could also be analysed both on the part of the level of thinking aloud, as Ericsson and Simon (1980) categorise the levels, and on the part of the level of interpretation, as Sengers and Gaver (2006) name the levels of users' assessments. In the first part, a rough division could be made between verbalising actions and observations, and explaining expectations, reasons and experiences. In the latter part, the utterances related to the operating of the system, its utility and value to the user would all be divided into their own groups, and the number of instances in these groups could be compared between different test settings.

7 Conclusions

This chapter outlines the themes presented in this thesis, and summarises the answers to the research questions. Finally, the research problem is addressed.

7.1 Usability testing in academic discussions

The first research question in this thesis is:

1. Which of the top academic HCI forums are active in discussing usability testing methodology?

The academic discussions on usability evaluation and usability testing, were very active in the 1990's as new methods were developed and assessed. The number of papers including usability evaluation increased for example in CHI conference even after that, but as the studies by Barkhuus and Rode (2007) show, most of the papers merely apply the evaluation methods instead of studying and developing the methods. After this declining, research on usability evaluation methods has again enforced, and the CHI 2013 conference included even three paper sessions on evaluation methods[28]. Table 44 shows the number of references used in this thesis from various HCI forums.

Table 44: Number of references in this thesis from various HCI forums and their time frame.

Forum	Number of references	References 2010-	Time frame
ACM Transactions on Computer-Human Interaction	6	1	1996-2012
Applied Ergonomics	9	7	1998-2013
Behaviour & Information Technology	22	8	1991-2014
Ergonomics	5	1	1990-2012
Human-Computer Interaction	6	1	1998-2010
Human Factors: The Journal of the Human Factors and Ergonomics Society	2	0	1992-1994
IEEE Transactions on Professional Communication	7	4	1989-2012
Interacting with Computers	8	1	1994-2011
International Journal of Human-Computer Studies (Previously International Journal of Man-Machine Studies)	14	1	1986-2012
International Journal of Human-Computer Interaction	18	6	1995-2013
CHI conference	34	7	1982-2014
Human Factors and Ergonomics Society Annual Meeting	12	2	1990-2013
NordiCHI conference	11	2	2004-2010
Total	154	41	1982-2014

The time frame for the references, and especially the references after 2010 show that academic discussions on usability evaluation and testing are still active in the top HCI forums. The number of references also demonstrates

[28] http://chi2013.acm.org/program/by-day/

nicely how central role the conferences, especially the CHI conferences have in the HCI field. In addition, the table shows the close relation between usability evaluation methods and studies of human factors as the large number of references in Behaviour & Information Technology, and Human Factors and Ergonomics Society Annual Meeting.

If we take a closer look on the references used in this thesis on the part of the experiments on thinking aloud and moderator presence in usability test settings, and examine the forums in which these studies have been published since the year 2000, the CHI conference proves to be an important forum also for these discussions along with Behaviour & information Technology. Table 45 shows the title of these articles as well as their publication forum.

Table 45: Experiments addressing the effects of thinking aloud (TA) or moderator presence (MP) in usability testing published since the year 2000.

Authors	Title	Publication forum	TA	MP
Van den Haak, de Jong & Schellens 2003	Retrospective *vs.* concurrent think-aloud protocols: Testing the usability of an online library catalogue	Behaviour & Information Technology	X	
Schulte-Mecklenbeck & Huber 2003	Information search in the laboratory and on the Web: With or without an experimenter	Behavior Research Methods, Instruments, & Computers		X
Krahmer & Ummelen 2004	Thinking about thinking aloud: A comparison of two verbal protocols for usability testing	IEEE Transactions on Professional Communication	X	
Guan, Lee, Cuddihy & Ramey 2006	The validity of the stimulated retrospective think-aloud method as measured by eye tracking	SIGCHI conference on Human Factors in computing systems (CHI)	X	
Eger, Ball, Stevens & Dodd 2007	Cueing retrospective verbal reports in usability testing through eye-movement replay	British HCI Group Annual Conference on People and Computers (BCS-HCI)	X	X
Hertzum, Hansen & Andersen 2009	Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload?	Behaviour & Information Technology	X	
Sonderegger & Sauer 2009	The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures	Ergonomics		X
Olmsted-Hawala, Murphy, Hawala & Ashenfelter 2010	Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability	SIGCHI conference on Human Factors in computing systems (CHI)	X	
Olmsted-Hawala, Murphy, Hawala & Ashenfelter 2010	Think-aloud protocols: Analyzing three different think-aloud protocols with counts of verbalized frustrations in a usability study of an information-rich web site	IEEE International Professional Communication Conference (IEEE IPCC)	X	
Cooke 2010	Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach	IEEE Transactions on Professional Communication	X	
Zhao & McDonald 2010	Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods	Nordic Conference on Human-Computer Interaction (NordiCHI)	X	
Hertzum & Holmegaard 2013	Thinking aloud in the presence of interruptions and time constraints	International Journal of Human-Computer Interaction	X	
McDonald & Petrie 2013	The effect of global instructions on think-aloud testing	SIGCHI conference on Human Factors in computing systems (CHI)	X	
Zhao, McDonald & Edwards 2014	The impact of two different think-aloud instructions in a usability test: A case of just following orders?	Behaviour & Information Technology	X	

7.2 Overview of usability test challenges

The second research question in this thesis is:

2. What problems and challenges have been reported on present usability testing methods in these forums?

The challenges and problems reported on usability testing could be categorised to sampling problems; taking the context of use, utility and long term use into account; and misuse of the term *"usability testing"*. These topics are briefly discussed in the following subchapters, along with some suggestions for solutions presented in this thesis.

7.2.1 Problems related to sampling

There are several types of sampling problems in usability testing. The participants must be selected with care so that they are representative users and fit to the focus chosen for the evaluation. The selection criteria should be clearly reported as well as the backgrounds of the test users so that the test users' fit to the test can be assessed. This data is also needed when considering potential effects of users' background on the results of the evaluation.

The number of required test users has been a subject of active debate for a long time in usability research. Some studies indicate that 5 users is enough (*e.g.* Virzi 1992; J. Nielsen 1994a), but J.R. Lewis (2001b), for example, gives a wider estimate as a formula $1-(1-p)^n$, where p is the problem-discovery rate, and n is the sample size of test users. Especially with complex systems, this formula gives considerably larger numbers of required test users than just five. Lindgaard and Chattratichart (2007), on the other hand, suggest making more test tasks with less users to reveal as many usability problems as possible, and the analyses by Skov and Stage (2012) complement this finding by emphasising the quality and relevance of the test tasks.

On the part of the type of the test tasks, Dicks (2002) and Hornbæk (2006) call for tasks that require cognitive problem solving instead of mere motoric actions. This way, the users would have better opportunity to assess the utility of the system in addition to its ease of use. J.R. Lewis (2001a), on his part, reminds that even the convention of giving only doable tasks in user testing may bias the results, as the users are more convinced in finding the required functionality in a test than in a real use situation.

Informal walkthrough and contextual walkthrough take these issues of test tasks into account by letting the test users explore the system on their own and try the functionalities they like or are interested in. Also Cordes (2001) recommends using test tasks that the test users bring along to the test, but he recommends to collect them in advance to give the test moderators more control of the sessions, and to enable statistical analyses. Respectively, the checklist used in informal walkthrough gives means to have all the test users try out a certain set of test tasks, and thereby, enables quantitative analyses.

Also critical incidents reporting (Hartson & Castillo 1998), and Experience Clip (Isomursu *et al.* 2004) give room for users in deciding what to do with the system, and even let the users decide what to report to the evaluators. As the

critical incidents reporting does not require the evaluators to be near the users, it enables also longitudinal studies.

After selecting the test participants and the test tasks, there are sampling challenges also in analysing and interpreting the data. As the studies by Hertzum and Jacobsen (2001); Molich *et al.* (2004); Molich and Dumas (2008); and Hornbæk and Frøkjær (2008b) point out, different evaluators interpret the data differently, and end up to different conclusions. Therefore, Holleran (1991) recommends using quantitative measures with statistical analyses to complement and support subjective interpretations of qualitative data, as well as using several methods in the studies for getting a variety of data. Still, as the experiments before and the experiment in this thesis show, even specific format and instructions for reporting the problems does not make the evaluators' reports alike. Therefore, the moderator effect is something that needs to be taken into account in interpreting the results, and several evaluators are recommended to do the analyses to enable discussions on unclear situations.

7.2.2 Assessment of evolving use and utility

Problems in the assessment of long time use and evolving use, as well as the need for evaluating the usefulness of systems have been brought out in several articles (*e.g.* Cockton 2006; Sengers & Gaver 2006; Greenberg & Buxton 2008). For example, Greenberg and Buxton (2008), point out that the controlled studies generally used in the HCI field give very little room for the users to assess the utility of the system, or to figure out new ways of utilising the system in everyday use. Also Sengers and Gaver (2006) want the evaluation to take into account the new interpretations that users do of the system and its use, and to give users more room to explore the system and to assess it in various levels from operating the system to considering its value in their own life.

Contextual walkthrough and informal walkthrough try to tackle these issues by giving the test users room to explore and, thereby, assess the system in various levels. These methods can be used in different phases of system development, and also after the system has been taken into use to study the evolving use in real context or in laboratory settings. The critical incidents reporting (Hartson & Castillo 1998) also enables longitudinal studies in real use context, but it is based on self-reporting certain acute events with a simple form, so it does not give much room for assessing other issues besides usability.

On the part of the real use context, contextual walkthrough is built on the idea of making the evaluations in real use contexts with real tasks. Similarly, Contextual Assessment of Systems Usability (CASU) (Savioja *et al.* 2008), and ADA (Åborg *et al.* 2003) are used in real use context to study the usability of occupational systems, although the CASU method is used with simulated process scenarios for safety reasons. For leisure time systems, informal walkthroughs can be conducted in real use contexts with tasks selected by the users, as well as the Experience Clip by Isomursu *et al.* (2004).

On the part of the utility assessment, the new utility inspection method (UIM) by Johannessen and Hornbæk (2014) focuses specifically on utility.

Similarly, the colour coding used together with visual walkthrough (Juurmaa *et al.* 2013) concentrates in assessing the relevance of various user interface elements in the users' work or activities. Furthermore, most of the methods leaving the test users room to explore the system enable some evaluation of utility, as well as system acceptability and value. For example, informal and contextual walkthroughs and the Experience Clip by Isomursu *et al.* (2004) allow this type of exploration.

To give users more time to adapt and find their own way of using the systems, Cockton (2006) recommends using self-reporting methods to study the use in real context and over time. For example, the longitudinal studies by Sonderegger *et al.* (2012), show that the effects of aesthetics on perceived usability and emotions wane already in a week, so longitudinal studies could indeed yield more realistic results compared to the one-off usability tests.

7.2.3 Misuses of term "*usability testing*"

Dicks (2002) brings forth the misuse of the term "*usability testing*" in usability practices. In addition to its real meaning, *i.e.*, planned tests with representative users, the term is sometimes used for all the user testing methods including interviews and questionnaires, and even for usability inspection methods without any user involvement. Therefore, accurate and systematic use of the top terms, *i.e.*, usability evaluation, user testing, usability inspections and usability testing, is essential in teaching the topic, in discussions with the company representatives, and especially in reporting and reviewing in academic forums.

Greenberg and Buxton (2008) bring out also another misuse of usability testing, namely the use by rule. Although right terms are used, the methods are sometimes applied without thought and clear goals just to be able to report that the design has been tested with users, and thereby, to convince the article reviewers of the validity of the study. They also point out that usability testing, even when correctly applied, is not always the right choice for gathering users' comments and feedback. For example, the design may be on a too rough level, and should be introduced to the users in some more open way than traditional usability testing. (Greenberg & Buxton 2008) For instance, pluralistic usability walkthrough, especially in its modified version (Riihiaho 2002), gives this kind of opportunity for developers to present their designs, and for users to comment the designs.

7.2.4 Assumptions in method comparisons

The article by Hornbæk (2010) summarises many of the problems that the comparisons of various usability evaluation methods often face, but are rarely brought clearly forth in the studies or discussed. For example, the number of problems found is quite a general measurement for setting the methods in order by effectiveness and efficiency, although the problems may be reported with very different specificity, and even without knowing about their relevancy in real use. A *"single best*" usability evaluation method is assumed to exist

(Hornbæk 2010), and usually, it is usability testing that is considered as this baseline method[29]. For example, Cuomo and Bowen (1994) study which of the problems found with other methods can be confirmed by usability tests, and Desurvire (1994) uses the laboratory testing as a benchmark for the other methods. Thereby, usability testing is often considered as a representative of real use, dictating which problems are real and which ones are false alarms.

Furthermore, Hornbæk (2010) and Woolrych *et al.* (2011) point out that the results of the evaluations are very much dependent on the evaluators' expertise and the way the methods are applied, which makes it even more challenging to compare different methods. As Woolrych *et al.* (2011) recommend, the evaluation methods should be considered as "*mixes of ingredients*" instead of recipes leading to similar results whenever applied. Different methods are good at different situations, so they should be selected, mixed and modified according to the goals and context of the evaluation, instead of comparing them with simple metrics in a one sample situation.

7.3 Methods for usability testing

The third research question in this thesis is:

3. What methods and modifications of usability testing are reported?

This thesis presents quite many methods for usability testing in addition to the traditional usability testing. Table 46 summarises these methods briefly by presenting the level in which the method permits users to assess the system (operation, utility or value); the time interval that is studied; the environment the method is intended for; the source of the test tasks; the number of test users in one test session; and the level of prototype required (paper prototype, functional prototype or working system). In addition to this key information, the reference in this thesis as well as the original reference for the methods are presented in the table. Some methods, such as the usability roundtable (M.B. Butler 1996) and ADA (Åborg *et al.* 2003), are left out of the summary, because the first one focuses on gathering user requirements instead of usability testing, and the latter is intended only for experienced occupational health experts to be applied when visiting the employees' workplaces.

The summary clearly shows that most of the methods require something that the users can actually try out, *i.e.*, a functional prototype or even a working system. Only the pluralistic usability walkthrough and visual walkthrough methods are intended to be used with paper prototypes. As Greenberg and Buxton (2008) point out, with paper prototypes and early sketches, other methods, such as focus groups, may be more applicable than usability testing. Similarly to focus groups, also pluralistic usability walkthrough gives room for group discussions, so these are good alternatives to be considered in the early phases. Focus groups may be more appropriate for the first ideas focusing on the set of functionalities, whereas pluralistic usability walkthroughs are recommended for evaluating the way these functionalities will be operated.

[29] This issue was discussed also in my licentiate's thesis (Riihiaho 2000, p. 58)

Furthermore, the summary shows that quite few of the methods support the assessment of other issues besides the usability of the system. In practice, the methods allowing the users to explore the system on their own, or letting the users select their own tasks, also leave some room for assessing the usefulness of the system, and even the value of the system in daily life and social contexts.

Table 46: Summary of methods for usability testing and their key information.

Method (reference)	Special focus	Level of evaluation	Time frame	Context	Tasks	Users per session	Level of prototype
Traditional usability testing (Ch. 3)		Operation	Use session	Controlled	Predefined	Single	Functional
Question asking protocol (Ch. 3.3.1, Kato 1986)	User asks questions from a tutor	Operation	Use session	Controlled	Predefined	Single	Functional
Cooperative evaluation (Ch. 3.3.2, P.C. Wright & Monk 1991)	System developer as a test moderator	Operation	Use session	Controlled	Predefined	Single	Functional
Cooperative Usability Testing (Ch. 3.3.3, Frøkjær & Hornbæk 2005, Følstad & Hornbæk 2010)	Interpretation sessions with users	Operation, utility	Use session	Controlled	Free or predefined	Single	Functional
Critical incidents reporting (Ch. 3.3.4, Hartson & Castillo 1998)	Self reporting triggered by critical incidents	Operation	Use over time	Real use context	Real tasks	Single	Working system
Experience Clip (Ch. 3.3.5, Isomursu *et al.* 2004)	One user testing and other recording	Operation, utility, value	Brief use on own	Real use context	Real tasks	Pair	Functional
Paired-user testing (Ch. 4.1)		Operation	Use session	Controlled	Predefined	Pair	Functional
Peer-tutoring (Ch. 4.1.2, Höysniemi *et al.* 2003)	Learning	Operation	Use session	Controlled	"*Teach the tutee to use the system*"	Pair or more	Functional
Pluralistic usability walkthrough (Ch. 4.2)		Operation, utility	Use session	Controlled	Predefined	Two or more	Paper prototype
Modified visual walkthrough (Ch. 4.3.1)	Information and functionality relevance	Operation, utility	Use session	Controlled or free	Free or predefined	Single	Printouts
Informal walkthrough (Ch. 4.4)		Operation, utility, value	Use session	Controlled or free	Free and some predefined tasks	Single or more	Functional
Contextual walkthrough (Ch. 4.5)	Work-related systems	Operation, utility, value	Use session	Real use context	Real tasks from the context	Single or more	Functional
Contextual Assessment of Systems Usability (Ch. 4.5.3, Savioja *et al.* 2008)	Work in control rooms	Operation, utility	Use session	Real use context	Simulated processes	Single or more	Functional

7.4 Contextual factors of usability testing

The fourth research question in this thesis is:

4. What contextual factors of usability testing have been identified, studied and reported?

In addition to the variety of usability testing methods, each of these methods includes several contextual factors that affect the outcomes of the method. The effects of these factors have been studied in various experiments, and the experiment in this thesis complements this set. The studied factors include the test environment; the presence of a test moderator or an observer; number of test users in a session and overall in the evaluation; the number and type of test tasks; the test users' level of expertise in various aspects; the level of the prototype; aesthetics of the system; level of usability; having parallel tasks to focus; users' expectations of the system; price of the evaluated system; and use of the thinking aloud method.

Table 47 summarises the studies of contextual factors discussed in this thesis, and shows their independent and dependent variables. The table is not all-inclusive as some of the studies have also other variables that are left out of the table since they are not in focus in this thesis. The studies of the effects of thinking aloud are outlined separately in Table 48. As can be seen from these tables, performance measures, such as time on tasks, are very popular variables in the studies in addition to subjective ratings and number of usability problems found. The estimated workload, usually measured with the NASA task load index, is also a very general measurement in these type of studies.

Methodological issues, such as a possible bias of the presence of a test moderator and the effects of thinking aloud method, are mentioned already in the pitfalls listed by Holleran in 1991. Still, these issues have not received much attention in the later studies. The absence of these studies backed up my original plans for the experiment to focus on the effects of moderator presence and the use of the thinking aloud method in usability testing.

Table 47: Studies of various contextual factors in usability test settings, and their independent and dependent variables.

Reference / independent variables	Environment	Moderator	Pair of users	Users' expertise	Level of proto	Aesthetics	Usability	Dual task	Expectations	Price of product	Dependent variables
Davis *et al.* 1968		X									Performance
Hackman & Biers 1992		X	X								Performance, subjective rating, quality of thinking aloud
Held & Biers 1992		X		X							Subjective rating
Barker & Biers 1994	X			X							Amount of errors made, subjective rating
Archer & Yuan 1995					X						Number of key presses, subjective rating
Virzi *et al.* 1996					X						Problems found
Catani & Biers 1998					X						Number and severity of problems found, subjective rating
Tractinsky *et al.* 2000						X	X				Perceived aesthetics and usability
Schulte-Mecklenbeck & Huber 2003	X	X									Amount of information read, number of dropouts
Kjeldskov & Stage 2004	X							X			Number and severity of problems found, performance time, workload
Kaikkonen *et al.* 2005	X										Number, type and severity of problems found, performance time, overall user performance
Kjeldskov *et al.* 2005b				X							Number of solved tasks and problems found, total task completion time, workload
Ben-Bassat *et al.* 2006						X	X				Performance, subjective rating
Duh *et al.* 2006	X										Number and severity of problems found, performance time, overall user performance
C.M. Nielsen *et al.* 2006	X										Number, type and severity of problems found, workload, overall satisfaction
Sauer & Sonderegger 2009					X	X					Performance, subjective rating
Sonderegger & Sauer 2009		X									Performance, subjective rating, physiological measures
Sauer *et al.* 2010				X	X						Number and severity of problems found, performance
Raita & Oulasvirta 2011									X		Performance, subjective rating, workload
Sauer & Sonderegger 2011a	X							X			Water and electricity use, perception of instructions and several other measures
Sauer & Sonderegger 2011b						X			X		Performance, subjective rating
Mugge & Schoormans 2012				X		X					Expected usability
Sonderegger *et al.* 2012						X	X				Performance, subjective rating, perceived aesthetics, emotion
Boothe *et al.* 2013					X						Number of problems found, subjective ratings
Sonderegger & Sauer 2013										X	Subjective rating
Riihiaho (this thesis)		X									Performance, number of problems found, user rating, evaluators' certainty on judgements, users' feedback on test settings

Table 48: Studies of the effect of thinking aloud, and their independent and dependent variables. The version of thinking aloud has varied between classic thinking aloud, more interactive, relaxed or explicit thinking aloud, retrospective thinking aloud, and working in silence.

Reference / independent variables or conditions	Classic thinking aloud	Interactive, relaxed or explicitly instructed TA	Retrospective	Silence	Dependent variables
Gagné & Smith 1962	X			X	Performance
Davis *et al.* 1968	X			X	Performance
Stinessen 1985	X			X	Performance
Russo *et al.* 1989	X			X	Performance
Rhenius & Deffner 1990	X			X	Overlap with eye-tracking data
R.B. Wright & Converse 1992	X			X	Task times, errors made
Ohnemus & Biers 1993	X		X		Performance, subjective rating, amount and quality of verbalising
Van den Haak *et al.* 2003	X		X		Number and type of problems found, performance, completion rate, experiences with thinking aloud
Krahmer & Ummelen 2004	X	X			Performance, amount of words uttered, quality of thinking aloud, number of problems found
Guan *et al.* 2006	X		X		Overlap with eye-tracking data
Eger *et al.* 2007		X	X	X	Task times, completion rate, number and type of problems found, subjective rating
Hertzum *et al.* 2009	X	X		X	Task times, eye and hand movements, workload, accuracy
Olmsted-Hawala *et al,* 2010a	X	X		X	Task solution rate, task times, subjective rating
Olmsted-Hawala *et al.* 2010b	X	X			Number of verbalised and non-verbalised frustrations and positive comments
Cooke 2010	X				Overlap with eye-tracking data
Zhao & McDonald 2010	X	X			Nature and amount of utterances
Hertzum & Holmegaard 2013	X			X	Task solution rate, task times, interruption performance, subtask behaviour, eye movements, mental workload
McDonald & Petrie 2013	X	X		X	Performance, workload, behaviour
Zhao *et al.* 2014	X	X			Performance, workload, perceptions of behaviour, number and type of utterances, number, type, severity and source of problems
Riihiaho (this thesis)		X		X	Performance, number of problems found, user rating, evaluators' certainty on judgements, users' feedback on test settings

Table 48 shows that performance measures, such as time on tasks and accuracy, have been very central measures also in the experiments on thinking aloud. Several tests have also utilised the eye-tracking systems to compare the users' verbal reports with the eye-tracking data, and thereby, to verify that the reports and behaviour do indeed match most of the time. Quite few studies, however, have asked the users directly to assess the settings and the naturalness of thinking aloud when performing the test tasks.

7.5 Effects of relaxed thinking aloud and moderator presence

The fifth research question in this thesis is:

5. What effect does concurrent relaxed thinking aloud method and the presence of a test moderator have in a usability test setting on users' performance and preferences, as well as on the evaluators' analyses?

7.5.1 Effects of relaxed thinking aloud

The studies of thinking aloud in the 1960's mostly focus on problem solving, and instead of just verbalising their thoughts in their working memory, the participants are asked to state reasons for every move they make with the Tower of Hanoi problem or some similar problems (*e.g.* Gagné & Smith 1962; Davis *et al.* 1968). Thereby, the thinking aloud in these experiments is on level 3 in which the models by Ericsson and Simon (1980) claim the thinking aloud to change the participants' normal behaviour. Indeed, these studies show that the participants perform better in the last, more complex tasks, if they have been thinking aloud in the previous tasks.

On the part of the level 1 and 2 thinking aloud in which participants verbalise their thoughts from the working memory without explaining reasons for their actions or choices, the model by Ericsson and Simon (1980) predicts that the thinking aloud has no effect on the performance except for possibly slowing it down. For example, the studies by Rhenius and Deffner (1990) on problem solving support this prediction, as their results show no difference in the accuracy of problem solving, but only longer solution times for the ones thinking aloud.

Most of the studies in the usability test settings support the model of Ericsson and Simon (1980). For example, in the studies by Hertzum *et al.* (2009), the test users' performance is slower in the conditions with thinking aloud compared to the silent working, but with classic thinking aloud, no other effect is found. Relaxed thinking aloud, on the other hand, changes the users' behaviour also in other ways by increasing general browsing of the web pages, and by increasing the experienced mental workload when compared to working silently. Also the results of the study by R.B. Wright and Converse (1992) show significant differences between the test groups. In this experiment, half of the users are asked to give reasons for each step in completing a task requiring level 3 thinking aloud, and the others perform silently. The participants with thinking aloud make fewer errors, and use less time on tasks, and this effect increases when the tasks become more complex. Thereby, the changes in users' behaviour again support the model by Ericsson and Simon.

Also the meta-analysis by Fox *et al.* (2011) points out the difference between level 1 or 2 thinking aloud and level 3 thinking aloud, as the results with the first ones show no statistical effect on users' performance other than possibly slowing it down, whereas level 3 thinking aloud requiring users to explain and argument their actions significantly improves the users' performance compared to silent conditions. However, the study by Hertzum and Holmegaard (2013) shows that even level 1 and 2 thinking aloud interacts with interruptions, so that visual interruptions significantly improve the task completion rate with thinking aloud compared to silent performance.

The results by Zhao *et al.* (2014), on their part, seem somewhat contradictory to the meta-analysis by Fox *et al.* (2011), as they find no difference in task performance between groups that are instructed to state their thoughts aloud, and groups that are given a more specific instruction to give "*explanations and content that is relevant to user experience*". When analysing the users' self-

assessments of the type of thinking aloud, however, participants in both conditions feel they are explaining quite a lot about the reasons for their difficulties. Thereby, both the conditions involve level 3 thinking aloud, and the users' behaviour might be different from normal in both conditions. As the users with specific instructions are asked to explain more than in the other condition, it is somewhat expected to get higher ratings for the mental workload in this condition. (Zhao *et al.* 2014)

On the part of the participant experiences, the study by Eger *et al.* (2007) show that the test users consider the concurrent thinking aloud as more unpleasant than performing in silence, and think that it slows down their performance although no statistical differences are found in the task times. The studies by Hertzum *et al.* (2009), and McDonald and Petrie (2013) support these findings, as the participants in these studies assess the mental demand significantly higher in the conditions with thinking aloud compared to silent performance.

In the experiment in this thesis, the level of thinking aloud is not systematically analysed, but according to the users' feedback, it mostly stays on levels 1 and 2. Thereby, the effect of slowing down the performance is quite expected in the thinking aloud conditions, and no other effects are expected or gained on the part of the users. From the evaluators' point of view, however, it is expected that the evaluators would be more doubtful about their judgements on usability problems, and especially about the causes of the problems, in test conditions where users perform silently. Nevertheless, the results show no significant differences between the evaluators' certainty in thinking aloud and silent conditions. Furthermore, the presence of a moderator seems to have more effect on this issue, but not statistically significant.

Still, at least with simple systems, it is worth considering whether to have users think aloud or let them work silently, as several users praise the opportunity to work silently in the same way as in a normal use situation. Since there is also a risk of reactivity in thinking aloud if the users start explaining and reasoning their actions, leaving all the comments and thoughts to be discussed only after each task is a reasonable alternative for the concurrent thinking aloud. Similarly to the retrospective thinking aloud studies, the users can be prompted in these post-task or post-test discussions with the goal of the task, or the solution they found. The discussions related to certain tasks should, anyhow, take place right after the tasks and not only at the end of the test sessions, because the users soon forget the task-specific details, and are not usually able to recollect them after additional tasks without the help of video clips or other concrete support for recall. Based on the studies by Trudel and Payne (1995), however, even these reviews between the tasks can effect the users' learning, and thereby, improve their performance in repetition tasks.

The discussions with usability practitioners brought forth also the customer's point of view to the thinking aloud method. If the representatives of the customer company want to observe the test sessions, they are quite lost with the users' problems if the users do not explain their thoughts and wonders aloud. Experienced usability evaluators may be able to interpret the prob-

lems from the users' gestures and facial expressions, and on the basis of their application domain knowledge, but the customer representatives usually cannot. Thus, an excellent opportunity may be wasted to motivate the developers to make the required changes to the system, and also to build empathy with the users. Table 49 summarises some of the benefits and disadvantages of the relaxed concurrent thinking aloud discussed in this thesis.

Table 49: Pros and cons of relaxed concurrent thinking aloud in usability testing.

Benefits	Disadvantages
+ Gives information about the user's experiences + Gives explanations of the causes of the problems + Makes discussions with the moderator more natural + Silence usually reveals serious problems	- May change the user's behaviour either improving or impairing the performance in accuracy - May slow down the user's performance - Negative effect on completion rates - User's perceive as unnatural - Increases the user's mental demand

7.5.2 Effects of moderator presence

The studies of problem solving by Davis *et al.* (1968) show that the presence of a test moderator has either an insignificant effect or a positive effect in the test participants' performance. In usability test settings, Held and Biers (1992) find that expert users give more negative ratings of the system if a moderator is present, whereas with novice users, the effect is positive but not significant. In another experiment by Hackman and Biers (1992), no differences are found in the users' performance regardless of them working alone, in pairs or with a moderator present. The results of Schulte-Mecklenbeck and Huber (2003), on the other hand, indicate that the presence of a test moderator can encourage participants to search for more information than on their own, and also to complete the test with more certainty compared to users performing the tasks on their own. The study by Sonderegger and Sauer (2009) also suggest that a good rapport between the moderator and the test user may enhance performance. The study by Eger *et al.* (2007), on the other hand, indicates that the moderator's presence has a negative effect on the test users if they are asked to think aloud, making the experience more unnatural and unpleasant than in retrospective conditions.

The experiment in this thesis, on its part, indicates that the presence of a test moderator does not have a significant effect on time on task, number of usability problems faced, or the perceived naturalness of performing silently or using thinking aloud. However, the effect on users' subjective ratings is significant, as the users having the moderator in the same room while doing the test tasks rate the system preferences significantly higher than the users working alone in the test room. Still, the participants in the experiment value the presence of the moderator, and consider it as a natural part of usability testing. Some participants also comment in the post-test questionnaire that during the test, they were unsure if they had done the tasks right and learned only in the end discussions that some of the tasks were indeed solved incorrectly. With the moderator present, these situations could have been detected right after the tasks, and also the reasons for the problems could have been discussed with

more detail than in the end discussions. Therefore, from the evaluator's or moderator's point of view, being next to the user and being able to ask for clarifying questions when the actions are still fresh in mind is an elementary part of finding the problems and their reasons, as well as their relevancy in the evaluated system. Table 50 summarises some of the benefits and disadvantages of the moderator presence discussed in this thesis.

Table 50: Pros and cons of having a test moderator next to the test user in a usability test.

Benefits	Disadvantages
+ Enables discussions with the user within and between the test tasks + May motivate the test users to do more, and to finish all the test tasks + Users may get instant feedback on their task solution	- May change the user's behaviour by improving the performance in accuracy - May affect on the users' feedback on the system - Moderator's reminders to think aloud may disturb and even frustrate the test users

7.5.3 Considerations on contextual factors of usability testing

Based on the results of various studies of contextual factors of usability testing, no straightforward instructions or standard procedures can be given in selecting suitable methods and settings. Some guidelines, however, can be provided. For example, Table 46 in this thesis gives some ideas on which methods could be applied if the goals for the evaluation are clear. The main decisions regarding the test settings include the test environment and the source of the test tasks. If the users are given the opportunity to explore and try out their own tasks, they are also more able to assess the system from other points of view in addition to usability, such as utility and social value.

The studies of various contextual factors affecting the results of usability tests demonstrate how even very small changes can have substantial effects on the results. For example, the study by Raita and Oulasvirta (2011) shows how test users try for more and do not give up so easily with the tasks if the leaflets they have seen about the system have indicated that the system is quite complex to use. The ones who have read positive comments, on the other hand, give up more easily, as the system does not meet their expectations. Cordes (2001), on his part, varies the instructions given to the test users telling half of them that the system does not necessarily support all the functions required to perform the tasks, whereas the other half is not forewarn about this possibility. After these instructions, the ones that have been forewarn, give up more often and more quickly, whereas the others try significantly longer to find the right solution (Cordes 2001). In usability tests, it is quite common that the users persistently try to get the tasks done, and admit even themselves that in normal use situations, they would probably have given up trying.

On the part of learning effects, the studies by Trudel and Payne (1995), show that when test participants are learning to use a new system, the ones who review their actions every now and then, perform better than the ones doing their tasks without breaks or verbalisations. Also focusing to one task at a time improves the learning and performance. These results indicate that concurrent thinking aloud and even retrospective thinking aloud between the

test tasks may improve learning and bias the results of the subsequent tasks, as well as the convention of giving one test task at a time for focus to the users instead of several parallel tasks to explore in parallel.

In addition to the instructions and information given about the system before its use, also quite small factors of the system, such as the price, apparent novelty, or the level of finishing affect the results. For example, Sonderegger and Sauer (2013) show that Swiss users expect more expensive systems to be also more usable, whereas East German users consider more expensive systems to be more complex, and, thereby, more complicated to use. Apparent novelty also changes the users' assessments on the usability of the product. For example, participants in the studies by Mugge and Schoormans (2012) assess the usability of a novel looking product as poorer than the traditional looking product just based on the looks and the product information. With prototypes, the level of fidelity can also affect the perceived attractiveness of the system. In the studies by Sauer and Sonderegger (2009), the finished product is assessed as less attractive than the low- and medium-fidelity prototypes with the same contents. Thereby, it seems that the test users fill in the gaps between the prototype and the finished product with positive assumptions, and therefore rate them better than the finished version. On the part of the prototypes, it is also recommended to have several design alternatives for the users to compare in the early phase of development (Tohidi *et al.* 2006a). To get ideas from the users, Tohidi *et al.* (2006b) also recommend to give the users an opportunity to sketch their own designs on a sheet of paper.

7.6 State of usability testing and its validity

Finally, the research problem in this thesis is:

> Are the established practices of usability testing appropriate and valid, that is they do not change the phenomena they are studying?

The problems and challenges of usability testing reported in academic forums focus on sampling; context of use; use over time; assessment of utility and value; and misuse of the term "usability testing" and the way it is applied. Many of these problems are not academically challenging as they require more the attention of the practitioners in the field rather than further research. For example, the term usability testing must be used accurately and systematically only in the meaning of testing with users instead of all the evaluation methods; and the method must be applied with thought instead of "by the rule" just to get a paper published.

On the part of the sampling problems, various instructions, guidelines and even formulas are available to help in selecting test users and test tasks. For example, the formulas for estimating the required number of test participants (J.R. Lewis 2001b) can be used if it is essential to find as many problems as possible. However, the same effect of increasing the number of problems found can be achieved by having more test tasks (Lindgaard & Chattratichart 2007) and by improving the quality and relevance of these test tasks (Skov &

Stage 2012). Despite the number of test participants, they must be selected with care to represent the real users, and from the groups that have been selected as the focus for the current evaluation.

The sampling problems also include the problem with different kinds of evaluators interpreting things differently, and ending up to diverse results regardless of the methods and guidelines. Although there is no exhaustive solution to fix the evaluator effect, it is recommended to have several evaluators in the analysis phase to be able to combine various viewpoints, and to discuss on differences. Rapid analysis methods, such as Instant Data Analysis (Kjeldskov *et al.* 2004), can be used to streamline the analysis phase, and at the same time, to enable discussions about the problems as the users' actions are still fresh in mind. Indeed, the results of the study by Kjeldskov *et al.* (2004) are very encouraging as they find 85% of the critical usability problems in only 10% of the time needed for the traditional video analyses.

The dogmas by Hornbæk (2010) are academically interesting, but do not specifically require further research. Instead, they are something that especially the researchers need to be aware of, and take into account when trying to compare different usability evaluation methods. Comparisons as such are almost impossible to do, since the methods are developed for different kinds of purposes, and therefore, focus on different matters. The results depend on the complexity and completeness of the evaluated system; resources available including the evaluators' expertise; and many other factors. The relevancy and the severity of the problems found is finally determined only in the real use in the real use contexts, and this is rarely possible to measure. Therefore, some methods or results are needed as baselines for the others, but this again, is something that needs to be clearly reported. It also needs to be considered how well the baselines correspond to the real use situations. Furthermore, the impact on the development should be considered in these comparisons.

Some problems, such as utility assessment, can be solved to some extent with a set of suitable methods presented in this thesis, such as informal walkthrough, Experience Clip (Isomursu *et al.* 2004) and utility inspection method (Johannessen & Hornbæk 2014). These methods give the test participants some room to explore with the system, and to try out their own ideas. Thereby, the usefulness and even social value are more easy to assess than with a set of predefined test tasks. On the part of studying the evolving use of a new system, these methods can be used in repetition tests with a break of few weeks or even longer. Questionnaires and self-reporting, such as the critical incidents reporting tool by Hartson and Castillo (1998), can also be utilised for longitudinal studies. These methods do not fulfil all the requirements for longitudinal studies and utility assessment, but they do give an applicable starting point, and complement the brief one-off usability test sessions with predefined test tasks.

The most challenging academic problems in usability test setting, in my opinion, lay in the effects of central contextual factors of usability testing, namely the use of the thinking aloud method and the presence of a test moderator. However, the effects of these issues are still somewhat unclear, al-

though they are widely used in the usability testing practice. After the studies by R.B. Wright and Biers in 1992, it took over 10 years before academic studies of thinking aloud started to grow in number. Despite the several studies, the topic is so multifaceted that the results are still somewhat contradictory. As J.R. Lewis (2014) concludes his review of the thinking aloud studies: "*There is still work to do before we will have a deep understanding of the effects of current variations in TA protocols.*"

On the part of the effect of a moderator presence, only the studies by Held and Biers (1992), and Sonderegger and Sauer (2009) focus on usability testing. Therefore, the results of the experiment in this thesis are quite fundamental in the field of HCI, and complement very well the studies of the effects of various contextual factors in usability testing.

So, what does this thesis recommend to do on the part of the use of concurrent thinking aloud or the presence of a test moderator? Similarly to the review by J.R. Lewis (2014) it recommends to make a clear division between summative and formative usability testing. Summative evaluation may require strict test procedures with exactly similar tasks and settings for all the test participants, but formative evaluation allows more freedom and flexibility to the methods and practices. As Nørgaard and Hornbæk (2006) conclude: "*We encourage evaluators to change set-up or make alterations to the prototype in the middle of a test if they believe it will help them answer important questions about the use of the system.*"

Thereby, despite the risk of a positive effect in subjective ratings, this thesis recommends to have a moderator near the test user in formative usability testing, since being near the user gives the moderator an exquisite opportunity to ask clarifying questions after each task, when the experiences are still fresh in user's mind. On the part of the thinking aloud, it is worth considering whether to use thinking aloud with the risk of changing users' behaviour, as the concurrent relaxed thinking aloud in this experiment was mostly reported as an unnatural extra effort by the users, without special benefit to the evaluators. Still, it gives more information about the problems for the customers observing the tests, and thereby, may motivate the designers to make the required changes to the system. Concurrent thinking aloud also facilitates natural discussions between the test user and the moderator during the test session. Therefore, if performance measurements are required, classic concurrent thinking aloud should be used with minimal interventions. In formative testing, however, the more explanatory relaxed thinking aloud can be used, as long as its potential effects on users' performance are kept in mind in addition to the potential positive effect of the moderator presence on users' subjective ratings.

References

Akers, D., Simpson, M., Jeffries, R. and Winograd, T. (2009) Undo and erase events as indicators of usability problems. In *Proceedings of the 27th international conference on Human factors in computing systems* (CHI '09). ACM, New York, USA, pp. 659-668. ISBN: 978-1-60558-246-7, DOI= 10.1145/1518701.1518804.

Akers, D., Jeffries, R., Simpson, M. and Winograd, T. (2012) Backtracking events as indicators of usability problems in creation-oriented applications. *ACM Transactions on Computer-Human Interaction*, Vol. 19, No. 2, Article 16, pp. 16:1-16:40. ISSN: 1073-0516, DOI= 10.1145/2240156.2240164.

Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E. and Moret-Bonillo, V. (2010) A context-of-use taxonomy for usability studies. *International Journal of Human-Computer Interaction*, Vol. 26, No. 10, pp. 941-970. ISSN: 1044-7318, DOI= 10.1080/10447318.2010.502099.

Archer, N.P. and Yuan, Y. (1995) Comparing telephone-computer interface designs: Are software simulations as good as hardware prototypes? *International Journal of Human-Computer Studies*, Vol. 42, No. 2, pp. 169-184. ISSN: 1071-5819, DOI= 10.1006/ijhc.1995.1008.

Baddeley, A. (1992) Working memory. *Science*, Vol. 255, No. 5044, pp. 556-559. ISSN: 0036-8075, DOI= 10.1126/science.1736359.

Bailey, R.W. (1996) *Human performance engineering: Designing high quality, professional user interfaces for computer products, applications, and systems*. 3rd ed. Prentice Hall, Upper Saddle River, USA, pp. 199-226.

Bangor, A., Kortum, P.T. and Miller, J.T. (2008) An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, Vol. 24, No. 6, pp. 574-594. ISSN: 10447318, DOI= 10.1080/10447310802205776.

Barker, R.T. and Biers, D.W. (1994) Software usability testing: Do user self-consciousness and the laboratory environment make any difference? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 38, No. 17, pp. 1131-1134. ISSN: 1541-9312, DOI= 10.1177/154193129403801713.

Barkhuus, L. and Rode, J.A. (2007) From mice to men – 24 years of evaluation in CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07). ACM Press, New York, USA. ISBN: 978-1-59593-593-9, DOI= 10.1145/1240624.2180963.

Barnum, C.M. (2011) *Usability testing essentials: Ready, set... test!* Elsevier/Morgan Kaufmann, Amsterdam, The Netherlands, 382 p. ISBN: 978-0-12-375092-1.

Bastien, J.M.C. and Scapin, D.L. (1995) Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction*, Vol. 7, No. 2, pp. 105-121. ISSN: 10447318, Accession number= 7389056.

Bawa, J. (1994) Comparative usability measurement: The role of the usability lab in PC Magazine UK and PC/ Computing. *Behaviour & Information Technology*, Vol. 13, No. 1-2, pp. 17-19. ISSN: 0144-929X, DOI= 10.1080/01449299408914579.

Ben-Bassat, T., Meyer, J. and Tractinsky, N. (2006) Economic and subjective measures of the perceived value of aesthetics and usability. *ACM Transactions on Computer-Human Interaction*, Vol. 13, No. 2, pp. 210-234. ISSN: 1073-0516, DOI= 10.1145/1165734.1165737.

Benyon, D., Turner, P. and Turner, S. (2005) *Designing interactive systems: People, activities, contexts, technologies*. Addison-Wesley, Harlow, UK, 789 p. ISBN: 0-321-11629-1.

Bevan, N. and Macleod, M. (1994) Usability measurement in context. *Behaviour & Information Technology*, Vol. 13, no. 1-2, pp. 132-145. ISSN: 0144-929X, DOI= 10.1080/01449299408914592.

Beyer, H.R. and Holtzblatt, K. (1995) Apprenticing with the customer. *Communications of the ACM*, Vol. 38, No. 5, pp. 45-52. ISSN: 0001-0782, DOI = 10.1145/203356.203365.

Beyer, H.R. and Holtzblatt, K. (1998) *Contextual design: Defining customer-centered systems*. Morgan Kaufmann Publishers, San Francisco, USA, 472 p. ISBN: 1-55860-411-1.

Bias, R. (1994) The pluralistic usability walkthrough: Coordinated empathies. In J. Nielsen & R.L. Mack (Eds.) *Usability inspection methods*. John Wiley & Sons, New York, USA, pp. 63-76. ISBN: 0-471-01877-5.

Bloomer, S., Croft, R. and Kieboom, H. (1997) Strategic usability: Introducing usability into organisations. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '97). ACM, New York, USA, pp. 156-157. ISBN: 0-89791-926-2, DOI= 10.1145/1120212.1120320.

Booth, P.A. (1989) *An introduction to human-computer interaction*. Lawrence Erlbaum Associates, Hove, UK. 268 p. ISBN: 0-86377-123-8.

Boothe, C., Strawderman, L. and Hosea, E. (2013) The effects of prototype medium on usability testing. *Applied Ergonomics*, Vol. 44, No. 6, pp. 1033-1038. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2013.04.014.

Boren, T. and Ramey, J. (2000) Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, Vol. 43, No. 3, pp. 261-278. ISSN: 0361-1434, DOI=10.1109/47.867942.

Bowers, V.A. and Snyder, H.L. (1990) Concurrent versus retrospective verbal protocol for comparing window usability. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 34, No. 17, pp. 1270-1274. ISSN: 1541-9312, DOI= 10.1177/154193129003401720.

Brooke, J. (1996) SUS: A “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.) *Usability evaluation in industry*. Taylor & Francis, London, UK, pp. 189–194. ISBN: 0748404600.

Brooke, J. (2013) SUS: A retrospective. *Journal of Usability Studies*, Vol. 8, No. 2, pp. 29-40. ISSN: 1931-3357. Available on http://www.usabilityprofessionals.org/upa_publications/jus/2013february/JUS_Brooke_February_2013.pdf.

Brooks, P. (1994) Adding value to usability testing. In J. Nielsen & R.L. Mack (Eds.) *Usability inspection methods*. John Wiley & Sons, New York, USA, pp. 255-271. ISBN: 0-471-01877-5.

Brown, L. and Gardner, J. (1985) Using citation analysis to assess the impact of journals and articles on contemporary accounting research (CAR). *Journal of Accounting Research*, Vol. 23, No. 1, pp. 84–109. ISSN: 00218456. Available on http://www.jstor.org/stable/2490908.

Bruun, A. and Stage, J. (2014) Barefoot Usability Evaluations. *Behaviour & Information Technology*. ISSN: 0144-929X, DOI= 10.1080/0144929X.2014.883552, Published online: February 17th 2014.

Butler, K.A. (1996) Usability engineering turns 10. *interactions*, Vol. 3, no. 1, pp. 58-75. ISSN: 1072-5520, DOI= 10.1145/223500.223513.

Butler, M.B. (1996) Getting to know your users: Usability roundtables at Lotus Development. *interactions*, Vol. 3, no. 1, pp. 23-30. ISSN: 1072-5520, DOI= 10.1145/223500.223507.

Bødker, S. (2006) When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction: Changing roles* (NordiCHI '06), A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh & D. Svanaes (Eds.). ACM, New York, USA, pp. 1-8. ISBN: 1-59593-325-5, DOI= 10.1145/1182475.1182476.

Bødker, S. and Grønbæk, K. (1991) Design in action: From prototyping by demonstration to cooperative prototyping. In J. Greenbaum & M. Kyng (Eds.) *Design at work: Cooperative design of computer systems*. Lawrence Erlbaum Associates, Hillsdale, USA, pp. 197-218. ISBN: 0-8058-0612-1.

Bødker, S. and Madsen, K. (1998) Context: An active choice in usability work. *interactions*, Vol. 5, no. 4, pp. 17-25. ISSN: 1072-5520, DOI= 10.1145/278465.278469.

Camilli, G. and Hopkins, K. D. (1978) Applicability of chi-square to 2×2 contingency tables with small expected cell frequencies. *Psychological Bulletin*, Vol. 85, No. 1, pp. 163-167. ISSN: 0033-2909, DOI= 10.1037/0033-2909.85.1.163.

Carroll, J.M. (1997) Human–computer interaction: Psychology as a science of design. *International Journal of Human-Computer Studies*, Vol. 46, No. 4, pp. 501-522. ISSN: 1071-5819, DOI= 10.1006/ijhc.1996.0101.

Catani, M.B. and Biers, D.W. (1998) Usability Evaluation and Prototype Fidelity: Users and Usability Professionals. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*. Vol. 42, No. 19, pp. 1331-1335. ISSN: 1541-9312, DOI= 10.1177/154193129804201901.

Caulton, D.A. (2001) Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology*, Vol. 20, No. 1, pp. 1-7. ISSN: 0144-929X, DOI= 10.1080/01449290010020648.

Choros, K. and Muskala, M. (2009) Block map technique for the usability evaluation of a website. In N.T. Nguyen, R. Kowalczyk & S-M. Chen (Eds.) *Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems*. Springer, Berlin, Germany, pp. 743-751. ISBN: 978-3-642-04440-3, DOI= 10.1007/978-3-642-04441-0_65.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q. and Yammiyavar, P. (2009) Cultural cognition in usability evaluation. *Interacting with Computers*, Vol. 21, No. 3, pp. 212-220. ISSN: 0953-5438, DOI= 10.1016/j.intcom.2009.05.003.

Cockton, G. (2006) Designing worth is worth designing. In *Proceedings of the 4th Nordic conference on Human-computer interaction: Changing roles* (NordiCHI '06), A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh & D. Svanaes (Eds.). ACM, New York, USA, pp. 165-174. ISBN: 1-59593-325-5, DOI= 10.1145/1182475.1182493.

Cooke, L. (2010) Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication*, Vol. 53, No. 3, pp. 202-215. ISSN: 0361-1434, DOI= 10.1109/TPC.2010.2052859.

Cordes, R.E. (2001) Task-selection bias: A case for user-defined tasks. *International Journal of Human-Computer Interaction*, Vol. 13, No. 4, pp. 411-419. ISSN: 1044-7318, Accession number = 6411688.

CORE (2010a) *Conference rankings*. The Computing Research and Education Association of Australasia. [e-publication] <http://core.edu.au/cms/images/downloads/conference/08sortrankacronymERA2010_conference_list.pdf> retrieved 6.10.2011.

CORE (2010b) *Journal rankings*. The Computing Research and Education Association of Australasia. [e-publication] <http://core.edu.au/index.php/categories/journals/12> retrieved 6.10.2011.

Culpepper, L.M. (1975) A system for reliable engineering software. In *Proceedings of the international conference on Reliable software (SIGPLAN '75)*. ACM, New York, USA, pp. 186-192. ISSN: 0362-1340, DOI= 10.1145/800027.808438.

Cuomo, D.L. and Bowen, C.D. (1994) Understanding usability issues addressed by three user-system interface evaluation techniques. *Interacting with Computers*, Vol. 6, No. 1, pp. 86-108. ISSN: 0953-5438, DOI= 10.1016/0953-5438(94)90006-X.

Davis, J.H., Carey, M.H., Foxman, P.N. and Tarr, D.B. (1968) Verbalization, experimenter presence, and problem solving. *Journal of Personality and Social Psychology*, Vol. 8, No. 3, pp. 299-302. ISSN: 0022-3514, DOI= 10.1037/h0025519.

Dehlholm, F. (1992) Picture analysis of screen images. In J. Leponiemi (Ed.) *Preceedings of the NordDATA `92 conference*. Pitky, Tampere, Finland. pp. 353-359.

Desurvire, H. (1994) Faster, cheaper!! Are usability inspection methods as effective as empirical testing? In J. Nielsen & R.L. Mack (Eds.) *Usability inspection methods*. John Wiley & Sons, New York, USA, pp. 173-202. ISBN: 0-471-01877-5.

Desurvire, H., Lawrence, D. and Atwood M. (1991) Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. *ACM SIGCHI Bulletin*, Vol. 23, no. 4, pp. 58-59. ISSN: 0736-6906, DOI= 10.1145/126729.1056062.

Dicks, R.S. (2002) Mis-usability: On the uses and misuses of usability testing. In *Proceedings of the 20th annual international conference on Computer documentation* (SIGDOC '02). ACM, New York, USA, pp. 26-30. ISBN: 1-58113-543-2, DOI= 10.1145/584955.584960.

Dix, A., Finlay, J., Abowd, G.D. and Beale, R. (2004) *Human-computer interaction*, 3rd ed. Harlow, Prentice Hall, 834 p. ISBN: 0130-461091.

Downey, L.L. (2007) Group usability testing: Evolution in usability techniques. *Journal of Usability Studies*, Vol. 2, No. 3, pp. 133-144. ISSN: 1931-3357. Available on http://www.usabilityprofessionals.org/upa_publications/jus/2007may/group-utests.html.

Duh, H.B.L., Tan, G.C.B. and Chen, V.H.H. (2006) Usability evaluation for mobile device: A comparison of laboratory and field tests. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services* (MobileHCI '06). ACM, New York, USA, pp. 181-186. ISBN: 1-59593-390-5, DOI= 10.1145/1152215.1152254.

Dumas, J.S. (2007) The great leap forward: The birth of the usability profession (1988-1993). *Journal of Usability Studies*, Vol. 2, No. 2, pp. 54-60. ISSN: 1931-3357. Available on http://www.upassoc.org/upa_publications/jus/2007_february/dumas_birth_of_usability_profession.pdf.

Dumas, J.S. and Loring, B.A. (2008) *Moderating usability tests: Principles and practices for interacting*. Elsevier/Morgan Kaufmann, Amsterdam, The Netherlands, 185 p. ISBN: 978-0-12-373933-9.

Dumas, J.S., Molich, R. and Jeffries, R. (2004) Describing usability problems: Are we sending the right message? *interactions*, Vol. 11, No. 4, pp. 24-29. ISSN: 1072-5520, DOI= 10.1145/1005261.1005274.

Dumas, J.S. and Redish, J.C. (1993) *A practical guide to usability testing*. Ablex Publishing Corporation, Norwood, USA, 412 p. ISBN: 0-89391-991-8.

Duncker, K. (1945) On problem solving. *Psychological Monographs*, Vol. 58, No. 270, pp. i-113. ISSN: 0096-9753, DOI= 10.1037/h0093599. (Original: Duncker, K. (1935). Zur Psychologie des produktiven Denkens. Springer, Berlin, Germany).

Eger, N., Ball, L.J., Stevens, R. and Dodd, J. (2007) Cueing retrospective verbal reports in usability testing through eye-movement replay. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it - Volume 1* (BCS-HCI '07). British Computer Society, Swinton, UK, pp. 129-137. ISBN: 978-1-902505-94-7, ACM id= 1531312.

Ehrlich, K., Butler, M.B. and Pernice, K. (1994) Getting the whole team into usability testing, *IEEE Software*, Vol. 11, No. 1, pp. 89-91. ISSN: 0740-7459, DOI= 10.1109/52.251216.

Elling, S., Lentz, L. and de Jong, M. (2012) Combining concurrent think-aloud protocols and eye-tracking observations: An analysis of verbalizations and silences. *IEEE Transactions on Professional Communication*, Vol. 55, No. 3, pp. 206-220, ISSN: 0361-1434, DOI= 10.1109/TPC.2012.2206190.

Ellis, T.J. and Levy, Y. (2008) Framework of problem-based research: A guide for novice researchers on the development of a research-worthy problem. *Informing Science Journal*, Vol. 11, pp. 17-33. ISSN: 1547-9684.

Ericsson, K.A. and Simon, H.A. (1980) Verbal reports as data. *Psychological Review*, Vol. 87, no. 3, pp. 215-251. ISSN: 0033-295X, DOI= 10.1037/0033-295X.87.3.215.

Ericsson, K. A. and Simon, H. A. (1984) *Protocol analysis: Verbal reports as data*. MIT-press, Cambridge, USA, 426 p. ISBN: 0-262-05029-3.

Eysenck, M.W. and Keane, M.T. (1990) *Cognitive psychology: A student's handbook*. Lawrence Erlbaum Associates, Hove, UK, 557 p. ISBN: 0-86377-154-8.

Finstad, K. (2006) The System Usability Scale and non-native English speakers. *Journal of Usability Studies*, Vol. 1, No. 4, pp. 185-188. ISSN: 1931-3357. Available on http://www.usabilityprofessionals.org/upa_publications/jus/2006_august/finstad_sus_non_native_speakers.pdf.

Fox, M.C., Ericsson, K. A. and Best, R. (2011) Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, Vol. 137, No. 2, pp. 316-344. ISSN: 0033-2909, DOI= 10.1037/a0021663.

Freeman, B. (2011) Triggered think-aloud protocol: Using eye tracking to improve usability test moderation. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI '11). ACM, New York, USA, pp. 1171-1174. ISBN: 978-1-4503-0228-9, DOI= 10.1145/1978942.1979117.

Frøkjær, E., Hertzum, M. and Hornbæk, K. (2000) Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '00). ACM, New York, USA, pp. 345-352. ISBN: 1-58113-216-6, DOI= 10.1145/332040.332455.

Frøkjær, E. and Hornbæk, K. (2005) Cooperative usability testing: Complementing usability tests with user-supported interpretation sessions. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '05). ACM, New York, NY, USA, pp. 1383-1386. ISBN: 1-59593-002-7, DOI= 10.1145/1056808.1056922.

Frøkjær, E. and Hornbæk, K. (2008) Metaphors of human thinking for usability inspection and design. *ACM Transactions on Computer-Human Interaction,* Vol. 14, No. 4, Article 20, pp. 20:1-20:33. ISSN: 1073-0516, DOI= 10.1145/1314683.1314688.

Følstad, A. (2007). Work-Domain experts as evaluators: Usability inspection of domain-specific work-support systems. *International Journal of Human-Computer Interaction*, Vol. 22, No. 3, pp. 217-245. ISSN: 1044-7318, DOI= 10.1080/10447310701373048.

Følstad, A. and Hornbæk, K. (2010) Work-domain knowledge in usability evaluation: Experiences with Cooperative Usability Testing. *The Journal of Systems and Software*, Vol. 83, No. 11, pp. 2019-2030. ISSN: 0164-1212, DOI= 10.1016/j.jss.2010.02.026.

Følstad, A., Law, E.L-C. and Hornbæk, K. (2010) Analysis in usability evaluations: An exploratory study. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (NordiCHI '10). ACM, New York, USA, pp. 647-650. ISBN: 978-1-60558-934-3, DOI= 10.1145/1868914.1868995.

Følstad, A., Law, E.L-C. and Hornbæk, K. (2012) Analysis in practical usability evaluation: A survey study. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, USA, pp. 2127-2136. ISBN: 978-1-4503-1015-4, DOI= 10.1145/2208276.2208365.

Gagné, R.M. and Smith, E.C. Jr. (1962) A study of the effects of verbalization on problem solving. *Journal of Experimental Psychology*, Vol. 63, No. 1, pp. 12-18. ISSN: 0096-3445, DOI= 10.1037/h0048703.

Gilmore, D., Cockton, G., Churchill, E., Kujala, S., Henderson, A. and Hammontree, M. (2008) Values, value and worth: Their relationship to HCI? In *CHI '08 extended abstracts on Human factors in computing systems (CHI EA '08)*. ACM, New York, USA, pp. 3933-3936. ISBN: 978-1-60558-012-8, DOI= 10.1145/1358628.1358960.

Gomoll, K. (1990) Some techniques for observing users. In *The art of human-computer interface design*, B. Laurel and S.J. Mountford (Eds.). Addison-Wesley Longman Publishing, Boston, USA, pp. 85-90. ISBN: 0201517973.

Gray, W.D. and Salzman, M.C. (1998) Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, Vol. 13, No. 3, (Experimental Comparisons of Usability Evaluation Methods: A Special Issue of Human-Computer Interaction, ISBN: 0-8058-9813-1), pp. 203-261. ISSN: 0737-0024, DOI= 10.1207/s15327051hci1303_2.

Greenberg, S. and Buxton, B. (2008) Usability evaluation considered harmful (some of the time). In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (CHI '08). ACM, New York, USA, pp. 111-120. ISBN: 978-1-60558-011-1, DOI= 10.1145/1357054.1357074.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A. and 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* (NordiCHI '04). ACM, New York, USA, pp. 207-215. ISBN: 1-58113-857-1, DOI= 10.1145/1028014.1028046.

Guan, Z., Lee, S., Cuddihy, E. and Ramey, J. (2006) The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '06), R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries & G.M. Olson (Eds.). ACM, New York, USA, pp. 1253-1262. ISBN: 1-59593-372-7, DOI= 10.1145/1124772.1124961.

van den Haak, M.J. and de Jong, M.D.T. (2005) Analyzing the interaction between facilitator and participants in two variants of the think-aloud method. In *Proceedings of the International Professional Communication Conference, 2005* (IPCC 2005). IEEE, New York, USA, pp. 323-327. ISBN: 0-7803-9027-X, DOI=

10.1109/
IPCC.2005.1494192.

van den Haak, M.J., de Jong, M.D.T. and Schellens, P.J. (2003) Retrospective *vs*. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, Vol. 22, No. 5, pp. 339-351. ISSN: 0144-929X, DOI= 10.1080/0044929031000.

Hackman G.S. and Biers, D.W. (1992) Team usability testing: Are two heads better than one? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 36, No. 16, pp. 1205-1209. ISSN: 1541-9312, DOI= 10.1177/154193129203601605.

Han, S.H., Yun, M.H., Kim, K-J. and Kwahk, J. (2000) Evaluation of product usability: Development and validation of usability dimensions and design elements based on empirical models. *International Journal of Industrial Ergonomics*, Vol. 26, No. 4, pp. 477-488. ISSN: 0169-8141, DOI= 10.1016/S0169-8141(00)00019-6.

Hansen, M. (1991) Ten steps to usability testing. In *Proceedings of the conference on 1991 ACM ninth annual international conference on systems documentation* (SIGDOC '91). ACM, New York, USA, pp. 135-139. ISBN: 0-89791-452-X, DOI= 10.1145/122778.122798.

Happ, A.J. (1994) Usability foresight: Strategic usability planning: A special interest group meeting report. SIGCHI *Bulletin*, Vol. 26, No. 1, pp. 17-21. ISSN: 0736-6906, DOI= 10.1145/181526.181527.

Hartmann, J., Sutcliffe, A. and De Angeli, A. (2008) Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer-Human Interaction*, Vol. 15, No. 4, Article 15, pp. 15:1-15:30. ISSN: 1073-0516, DOI=10.1145/1460355.1460357.

Hartson, H.R., Andre, T.S. and Williges, R.C. (2003) Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, Vol. 13, No. 4, pp. 145-181. ISSN: 1044-7318, DOI= 10.1207/S15327590IJHC1501_13.

Hartson, H.R. and Castillo, J.C. (1998) Remote evaluation for post-deployment usability improvement. In *Proceedings of the working conference on Advanced visual interfaces* (AVI '98), T. Catarci, M.F. Costabile, G. Santucci & L. Tarantino (Eds.). ACM, New York, USA, pp. 22-29. DOI= 10.1145/948496.948499.

Hassenzahl, M. (2004) The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, Vol. 19, No. 4, pp. 319-349. ISSN: 0737-0024, DOI: 10.1207/s15327051hci1904_2.

Hassenzahl, M. and Monk, A. (2010) The inference of perceived usability from beauty. *Human-Computer Interaction*, Vol. 25, No. 3, pp. 235-260. ISSN: 0737-0024, DOI= 10.1080/07370024.2010.500139.

Hassenzahl, M. and Sandweg, N. (2004) From mental effort to perceived usability: Transforming experiences into summary assessments. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '04). ACM, New York, USA, pp. 1283-1286. ISBN: 1-58113-703-6, DOI= 10.1145/985921.986044.

Held, J.E. and Biers, D.W. (1992) Software usability testing: Do evaluator intervention and task structure make any difference? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 36, No. 16, pp. 1215-1219. ISSN: 1541-9312, DOI= 10.1177/154193129203601607.

Hertzum, M. (2006) Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-*

Computer Interaction, Vol. 21, No. 2, pp. 125-146. ISSN: 1044-7318, DOI= 10.1207/s15327590ijhc2102_2.

Hertzum, M. (2010) Images of usability. *International Journal of Human-Computer Interaction*, Vol. 26, No. 6, pp. 567-600. ISSN: 1044-7318, DOI= 10.1080/10447311003781300.

Hertzum, M. and Clemmensen, T. (2012) How do usability professionals construe usability? *International Journal of Human-Computer Studies*, Vol. 70, No. 1, pp. 26–42. ISSN: 1071-5819, DOI= 10.1016/j.ijhcs.2011.08.001.

Hertzum, M. and Frøkjær, E. (1996) Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions on Computer-Human Interaction*, Vol. 3, No. 2, pp. 136-161. ISSN: 1073-0516, DOI= 10.1145/230562.230570.

Hertzum, M., Hansen, K. D., and Andersen, H.H.K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, Vol. 28, No. 2, pp. 165-181. ISSN: 0144-929X, DOI= 10.1080/01449290701773842.

Hertzum, M. and Holmegaard, K.D. (2013) Thinking aloud in the presence of interruptions and time constraints. *International Journal of Human-Computer Interaction*, Vol. 29, No. 5, pp. 351-364. ISSN: 1044-7318, DOI= 10.1080/10447318.2012.711705.

Hertzum, M. and Jacobsen, N.E. (2001) The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, Vol. 13, No. 4, pp. 421-443. ISSN: 1044-7318, DOI= 10.1207/S15327590IJHC1304_05.

Hertzum, M., Molich, R. and Jacobsen, N.E. (2014) What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology*, Vol. 33, No. 2, pp. 144-162. ISSN: 0144-929X, DOI= 10.1080/0144929X.2013.783114.

Hewett, T.T. (1986) The role of iterative evaluation in designing systems for usability. In *Proceedings of the Second Conference of the British Computer Society, human computer interaction specialist group on People and computers: designing for usability*, M.D. Harrison & A.F. Monk (Eds.). Cambridge University Press, New York, USA, pp. 196-214. ISBN: 0-521-33259-1.

Hewett, T.T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G. Verplank, W. (1992) *ACM SIGCHI Curricula for Human-Computer Interaction*, ACM Press, New York, USA. ISBN: 0-89791-474-0, also available on http://sigchi.org/cdg/.

Holleran, P.A. (1991) A methodological note on pitfalls in usability testing. *Behaviour & Information Technology*, Vol. 10, No. 5, pp. 345-357. ISSN: 0144-929X, DOI= 10.1080/01449299108924295.

Holtzblatt, K., Wendell, J. B. and Wood, S. (2005) *Rapid contextual design: a how-to guide to key techniques for user-centered design*. Morgan Kaufmann, Elsevier, Burlington, USA. 321 p. ISBN: 9780123540515.

Holtzblatt, K. (2011) What makes things cool? Intentional design for innovation. *interactions*, Vol. 18, No. 6, pp. 40-47. ISSN: 1072-5520. DOI= 10.1145/2029976.2029988.

Hornbæk, K. (2006) Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, Vol. 64, No. 2, pp. 79-102. ISSN: 1071-5819, DOI= 10.1016/j.ijhcs.2005.06.002.

Hornbæk, K. (2010) Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, Vol. 29, No. 1, pp. 97-111. ISSN: 0144-929X. DOI= 10.1080/01449290801939400.

Hornbæk, K. and 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* (CHI '05). ACM, New York, USA, pp. 391-400. ISBN: 1-58113-998-5, DOI= 10.1145/1054972.1055027.

Hornbæk, K. and Frøkjær, E. (2008a) Making use of business goals in usability evaluation: An experiment with novice evaluators. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (CHI '08). ACM, New York, USA, pp. 903-912. ISBN: 978-1-60558-011-1, DOI= 10.1145/1357054.1357197.

Hornbæk, K. and Frøkjær, E. (2008b) A study of the evaluator effect in usability testing. *Human-Computer Interaction*, Vol. 23, No. 3, pp. 251-277. ISSN: 0737-0024, DOI= 10.1080/07370020802278205.

Hornbæk, K. and Law, E.L-C. (2007) Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07). ACM, New York, USA, pp. 617-626. ISBN: 978-1-59593-593-9, DOI= 10.1145/1240624.1240722.

Hughes, M. (1999) Rigor in usability testing. *Technical Communication*, Vol. 46, No. 4, pp. 488-494. ISSN: 0049-3155, ProQuest document ID= 220995703.

Hwang, W. and Salvendy, G. (2010) Number of people required for usability evaluation: The 10±2 rule. *Communications of the ACM*, Vol. 53, No. 5, pp. 130-133. ISSN: 0001-0782, DOI= 10.1145/1735223.1735255.

Höysniemi, J., Hämäläinen, P. and Turkki, L. (2003) Using peer tutoring in evaluating the usability of a physically interactive computer game with children. *Interacting with Computers*, Vol. 15, No. 2, pp. 203-225. ISSN: 0953-5438, DOI= 10.1016/S0953-5438(03)00008-0.

Høegh, R.T. and Jensen, J.J. (2008) A case study of three software projects: Can software developers anticipate the usability problems in their software? *Behaviour & Information Technology*, Vol. 27, No. 4, pp. 307-312. ISSN: 0144-929X, DOI= 10.1080/01449290701766358.

Høegh, R.T., Nielsen, C.M., Overgaard, M., Pedersen, M.B. and Stage, J. (2006) The impact of usability reports and user test observations on developers' understanding of usability data: An exploratory study. *International Journal of Human-Computer Interaction*, Vol. 21, No. 2, pp. 173-196. ISSN: 1044-7318, DOI= 10.1207/s15327590ijhc2102_4.

Isbister, K. and Höök, K. (2009) On being supple: In search of rigor without rigidity in meeting new design and evaluation challenges for HCI practitioners. In *Proceedings of the 27th international conference on Human factors in computing systems* (CHI '09). ACM, New York, USA, pp. 2233-2242. ISBN: 978-1-60558-246-7, DOI= 10.1145/1518701.1519042.

ISO 9241-11 (1998) SFS-EN ISO 9241-11: Näyttöpäätteillä tehtävän toimistotyön ergonomiset vaatimukset. Osa 11: Käytettävyyden määrittely ja arviointi. *Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usa*bility. Suomen Standardoimisliitto SFS ry Helsinki, Finland, and European Committee for Standardization, Brussels, Belgium. 43 p.

ISO 9241-210 (2010) SFS-EN ISO 9241-210: Ihmisen ja järjestelmän vuorovaikutuksen ergonomia. Osa 210: Vuorovaikutteisten järjestelmien käyttäjäkeskeinen suunnittelu. *Ergonomics of human-system interaction. Part 210: Human-centred design for interactive systems*. Suomen Standardoimisliitto SFS ry Hel-

sinki, Finland, and European Committee for Standardization, Brussels, Belgium. 65 p.

ISO 13407 (1999) SFS-EN ISO 13407: Vuorovaikutteisten järjestelmien käyttäjäkeskeinen suunnitteluprosessi. *Human-centred design processes for interactive systems*. Suomen Standardoimisliitto SFS ry Helsinki, Finland, and European Committee for Standardization, Brussels, Belgium. 58 p.

ISO/IEC 9126-1 (2001) *Software engineering – Product quality – Part 1: Quality model*. International Organization for Standardization and International Electrotechnical Commission, Switzerland. 25 p.

ISO/IEC 25062 (2006) *Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Common Industry Format (CIF) for usability test reports*. International Organization for Standardization and International Electrotechnical Commission, Switzerland. 46 p.

ISO/TR 16982 (2002) *Ergonomics of human-system interaction – Usability methods supporting human-centred design*. International Organization for Standardization, Switzerland. 44 p.

Isomursu, M., Kuutti, K. and Väinämö, S. (2004) Experience clip: Method for user participation and evaluation of mobile concepts. In *Proceedings of the eighth conference on Participatory design: Artful integration: Interweaving media, materials and practices* (PDC 04), ACM, New York, USA, Vol. 1, pp. 83-92. ISBN: 1-58113-851-2, DOI= 10.1145/1011870.1011881.

Jacobsen, N.E., Hertzum, M. and John, B.E. (1998) The evaluator effect in usability studies: problem detection and severity judgments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 42, No. 19, pp. 1336-1340. ISSN: 1541-9312, DOI= 10.1177/154193129804201902.

Jeffries, R. (1994) Usability problem reports: Helping evaluators communicate effectively with developers. In J. Nielsen & R.L. Mack (Eds.) *Usability inspection methods*. John Wiley & Sons, New York, USA, pp. 273-294. ISBN: 0-471-01877-5.

Jeffries, R., Miller, J., Wharton, C. and Uyeda, K. (1991) User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the ACM CHI'91 Conference on Human Factors in Computing Systems*. ACM, New York, USA, pp. 119-124. ISBN: 0-89791-383-3, DOI= 10.1145/108844.108862.

Jick, T.D. (1979) Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, Vol. 24, No. 4, pp. 602-611. ISSN: 0001-8392.

Johannessen, G.H.J. and Hornbæk, K. (2014) Must evaluation methods be about usability? Devising and assessing the utility inspection method. *Behaviour & Information Technology*, Vol. 33, no. 2, pp. 194-205. ISSN: 0144-929X, DOI= 10.1080/0144929X.2012.751708.

John, B.E. and Marks, S.J. (1997) Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, Vol. 16, No. 4-5, pp. 188-202. ISSN: 0144-929X, DOI= 10.1080/014492997119789.

Julkaisufoorumi (2012) The Federation of Finnish Learned Societies. <http://www.tsv.fi/julkaisufoorumi/english.html?lang=en> n.d., retrieved 28. December 2012.

Juurmaa, K., Pitkänen, J., Riihiaho, S., Kantola, T. and Mäkelä, J. (2013) Visual walkthrough as a tool for utility assessment in a usability test. In Proceedings of the 27th International BCS Human Computer Interaction Conference (HCI 2013). Electronic Workshops in Computing (eWiC), pp. 1-6. ISSN: 1477-9358. Available on http://ewic.bcs.org/upload/pdf/ewic_hci13_short_paper11.pdf.

Jørgensen, A.H. (1990) Thinking-aloud in user interface design: a method promoting cognitive ergonomics. *Ergonomics*, Vol. 33, No. 4, pp. 501-507. ISSN: 0014-0139, DOI= 10.1080/00140139008927157.

Kaikkonen, A. (2009) *Internet on Mobiles: Evolution of Usability and User Experience*. TKK Dissertations 200, Helsinki University of Technology, Faculty of Information and Natural Sciences, Espoo, Finland. 123 + 100 p. ISBN: 978-952-248-189-4.

Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T. and Kankainen, A. (2005) Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability Studies*, Vol. 1, No. 1, pp. 4-16. ISSN: 1931-3357, available on http://www.upassoc.org/upa_publications/jus/2005_november/mobile.pdf.

Kallio, T. and Kekäläinen, A. (2004) Improving the effectiveness of mobile application design: User-pairs testing by non-professionals. In *Proceedings of Mobile Human-Computer Interaction - MobileHCI 2004*, S. Brewster & M. Dunlop (Eds.) *Lecture Notes in Computer Science*, Vol. 3160, Springer, Berlin, Germany, pp. 315-319. ISBN: 978-3-540-23086-1, DOI= 10.1007/978-3-540-28637-0_29.

Kanis, H. (2011) Estimating the number of usability problems. *Applied Ergonomics*, Vol. 42, No. 2, pp. 337-347. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2010.08.004.

Karat, C-M., Campbell, R. and 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* (CHI '92), P. Bauersfeld, J. Bennett & G. Lynch (Eds.). ACM, New York, USA, pp. 397-404. ISBN: 0-89791-513-5, DOI= 10.1145/142750.142873.

Karat, J. (1997) User-centered software evaluation methodologies. In M.G. Helander, T.K. Landauer & P.V. Prabhu (Eds.) *Handbook of human-computer interaction, 2nd ed.* Elsevier Science, Amsterdam, The Netherlands, pp. 689-704. ISBN: 978-0444886736.

Katerattanakul, P., Han, B. and Hong, S. (2003) Objective quality ranking of computing journals. *Communications of the ACM*, Vol. 46, No. 10, pp. 111-114. ISSN: 0001-0782, DOI= 10.1145/944217.944221.

Kato, T. (1986) What "question-asking protocols" can say about the user interface. *International Journal of Man-Machine Studies*, Vol. 25, No. 6, pp. 659-673. ISSN: 0020-7373. DOI= 10.1016/S0020-7373(86)80080-3.

Keenan, S. L., Hartson, H. R., Kafura, D. G., and Schulman, R. S. (1999) The usability problem taxonomy: A framework for classification and analysis. *Empirical Software Engineering*, Vol. 4, No. 1, pp. 71-104. ISSN: 1382-3256, DOI= 10.1023/A:1009855231530.

Keinonen, T., Nieminen, M.H.T., Riihiaho, S. and Säde, S. (1996). *Designing Usable Smart Products*. Helsinki University of Technology, Technical Report, TKO-C81.

Kennedy, S. (1989) Using video in the BNR usability lab. *SIGCHI Bulletin*, Vol. 21, No. 2, pp. 92-95. ISSN: 0736-6906, DOI= 10.1145/70609.70624.

van Kesteren, I.E.H., Bekker, M.M., Vermeeren, A.P.O.S. and Lloyd, P.A. (2003) Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In *Proceedings of the 2003 conference on Interaction design and children* (IDC '03), S. MacFarlane, T. Nicol, J. Read & L. Snape (Eds.). ACM, New York, USA, pp. 41-49. ISBN: 1-58113-732-X, DOI= 10.1145/953536.953544.

Kim, H., Kim, J., Lee, Y., Chae, M. and Choi, Y. (2002) An empirical study of the use contexts and usability problems in mobile Internet. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (HICSS'02), pp. 1767-1776. ISBN: 0-7695-1435-9, DOI= 10.1109/HICSS.2002.994090.

Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howard, S., Balbo, S. and Davies, J. (2005a) Evaluating the usability of a mobile guide: The influence of location, participants and resources. *Behaviour & Information Technology*, Vol. 24, No. 1, pp. 51-65. ISSN: 0144-929X, DOI= 10.1080/01449290512331319030.

Kjeldskov, J., Skov, M.B. and Stage, J. (2004) Instant data analysis: Conducting usability evaluations in a day. In *Proceedings of the third Nordic conference on Human-computer interaction* (NordiCHI '04). ACM, New York, USA, pp. 233-240. ISBN: 1-58113-857-1, DOI= 10.1145/1028014.1028050.

Kjeldskov, J., Skov, M.B. and Stage, J. (2005b) Does time heal? A longitudinal study of usability. In *Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future* (OZCHI '05). Computer-Human Interaction Special Interest Group (CHISIG) of Australia, Narrabundah, Australia, pp. 1-10. ISBN: 1-59593-222-4.

Kjeldskov, J. and Stage, J. (2004) New techniques for usability evaluation of mobile systems. *International Journal of Human-Computer Studies*, Vol. 60, No. 5–6, pp. 599-620. ISSN: 1071-5819, DOI= 10.1016/j.ijhcs.2003.11.001.

Koivunen, M-R., Nieminen, M.H.T. and Riihiaho, S. (1995) Launching the Usability Approach: Experience at Helsinki University of Technology. In *SIGCHI Bulletin*, Vol. 27, No. 2, pp. 54-60. ISSN: 0736-6906, DOI= 10.1145/202511.202526.

Krahmer, E. and Ummelen, N. (2004) Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*, Vol. 47, No. 2, pp. 105-117. ISSN: 0361-1434. DOI= 10.1109/TPC.2004.828205.

Lárusdóttir, M., Cajander, Å. and Gulliksen, J. (2013) Informal feedback rather than performance measurements – user-centred evaluation in Scrum projects. *Behaviour & Information Technology*. ISSN: 0144-929X, DOI= 10.1080/0144929X.2013.857430, Published online: December 6th 2013.

Lavery, D., Cockton, G. and Atkinson, M. P. (1997) Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, Vol. 16, No. 4-5, pp. 246-266. ISSN: 0144929X, DOI= 10.1080/014492997119824.

Lazar, J., Feng, J.H. and Hochheiser, H. (2010) *Research methods in human-computer interaction*. John Wiley & Sons, Glasgow, UK, 426 p. ISBN: 978-0-470-72337-1.

Lepistö, A. and Ovaska, S. (2004) Usability evaluation involving participants with cognitive disabilities. In *Proceedings of the third Nordic conference on Human-computer interaction* (NordiCHI '04). ACM, New York, USA, pp. 305-308. ISBN: 1-58113-857-1, DOI= 10.1145/1028014.1028061.

Levy, Y. and Ellis, T.J. (2006) A systems approach to conduct an effective literature review in support of information systems research. *Informing Science Journal*, Vol. 9, pp. 181-212. ISSN: 1547-9684.

Lewis, C.H. and Mack, R. (1982) Learning to use a text processing system: Evidence from "thinking aloud" protocols. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems* (CHI '82). ACM, New York, USA, pp. 387-392. DOI= 10.1145/800049.801817.

Lewis, C.H., Polson, P.G., Wharton, C. and Rieman, J. (1990) Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Pro-*

ceedings of the ACM CHI'90 Conference on Human Factors in Computing Systems (CHI '90), J.C. Chew & J. Whiteside (Eds.). ACM, New York, USA, pp. 235-242. ISBN: 0-201-50932-6, DOI= 10.1145/97243.97279.

Lewis, C.H. and Wharton, C. (1997) Cognitive walkthroughs. In M. Helander, T.K. Landauer & P. Prabhu (Eds.) *Handbook of human-computer interaction, 2nd ed.* Elsevier Science, Amsterdam, The Netherlands, pp. 717-732. ISBN: 978-0444886736.

Lewis, J.R. (1994) Sample sizes for usability studies: Additional considerations. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 36, no. 2, pp. 368-378. ISSN: 0018-7208, DOI= 10.1177/001872089403600215.

Lewis, J.R. (2001a) Introduction: Current issues in usability evaluation. *International Journal of Human-Computer Studies*, Vol. 13, No. 4, pp. 343-349. ISSN: 1071-5819, DOI= 10.1207/S15327590IJHC1304_01.

Lewis, J.R. (2001b) Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Studies*, Vol. 13, No. 4, pp. 445-479. ISSN: 1071-5819, DOI= 10.1207/S15327590IJHC1304_06.

Lewis, J.R. (2014) Usability: Lessons learned ... and yet to be learned. *International Journal of Human-Computer Interaction*, Vol. 30, No. 9, pp. 663-684. ISSN: 1044-7318, DOI= 10.1080/10447318.2014.930311.

Lewis, J.R. and Sauro, J. (2009) The Factor Structure of the System Usability Scale. In *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009* (HCD 09), M. Kurosu (Ed.). Springer-Verlag, Berlin, Germany, pp. 94-103. ISBN: 978-3-642-02805-2, DOI= 10.1007/978-3-642-02806-9_12.

Lindgaard, G. and Chattratichart, J. (2007) Usability testing: What have we overlooked? In *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI '07). ACM, New York, USA, pp. 1415-1424. ISBN: 978-1-59593-593-9, DOI= 10.1145/1240624.1240839.

Lowry, P.B., Romans, D. and Curtis, A. (2004) Global journal prestige and supporting disciplines: A scientometric study of information systems journals. *Journal of the Association for Information Systems*, Vol. 5, No. 2, pp. 29-77. ISSN: 1536-9323.

Mack, R.L., Lewis, C.H. and Carroll, J.M. (1983) Learning to use word processors: Problems and prospects. *ACM Transactions on Information Systems*, Vol. 1, No. 3, pp. 254-271. ISSN: 1046-8188, DOI= 10.1145/357436.357440.

Mackay, W.E. and Fayard, A-L. (1997) HCI, natural science and design: A framework for triangulation across disciplines. In *Proceedings of the 2nd conference on Designing interactive systems: Processes, practices, methods, and techniques* (DIS '97), S. Coles (Ed.). ACM, New York, USA, pp. 223-234. ISBN: 0-89791-863-0, DOI= 10.1145/263552.263612.

Maguire, M. (2001) Context of Use within usability activities. *International Journal of Human-Computer Studies*, Vol. 55, No. 4, pp. 453-483. ISSN: 1071-5819, DOI= 10.1006/ijhc.2001.0486.

Maier, N.R.F. (1931) Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, Vol. 12, No. 2, pp. 181-194. ISSN: 0735-7036, DOI= 10.1037/h0071361.

O'Malley, C.E., Draper, S.W. and Riley, M.S. (1984) Constructive interaction: A method for studying human-computer-human interaction. In B. Shackel (Ed.) *Human-computer interaction - INTERACT'84*. Elsevier Science Publishers, Vol. 84, pp. 269-274. ISBN: 0-444-87773-8.

Markopoulos, P. and Bekker, M. (2003) On the assessment of usability testing methods for children. *Interacting with computers*, Vol. 15, No. 2, pp. 227-243. ISSN: 0953-5438, DOI= 10.1016/S0953-5438(03)00009-2.

McDonald, S., Edwards, H. M. and Zhao, T. (2012) Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication*, Vol. 55, No. 1, pp. 2-19. ISSN: 0361-1434, DOI= 10.1109/TPC.2011.2182569.

McDonald, S., McGarry, K. and Willis, L.M. (2013a) Thinking-aloud about web navigation: The relationship between think-aloud instructions, task difficulty and performance. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting - 2013*, Vol. 57, No. 1, pp. 2037-2041. ISBN: 978-0-945289-43-2, DOI= 10.1177/1541931213571455.

McDonald, S., Monahan, K. and Cockton, G. (2006) Modified contextual design as a field evaluation method. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (NordiCHI '06), A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh & D. Svanaes (Eds.). ACM, New York, USA, pp. 437-440. ISBN: 1-59593-325-5, DOI= 10.1145/1182475.1182531.

McDonald, S. and Petrie, H. (2013) The effect of global instructions on think-aloud testing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). ACM, New York, USA, pp. 2941-2944. ISBN: 978-1-4503-1899-0, DOI= 10.1145/2470654.2481407.

McDonald, S., Zhao, T. and Edwards, H.M. (2013b) Dual verbal elicitation: The complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human-Computer Interaction*, Vol. 29, No. 10, pp. 647-660, ISSN: 1044-7318, DOI= 10.1080/10447318.2012.758529.

Melkus, L.A. (1985) The benefits of laboratory testing for usability. In *Proceedings of the twenty-first annual conference on Computer personnel research (SIGCPR '85)*, J.C. Wetherbe (Ed.). ACM, New York, USA, pp. 91-96. ISBN: 0-89791-156-3, DOI= 10.1145/16687.16698.

Molich, R. and Dumas, J.S. (2008) Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, Vol. 27, No. 3, pp. 263-281. ISSN: 0144929X, DOI= 10.1080/01449290600959062.

Molich, R., Ede, M.R., Kaasgaard, K. and Karyukin, B. (2004) Comparative usability evaluation. *Behaviour & Information Technology*, Vol. 23, No. 1, pp. 65-74. ISSN: 0144929X, DOI= 10.1080/0144929032000173951.

Monahan, K., Lähteenmäki, M., McDonald, S. and Cockton, G. (2008) An investigation into the use of field methods in the design and evaluation of interactive systems. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1* (BCS-HCI '08). British Computer Society, Swinton, UK, pp. 99-108. ISBN: 978-1-906124-04-5, ACM id= 1531528.

Monk, A. (2004) The Product as a Fixed-Effect Fallacy. *Human-Computer Interaction*, Vol. 19, No. 4, pp. 371-375. ISSN: 0737-0024, DOI= 10.1207/s15327051hci1904_6.

Moshagen, M., Musch, J. and Göritz, A.S. (2009) A blessing, not a curse: Experimental evidence for beneficial effects of visual aesthetics on performance. *Ergonomics*, Vol. 52, No. 10, pp. 1311-1320. ISSN: 0014-0139, DOI= 10.1080/00140130903061717.

Mugge, R. and Schoormans, J.P.L. (2012) Product design and apparent usability. The influence of novelty in product appearance. *Applied Ergonomics*, Vol. 43, No. 6, pp. 1081-1088. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2012.03.009.

Nielsen, C.M., Overgaard, M., Pedersen, M.B., Stage, J. and 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* (NordiCHI '06), A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh & D. Svanaes (Eds.). ACM, New York, USA, pp. 272-280. ISBN: 1-59593-325-5, DOI= 10.1145/1182475.1182504.

Nielsen, J. (1993) *Usability Engineering*. Academic Press, Boston, USA. 358 p. ISBN: 0-12-518405-0.

Nielsen, J. (1994a) Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, Vol. 41, No. 3, pp. 385-397. ISSN: 1071-5819, DOI= 10.1006/ijhc.1994.1065.

Nielsen, J. (1994b) Heuristic evaluation. In J. Nielsen & R.L. Mack (Eds.) *Usability inspection methods*. John Wiley & Sons, New York, USA, pp. 25-62. ISBN: 0-471-01877-5.

Nielsen, J. (1995) Scenarios in discount usability engineering. In J.M. Carroll (Ed.) *Scenario-based design: Envisioning work and technology in system development*. John Wiley & Sons, New York, USA, pp. 59-83. ISBN: 0-471-07659-7.

Nielsen, J. and Levy, J. (1994) Measuring usability: Preference *vs.* performance. *Communications of the ACM*, Vol. 37, no. 4, pp. 66-75. ISSN: 0001-0782, DOI= 10.1145/175276.175282.

Nielsen, J. and Molich, R. (1990) Heuristic evaluation of user interfaces. In *Proceedings of the ACM CHI'90 Conference on Human Factors in Computing Systems* (CHI '90), J.C. Chew & J. Whiteside (Eds.). ACM, New York, USA, pp. 249-256. ISBN: 0-201-50932-6, DOI= 10.1145/97243.97281.

Nieminen, M.H.T. and Koivunen, M-R. (1995) Visual Walkthrough. In G. Allen, J. Wilkinson & P.C. Wright (Eds.) *HCI'95, People and Computers, Adjunct Proceedings*. The School of Computing & Mathematics, University of Huddersfield, UK, pp. 86-89.

Nieminen, M.P. (1996) *Designing user interface concepts for multimedia services*. Master's thesis, Helsinki University of Technology, Espoo, Finland. 100+8 p.

Nisbett, R.E. and Wilson, T.DC. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, Vol. 84, No. 3, pp. 231-259. ISSN: 0033-295X, DOI= 10.1037/0033-295X.84.3.231.

Nørgaard, M. and Hornbæk, K. (2006) What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the 6th conference on Designing Interactive systems* (DIS '06). ACM, New York, USA, pp. 209-218. ISBN: 1-59593-367-0, DOI= 10.1145/1142405.1142439.

Nørgaard, M. and Hornbæk, K. (2009) Exploring the value of usability feedback formats. *International Journal of Human-Computer Interaction*, Vol. 25, No. 1, pp. 49-74. ISSN: 1044-7318, DOI= 10.1080/10447310802546708.

Ohnemus, K.R. and Biers, D.W. (1993) Retrospective versus concurrent thinking-out-loud in usability testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 37, No. 17, pp. 1127-1131. ISSN: 1541-9312, DOI= 10.1177/154193129303701701.

Olmsted-Hawala, E.L., Murphy, E.D., Hawala, S. and Ashenfelter, K.T. (2010a) Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10). ACM, New York, USA, pp. 2381-2390. ISBN: 978-1-60558-929-9, DOI= 10.1145/1753326.1753685.

Olmsted-Hawala, E.L., Murphy, E.D., Hawala, S. and Ashenfelter, K.T. (2010b) Think-aloud protocols: Analyzing three different think-aloud protocols with counts of verbalized frustrations in a usability study of an information-rich web site. In *Proceedings of the IEEE International Professional Communication Conference* (IEEE IPCC'10), Enschede, Netherlands. IEEE, New York, USA, pp. 60–66. ISBN: 978-1-4244-8145-3, DOI= 10.1109/IPCC.2010.5529815.

Olson, G.M. and Moran, T.P. (1998) Commentary on "Damaged Merchandise?" *Human-Computer Interaction*, Vol. 13, No. 3, (Experimental Comparisons of Usability Evaluation Methods: A Special Issue of Human-Computer Interaction, ISBN: 0-8058-9813-1), pp. 263-323. ISSN: 0737-0024, DOI= 10.1207/s15327051hci1303_3.

Orne, M.T. (1962) On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, Vol. 17, No. 11, pp. 776-783. ISSN: 0003-066X, Accession Number= 00000487-196211000-00005.

Pike, M.F., Maior, H.A., Porcheron, M., Sharples, S.C. and Wilson, M.L. (2014) Measuring the effect of think aloud protocols on workload using fNIRS. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (CHI '14). ACM, New York, USA, pp. 3807-3816. ISBN: 978-1-4503-2473-1, DOI= 10.1145/2556288.2556974.

Polson, P.G., Lewis, C.H, Rieman, J. and Wharton, C. (1992) Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, Vol. 36, No. 5, pp. 741-773. ISSN: 0020-7373, DOI= 10.1016/0020-7373(92)90039-N.

Potosnak, K. (1988) Human factors – Recipe for a usability test. *IEEE Software*, Vol. 5, No.6, pp. 83-84. ISSN: 0740-7459, DOI= 10.1109/52.10008.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. (1994) *Human-computer interaction*. Addison-Wesley, Wokingham, UK, 773 p. ISBN: 0-201-62769-8.

Quinn, J.M. and Tran, T.Q. (2010) Attractive phones don't have to work better: Independent effects of attractiveness, effectiveness, and efficiency on perceived usability. In *Proceedings of the 28th international conference on Human factors in computing systems* (CHI '10). ACM, New York, USA, pp. 353-362. ISBN: 978-1-60558-929-9, DOI= 10.1145/1753326.1753380.

Raita, E. and Oulasvirta, A. (2011) Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers*, Vol. 23, No. 4, pp. 363-371. ISSN: 0953-5438, DOI= 10.1016/j.intcom.2011.04.002.

Redish, J. (2010) Technical Communication and Usability: Intertwined Strands and Mutual Influences. *IEEE Transactions on Professional Communication*, vol.53, no.3, pp.191-201. ISSN: 0361-1434, DOI= 10.1109/TPC.2010.2052861.

Redish, J., Bias, R.G., Bailey, R., Molich, R., Dumas, J. and Spool, J.M. (2002) Usability in practice: Formative usability evaluations - Evolution and revolution. *In CHI '02 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '02). ACM, New York, NY, USA, pp. 885-890. ISBN: 1-58113-454-1, DOI= 10.1145/506443.506647.

Reeves, B. and Nass, C. (1996) *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Center for the Study of Language and Information, Chicago, US; Cambridge University Press, New York, US, 305 p. ISBN: 1575860538.

Rhenius, D. and Deffner, G. (1990) Evaluation of concurrent thinking aloud using eye-tracking data. In *Proceedings of the Human Factors and Ergonomics Society*

Annual Meeting. Vol. 34, No. 17, pp. 1265-1269. ISSN: 1071-1813, DOI= 10.1177/154193129003401719.

Riihiaho, S. (2000) *Experiences with usability evaluation methods*. Licentiate's thesis. Helsinki University of Technology. Available on http://www.soberit.hut.fi/~sri/Riihiaho_thesis.pdf.

Riihiaho, S. (2002) The pluralistic usability walk-through method. *Ergonomics in Design: The Quarterly of Human Factors Applications*, Vol. 10, No. 3, pp. 23-27. ISSN: 1064-8046, DOI= 10.1177/106480460201000306.

Riihiaho, S. (2009) User testing when test tasks are not appropriate. In *European Conference on Cognitive Ergonomics: Designing beyond the Product – Understanding Activity and User Experience in Ubiquitous Environments (ECCE '09)*, L. Norros, H. Koskinen, L. Salo & P. Savioja (Eds.). VTT Technical Research Centre of Finland, Espoo, Finland, Article 21, pp. 228-235. ISBN: 978-951-38-6340-1.

Rohn, J.A., Spool, J., Ektare, M., Koyani, S., Muller, M. and Redish, J. (2002) Usability in practice: Alternatives to formative evaluations – Evolution and revolution. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, USA, pp. 891-897. ISBN: 1-58113-454-1, DOI= 10.1145/506443.506648.

Rosenbaum, S., and Kantner, L. (2007) Field usability testing: Method, not compromise. In *Proceedings of the IEEE International Professional Communication Conference, 2007* (IPCC 2007). IEEE, New York, USA, pp. 1-7. ISBN: 978-1-4244-1242-6, DOI= 10.1109/IPCC.2007.4464060.

Rosenbaum, S., Rohn, J.A. and Humburg, J. (2000) A toolkit for strategic usability: Results from workshops, panels, and surveys. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '00). ACM, New York, USA, pp. 337-344. ISBN: 1-58113-216-6, DOI= 10.1145/332040.332454.

Rubin, J. (1994) *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests*, John Wiley & Sons, New York, USA, 330 p. ISBN: 0471594032.

Rubin, J. and Chisnell, D. (2008) *Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests, 2nd ed.*, Wiley Publishing, Indianapolis, USA, 348 p. ISBN: 978-0-470-18548-3.

Rudd, J., Stern, K. and Isensee, S. (1996) Low *vs.* high-fidelity prototyping debate. *interactions*, Vol. 3, No. 1, pp. 76-85. ISSN: 1072-5520, DOI= 10.1145/223500.223514.

Russo, J.E., Johnson, E.J. and Stephens, D.L. (1989) The validity of verbal protocols. *Memory & Cognition*, Vol. 17, No. 6, pp. 759-769. ISSN: 0090-502X, DOI= 10.3758/BF03202637.

Salzman, M.C. and Rivers, S.D. (1994) Smoke and mirrors: Setting the stage for a successful usability test. *Behaviour & Information Technology*, Vol. 13, No. 1-2, pp. 9-16. ISSN: 0144929X, DOI= 10.1080/01449299408914578.

Sauer, J., Seibel, K. and Rüttinger, B. (2010) The influence of user expertise and prototype fidelity in usability tests. *Applied Ergonomics*, Vol. 41, No. 1, pp. 130-140. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2009.06.003.

Sauer, J. and Sonderegger, A. (2009) The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion. *Applied Ergonomics*, Vol. 40, No. 4, pp. 670-677. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2008.06.006.

Sauer, J. and Sonderegger, A. (2011a) Methodological issues in product evaluation: The influence of testing environment and task scenario. *Applied Ergonomics*,

Vol. 42, No. 3, pp. 487-494. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2010.09.005.

Sauer, J. and Sonderegger, A. (2011b) The influence of product aesthetics and user state in usability testing. *Behaviour & Information Technology*, Vol. 30, No. 6, pp. 787-796. ISSN: 0144-929X, DOI= 10.1080/0144929X.2010.503352.

Sauro, J. (2004) Premium usability: Getting the discount without paying the price. *interactions,* Vol. 11, No. 4, pp. 30-37. ISSN: 1072-5520, DOI= 10.1145/1005261.1005276.

Sauro, J. (2006) The user is in the numbers. *interactions*, Vol. 13, No. 6, pp. 22-25. ISSN: 1072-5520, DOI= 10.1145/1167948.1167971.

Sauro, J. and Kindlund, E. (2005) A method to standardize usability metrics into a single score. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI '05). ACM, New York, USA, pp. 401-409. ISBN: 1-58113-998-5, DOI= 10.1145/1054972.1055028.

Sauro, J. and Lewis, J.R. (2009) Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of the 27th international conference on Human factors in computing systems* (CHI '09). ACM, New York, USA, pp. 1609-1618. ISBN: 978-1-60558-246-7, DOI= 10.1145/1518701.1518947.

Sauro, J. and Lewis, J.R. (2011) When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, USA, pp. 2215-2224. ISBN: 978-1-4503-0228-9, DOI= 10.1145/1978942.1979266.

Savioja, P., Norros, L. and Salo, L. (2008) Evaluation of Systems Usability. In *Proceedings of the European Conference on Cognitive Ergonomics* (Madeira, Portugal, September 16-19, 2008). ECCE'08. Article no 26, 8 p. ISBN: 978-1-60558-399-0, DOI= 10.1145/1473018.1473051.

Sawyer, P., Flanders, A. and Wixon, D. (1996) Making a difference – The impact of inspections. In *Proceedings of the ACM CHI'96 Conference on Human Factors in Computing Systems* (CHI '96) pp. 376-382. ISBN: 0-89791-777-4, DOI= 10.1145/238386.238579.

Schrier, J.R. (1992) Reducing stress associated with participating in a usability test. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 36, No. 16, pp. 1210-1214. ISSN: 1541-9312, DOI= 10.1177/154193129203601606.

Schulte-Mecklenbeck, M. and Huber, O. (2003) Information search in the laboratory and on the Web: With or without an experimenter. *Behavior Research Methods, Instruments, & Computers*, Vol. 35, No. 2, pp. 227-235. ISSN: 0743-3808, DOI= 10.3758/BF03202545.

Sears, A. (1997) Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, Vol. 9, No. 3, pp. 213-234. ISSN: 1044-7318, DOI= 10.1207/s15327590ijhc0903_2.

Seffah, A., Donyaee, M., Kline, R.B. and Padda, H. K. (2006) Usability measurement and metrics: A consolidated model. *Software Quality Journal*, Vol. 14, No. 2, pp. 159-178. ISSN: 0963-9314, DOI= 10.1007/s11219-006-7600-8.

Sengers, P. and Gaver, B. (2006) Staying open to interpretation: Engaging multiple meanings in design and evaluation. In *Proceedings of the 6th conference on Designing Interactive systems* (DIS '06). ACM, New York, USA, pp. 99-108. ISBN: 1-59593-367-0, DOI= 10.1145/1142405.1142422.

Shi, Q. (2008) A field study of the relationship and communication between Chinese evaluators and users in thinking aloud usability tests. In *Proceedings of the 5th*

Nordic conference on Human-computer interaction: Building bridges (NordiCHI '08). ACM, New York, USA, pp. 344-352. ISBN: 978-1-59593-704-9, DOI= 10.1145/1463160.1463198.

Shneiderman, B. (1998) *Designing the user interface: Strategies for effective human-computer interaction,* 3rd ed. Addison-Wesley, Reading, USA, 639 p. ISBN: 0-201-69497-2.

Shrimpton-Smith, T., Zaman, B. and Geerts, D. (2008) Coupling the Users: The Benefits of Paired User Testing for iDTV. *International Journal of Human-Computer Interaction*, Vol. 24, No. 2, pp. 197-213. ISSN: 1044-7318, DOI= 10.1080/10447310701821558.

Silverman, D. (2000) *Doing qualitative research: A practical handbook*. SAGE Publications, London, UK, 316 p. ISBN: 0761958231.

Skov, M.B. and Stage, J. (2012) Training software developers and designers to conduct usability evaluations. *Behaviour & Information Technology*, Vol. 31, No. 4, pp. 425-435. ISSN: 0144929X, DOI= 10.1080/01449290903398208.

Sonderegger, A., and Sauer, J. (2009) The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. *Ergonomics*, Vol. 52, No. 11, pp. 1350-1361. ISSN: 0014-0139, DOI= 10.1080/00140130903067797.

Sonderegger, A. and Sauer, J. (2010) The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied Ergonomics*, Vol. 41, No. 3, pp. 403-410. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2009.09.002.

Sonderegger, A. and Sauer, J. (2013) The influence of socio-cultural background and product value in usability testing. *Applied Ergonomics*, Vol. 44, No. 3, pp. 341-349. ISSN: 0003-6870, DOI= 10.1016/j.apergo.2012.09.004.

Sonderegger, A., Zbinden, G., Uebelbacher, A. and Sauer, J. (2012) The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. *Ergonomics*, Vol. 55, No. 7, pp. 713-730. ISSN: 0014-0139, DOI= 10.1080/00140139.2012.672658.

Spencer, R.H. (1985) *Computer usability testing & evaluation*. Prentice-Hall, Upper Saddle River, USA, 224 p. ISBN: 0-13-164088-7.

Stinessen, L. (1985) The influence of verbalization on problem-solving. *Scandinavian Journal of Psychology*, Vol. 26, No. 1, pp. 342–347. ISSN: 1467-9450, DOI= 10.1111/j.1467-9450.1985.tb01173.x.

Sullivan, P. (1989) Beyond a narrow conception of usability testing. *IEEE Transactions on Professional Communication*, Vol. 32, No. 4, pp. 256-264. ISSN: 0361-1434, DOI= 10.1109/47.44537.

Säde, S., Nieminen, M.H.T. and Riihiaho, S. (1998) Testing usability with 3D paper prototypes - Case Halton System. *Applied Ergonomics* Vol. 29, No. 1, pp. 67-73. ISSN: 0003-6870, DOI= 10.1016/S0003-6870(97)00027-6.

Theofanos, M. and Quesenbery, W. (2005) Towards the design of effective formative test reports. *Journal of Usability Studies*, Vol. 1, No. 1, pp. 27-45. ISSN: 1931-3357, DOI= 10.1.1.102.6218.

Thomas, P. and Macredie, R.D. (2002) Introduction to the new usability. *ACM Transactions on Computer-Human Interaction,* Vol. 9, No. 2, pp. 69-73. ISSN: 1073-0516, DOI= 10.1145/513665.513666.

Tohidi, M., Buxton, W., Baecker, R. and Sellen, A. (2006a) Getting the right design and the design right. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (CHI '06), R. Grinter, T. Rodden, P. Aoki, E. Cutrell,

R. Jeffries & G.M. Olson (Eds.). ACM, New York, USA, pp. 1243-1252. ISBN: 1-59593-372-7, DOI= 10.1145/1124772.1124960.

Tohidi, M., Buxton, W., Baecker, R. and Sellen, A. (2006b) User sketches: A quick, inexpensive, and effective way to elicit more reflective user feedback. In *Proceedings of the 4th Nordic conference on Human-computer interaction: Changing roles* (NordiCHI '06), A. Mørch, K. Morgan, T. Bratteteig, G. Ghosh & D. Svanaes (Eds.). ACM, New York, USA, pp. 105-114. ISBN: 1-59593-325-5, DOI= 10.1145/1182475.1182487.

Tractinsky, N., Katz, A.S. and Ikar, D. (2000) What is beautiful is usable. *Interacting with Computers*, Vol. 13, No. 2, pp. 127-145. ISSN: 0953-5438, DOI= 10.1016/S0953-5438(00)00031-X.

Trudel, C-I. and Payne, S.J. (1995) Reflection and goal management in exploratory learning. *International Journal of Human-Computer Studies*, Vol. 42, No. 3, pp 307-339. ISSN: 1071-5819, DOI= 10.1006/ijhc.1995.1015.

Trudel, C-I. and Payne, S.J. (1996) Self-monitoring during exploration of an interactive device. *International Journal of Human-Computer Studies*, Vol. 45, No. 6, pp. 723-747. ISSN: 1071-5819, DOI= 10.1006/ijhc.1996.0076.

Tullis, T. and Albert, B. (2008) *Measuring the user experience: Collecting, analyzing, and presenting usability metrics*. Elsevier/Morgan Kaufmann, Amsterdam, The Netherlands, 317 p. ISBN: 978-0-12-373558-4.

Uldall-Espersen, T., Frøkjær, E. and Hornbæk, K. (2008) Tracing impact in a usability improvement process. *Interacting with Computers*, Vol. 20, No. 1, pp. 48-63. ISSN: 0953-5438, DOI= 10.1016/j.intcom.2007.08.001.

Valero, P. and Monk, A. (1998) Positioning HCI: Journals, descriptors and parent disciplines. *Behaviour & Information Technology*, Vol. 17, No. 1, pp. 3-9. ISSN: 0144929X, DOI= 10.1080/014492998119625.

Venturi, G., Troost, J. and Jokela, T. (2006) People, organizations, and processes: An inquiry into the adoption of user-centered design in industry. *International Journal of Human-Computer Interaction*, Vol. 21, No. 2, pp. 219-238. ISSN: 1044-7318, DOI= 10.1207/s15327590ijhc2102_6.

Vermeeren, A.P.O.S, van Kesteren, I.E.H. and Bekker, M.M. (2003) Managing the evaluator effect in user testing. In *Proceedings of IFIP Conference on Human-Computer Interaction* (INTERACT'03), G.W.M. Rauterberg, M. Menozzi & J. Wesson (Eds.). IOS Press, Amsterdam, The Netherlands, pp. 647-654. ISBN: 978-1-58603-363-7.

Virzi, R.A. (1992) Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 34, no. 4, pp. 457-468. ISSN: 0018-7208, DOI= 10.1177/001872089203400407.

Virzi, R.A. (1997) Usability inspection methods. In M. Helander, T.K. Landauer & P. Prabhu (Eds.) *Handbook of human-computer interaction, 2nd ed.* Elsevier Science, Amsterdam, The Netherlands, pp. 705-715. ISBN: 978-0444886736.

Virzi, R.A., Sokolov, J.L. and Karis, D. (1996) Usability problem identification using both low- and high-fidelity prototypes. In *Proceedings of the ACM CHI'96 Conference on Human Factors in Computing Systems* (CHI '96), M.J. Tauber (Ed.). ACM, New York, USA, pp. 236-243. ISBN: 0-89791-777-4, DOI= 10.1145/238386.238516.

Watson, J. B. (1920). Is thinking merely the action of language mechanisms? *British Journal of Psychology*, Vol. 11, pp. 87-104. ISSN: 00071269. Reprinted in 2009, Vol. 100, No. 1a, pp. 169-180. DOI= 10.1348/000712608X336095.

Werner, H. and Kaplan, B. (1963) *Symbol formation: an organismic-developmental approach to language and the expression of thought*. John Wiley & Sons, New York, USA, 530 p. ISBN: 9780471933724.

Wharton, C., Rieman, J., Lewis, C.H. and Polson, P. (1994) The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R.L. Mack (Eds.) *Usability inspection methods*. John Wiley & Sons, New York, USA, pp. 105-140. ISBN: 0-471-01877-5.

Whiteside, J., Bennett, J. and Holtzblatt, K. (1988) Usability engineering: Our experience and evolution. In M. Helander (Ed.) *Handbook of human-computer interaction*. Elsevier, Amsterdam, The Netherlands, pp. 791-817. ISBN: 978-0444705365.

Wichansky, A.M. (2000) Usability testing in 2000 and beyond. *Ergonomics*, Vol. 43, No. 7, pp. 998-1006. ISSN: 0014-0139, DOI= 10.1080/001401300409170.

Wildman, D. (1995) Getting the most from paired-user testing. *interactions*, Vol. 2, No. 3, pp. 21-27. ISSN: 1072-5520, DOI= 10.1145/208666.208675.

Wilson, C.E. (1998) Usability techniques: Pros and cons of co-participation in usability studies. *Usability Interface*, Vol. 4, No. 4 [Online]. Available on http://www.stcsig.org/usability/newsletter/9804-coparticipation.html.

Wilson, C.E. (2006) Triangulation: The explicit use of multiple methods, measures, and approaches for determining core issues in product development. *interactions*, Vol. 13, No. 6, pp. 46-47&63. ISSN: 1072-5520, DOI= 10.1145/1167948.1167980.

Wixon, D. (2003) Evaluating usability methods: Why the current literature fails the practitioner. *interactions*, Vol. 10, No. 4, pp. 28-34. ISSN: 1072-5520. DOI= 10.1145/838830.838870.

Wixon, D. and Wilson, C.E. (1997) The usability engineering framework for product design and evaluation. In M. Helander, T.K. Landauer & P. Prabhu (Eds.) *Handbook of human-computer interaction, 2nd ed.* Elsevier Science, Amsterdam, The Netherlands, pp. 653-688. ISBN: 978-0444886736.

Woolrych, A., Hornbæk, K., Frøkjær, E., and Cockton, G. (2011). Ingredients and meals rather than recipes: A proposal for research that does not treat usability evaluation methods as indivisible wholes. *International Journal of Human-Computer Interaction*, Vol. 27, No. 10, pp. 940-970. ISSN: 1044-7318, DOI= 10.1080/10447318.2011.555314.

Wright, P.C. and Monk, A.F. (1991a) The use of think-aloud evaluation methods in design. *ACM SIGCHI Bulletin*, Vol. 23, No. 1, pp. 55-57. ISSN: 0736-6906, DOI= 10.1145/122672.122685.

Wright, P.C. and Monk, A.F. (1991b) A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies, Vol.* 35, No. 6, pp. 891-912. ISSN: 0020-7373, DOI= 10.1016/S0020-7373(05)80167-1.

Wright, R.B. and Converse, S.A. (1992) Method bias and concurrent verbal protocol in software usability testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 36, No. 16, pp. 1220-1224. ISSN: 1541-9312, DOI= 10.1177/154193129203601608.

Yeo, A. (2000) Are usability assessment techniques reliable in non-Western cultures? *The Electronic Journal of Information Systems in Developing Countries*, Vol. 3, No. 1, pp. 1-21. ISSN: 1681-4835, available on www.ejisdc.org.

Youmans, R.J., Gonzalez, C.A., Figueroa, I.J. and Bellows, B. (2013) The effects of task-set switching on concurrent verbal protocol. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting - 2013*. Vol. 57, No. 1, pp. 1668-1672. ISBN: 978-0-945289-43-2, DOI= 10.1177/1541931213571370.

Zhao, T. and McDonald, S. (2010) Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (NordiCHI '10). ACM, New York, USA, pp. 581-590. ISBN: 978-1-60558-934-3, DOI= 10.1145/1868914.1868979.

Zhao, T., McDonald, S. and Edwards, H.M. (2014) The impact of two different think-aloud instructions in a usability test: A case of just following orders? *Behaviour & Information Technology*, Vol. 33, No. 2, pp. 162-182. ISSN: 0144929X, DOI= 10.1080/0144929X.2012.708786.

Åborg, C., Sandblad, B., Gulliksen, J. and Lif, M. (2003) Integrating work environment considerations into usability evaluation methods – the ADA approach. *Interacting with Computers*, Vol. 15, No. 3, pp. 453-471. ISSN: 0953-5438, DOI= 10.1016/S0953-5438(02)00060-7.

Appendix A: Data of empirical studies

The following table summarises the usability evaluation studies that the author has participated in Aalto University in 1993-2014. In the table, the acronym SW stands for software systems, and SP for smart products; Pro stands for systems developed for professional use, and Re for systems for recreational use or a wide set of users; RP stands for studies made by usability researchers in our research projects, and CA for studies made by students as course assignments. The number of users includes also the users in the pilot tests.

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
1	Elevator control system	1993	SP	Pro	CA	Pre-test interview, visual walkthrough, usability test, post-test interview, heuristic evaluation	X	4
2	Patient monitoring system text editor	1993	SW	Pro	CA	Usability test, post-test questionnaire, post-test interview	X	3
3	Reverse vending machine	1993	SP	Re	CA	Usability test, post-test interview, observation	X	11
4	Simulator for patient ventilation	1993	SW	Pro	CA	Pre-test interview, usability test, post-test interview	X	6
5	System for configuring an anaesthesia record keeper	1993	SW	Pro	CA	Usability test, post-test interview	X	4
6	Television (everyday use)	1993	SP	Re	CA	Pre-test interview, usability test, post-test interview	X	8
7	Television (taking into use)	1993	SP	Re	CA	Usability test, post-test questionnaire, post-test interview, heuristic evaluation	X	13
8	X-ray fluorescence analyser	1993	SP	Pro	CA	Usability test, post-test questionnaire, post-test interview	X	5
9	Anaesthesia record keeper	1994	SP	Pro	CA	Pre-test questionnaire, usability test, post-test questionnaire	X	6
10	CASE tool	1994	SW	Pro	CA	Pre-test questionnaire, usability test, post-test interview	X	6
11	Document management system	1994	SW	Pro	CA	Pre-test questionnaire, usability test, post-test questionnaire, post-test interview, heuristic evaluation	X	6
12	Gaming slot machine	1994	SP	Re	CA	Usability test, paired-user testing, post-test interview	X	8

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
13	Information management system	1994	SW	Pro	CA	Interview, pre-test interview, usability test, post-test interview, heuristic evaluation	X	4
14	Network navigator application	1994	SW	Pro	CA	Usability test, post-test questionnaire, post-test interview, heuristic evaluation	X	4
15	System for ordering elevator components	1994	SW	Pro	CA	Pre-test interview, usability test, post-test interview, heuristic evaluation	X	4
16	System for PC-based elemental analysis	1994	SW	Pro	CA	Pre-test questionnaire, usability test, post-test interview, heuristic evaluation	X	8
17	Television	1994	SP	Re	CA	Pre-test questionnaire, usability test, post-test interview, post-test questionnaire, heuristic evaluation	X	8
18	Television (everyday use)	1994	SP	Re	CA	Usability test, post-test questionnaire, post-test interview, heuristic evaluation	X	7
19	Television (taking into use)	1994	SP	Re	CA	Usability test, post-test interview, heuristic evaluation	X	9
20	Transmission management system	1994	SW	Pro	CA	Pre-test interview, usability test, post-test questionnaire, post-test interview, heuristic evaluation	X	5
21	Video recorder	1994	SP	Re	CA	Usability test, paired-user testing, post-test interview, heuristic evaluation	X	8
22	2 televisions	1995	SP	Re	CA	Pre-test questionnaire, usability test, post-test questionnaire	X	8
23	2 televisions	1995	SP	Re	CA	Pre-test questionnaire, visual walkthrough, usability test, post-test questionnaire, post-test interview, heuristic evaluation	X	7
24	2 video tape recorders	1995	SP	Re	CA	Usability test, post-test interview	X	7
25	2 video tape recorders	1995	SP	Re	CA	Usability test, post-test questionnaire, post-test interview	X	6
26	4 mobile phones	1995	SP	Re	RP	Pre-test questionnaire, usability test, post-test interview	X	17
27	Elevator monitoring and commanding system	1995	SW	Pro	RP	Heuristic evaluation, pluralistic usability walkthrough	-	2
28	Geographical information system for broadcasting information	1995	SW	Pro	CA	Pre-test interview, usability test, post-test questionnaire, post-test interview	X	4
29	Hand radio	1995	SP	Pro	RP	Pre-test questionnaire, usability test, post-test interview	X	14
30	Heart rate monitor	1995	SP	Re	RP	Pre-test questionnaire, usability test, post-test interview	X	3
31	Online newspaper	1995	SW	Re	CA	Pre-test questionnaire, usability test, post-test interview, post-test questionnaire	X	6
32	Online newspaper	1995	SW	Re	CA	Visual walkthrough, usability test, post-test questionnaire, post-test interview, heuristic evalu-	X	6

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
						ation		
33	Project coordinating system	1995	SW	Pro	CA	Pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	7
34	Remote control and support system	1995	SW	Pro	CA	Pre-test questionnaire, visual walkthrough, usability test, post-test questionnaire, post-test interview, heuristic evaluation	X	6
35	System for monitoring and coupling power-distribution networks	1995	SW	Pro	CA	Pre-test interview, usability test, post-test interview, heuristic evaluation	X	6
36	System to analyse telecommunications network alarm sequences	1995	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, visual walkthrough, usability test, post-test questionnaire, post-test interview	X	4
37	System to support university studies	1995	SW	Re	CA	Usability test, post–test interview	X	4
38	2 reverse vending machines	1996	SP	Re	RP	Cognitive walkthrough, usability test	X	20
39	3 office phones	1996	SP	Pro	RP	Pre-test questionnaire, usability test, paired-user testing, post-test questionnaire	X	6
40	Headquarter document system	1996	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	11
41	Mechanical and electrical engineering tool	1996	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, pluralistic usability walkthrough, post-test interview	-	3
42	Name service maintenance system	1996	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	4
43	News-on-demand digital media concept	1996	SW	Re	RP	Pre-test questionnaire, visual walkthrough, informal walkthrough, post-test interview	X	6
44	Patient monitor	1996	SP	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	-	4
45	Smart faucet	1996	SP	Re	RP	Usability test	X	5
46	System for marketing and customer service for power plants	1996	SW	Pro	CA	Heuristic evaluation, usability test, post-test interview, post-test questionnaire	X	3
47	System for registering course grades	1996	SW	Pro	RP	Pluralistic usability walkthrough	-	2
48	System for risk policy	1996	SW	Pro	RP	Pre-test questionnaire, usability test, post-test interview	X	3
49	Techno-economic design system	1996	SW	Pro	CA	Expert review, usability test, post-test interview, post-test questionnaire	X	4

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
50	Elevator monitoring and commanding system	1997	SW	Pro	RP	Pre-test questionnaire, visual walkthrough, usability test, post-test interview	X	2
51	Geographic information system for buildings and natural resources	1997	SW	Pro	CA	Heuristic evaluation, pre-test interview, usability test, post-test interview	X	5
52	Multimedia system for supporting maintenance work	1997	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	4
53	System for controlling traffic certificates	1997	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	6
54	System for improving teamwork in organisations	1997	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	7
55	System for secure data transfer	1997	SW	Re	CA	Heuristic evaluation, pre-test interview, usability test, post-test interview	X	5
56	System for web based documentation	1997	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	6
57	Client application for visualisation system	1998	SW	Pro	CA	Heuristic evaluation, usability test, post-test interview	X	5
58	Heart rate monitor	1998	SP	Re	CA	Heuristic evaluation, use diary, pre-test interview, visual walkthrough, usability test, post-test interview, survey (N=10)	X	14
59	Support system for call centres	1998	SW	Pro	RP	Contextual walkthrough	-	4
60	Support system for call centres	1998	SW	Pro	CA	Interview, heuristic evaluation, contextual walkthrough	-	4
61	Television with Internet connection	1998	SP	Re	CA	Heuristic evaluation, pre-test interview, usability test, paired-user testing, post-test interview	X	7
62	Web based survey editor	1998	SW	Re	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	4
63	Web store	1998	SW	Re	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview, post-test questionnaire, pluralistic usability walkthrough (N=20)	X	24
64	Field terminal for maintenance men	1999	SP	Pro	CA	Heuristic evaluation, pre-test interview, usability test, post-test interview	X	5
65	Gaming slot machine	1999	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, contextual interview (N=?), pre-test questionnaire, paired-user testing, usability test, post-test questionnaire, post-test interview	X	7
66	Heart rate monitor	1999	SP	Re	CA	Cognitive walkthrough, heuristic evaluation, usability test, post-test interview	X	5

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
67	Interactive hotel system	1999	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, usability test, post-test questionnaire, post-test interview	X	8
68	Mobile phone	1999	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, pre-test interview, paired-user testing, usability test, post-test interview	X	6
69	Teller terminal	1999	SP	Pro	CA	Heuristic evaluation, usability test, post-test questionnaire, post-test interview	X	7
70	2 gaming slot machines	2000	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, pre-test questionnaire, paired-user testing, group testing, usability test, post-test interview	X	8
71	Discussion group service for organisations	2000	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	4
72	Heart rate monitor	2000	SP	Re	CA	Cognitive walkthrough, pre-test questionnaire, usability test, post-test interview, use diary	X	7
73	Management tool for elevator display	2000	SW	Pro	CA	Heuristic evaluation, paired-user testing, post-test questionnaire, post-test interview	-	8
74	Mobile phone	2000	SP	Re	CA	Cognitive walkthrough, survey (N=20), usability test (N=5), post-test interview, use diary (N=4)	X	29
75	System for analysing electrocardiogram data	2000	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	4
76	3 gaming slot machines	2001	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, pre-test questionnaire, paired-user testing, informal walkthrough, usability test, post-test questionnaire, post-test interview	X	6
77	3 gaming slot machines	2001	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, paired-user testing, informal walkthrough, post-test questionnaire	X	6
78	Central monitoring software	2001	SW	Pro	CA	Heuristic evaluation, usability test, post-test interview, post-test questionnaire	X	4
79	Configuration tool for remote monitoring	2001	SW	Pro	CA	Heuristic evaluation, usability test, post-test questionnaire, post-test interview	X	3
80	Heart rate monitor	2001	SP	Re	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	5
81	Heart rate monitor	2001	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, usability test, post-test questionnaire, post-test interview	X	4

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
82	Scheduling system for building projects	2001	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, usability test, post-test interview	X	5
83	Web portal for technology and innovation programmes	2001	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	5
84	Web service for supplementary education	2001	SW	Re	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	5
85	Company intranet for managers of casino games	2002	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, usability test, post-test questionnaire, post-test interview	X	5
86	Company intranet: guidelines and forms	2002	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, pre-test interview, usability test, post-test interview, card sorting, post-test questionnaire	X	7
87	Elevator destination control system	2002	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, usability test, post-test questionnaire, post-test interview	X	5
88	Heart rate monitor (everyday use)	2002	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, pre-test interview, usability test, post-test interview	X	5
89	Heart rate monitor (taking into use)	2002	SP	Re	CA	Heuristic evaluation, pre-test interview, usability test, post-test interview	X	5
90	Patient monitoring system in a personal digital assistant	2002	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	4
91	Scheduling and resource management system for building projects	2002	SW	Pro	CA	Heuristic evaluation, usability test, post-test interview, post-test questionnaire	X	5
92	System for modelling and detailing steel structures	2002	SW	Pro	CA	Heuristic evaluation, usability test, card sorting, post-test questionnaire, post-test interview, pluralistic usability walkthrough	X	8
93	Web service for adult education	2002	SW	Re	CA	Heuristic evaluation, pre-test interview, usability test, visual walkthrough, post-test questionnaire, post-test interview	X	5
94	Anaesthesia monitor	2003	SP	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview, post-test questionnaire	X	5
95	Calendar service for a heart rate monitor	2003	SW	Re	CA	Heuristic evaluation, pre-test interview, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	6
96	Gaming slot machine for group play	2003	SP	Re	CA	Heuristic evaluation, informal walkthrough, post-test interview, post-test questionnaire	X	9
97	Heart rate monitor	2003	SP	Re	CA	Cognitive walkthrough, heuristic evaluation, usability test, peer tutoring, post-test questionnaire, post-test interview, probe	X	6

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
98	Heart rate monitor	2003	SP	Re	CA	Heuristic evaluation, usability test, post-test questionnaire, post-test interview, symbol and terminology test (N=5)	X	10
99	System for authoring mixed reality	2003	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire	X	5
100	System for authoring mixed reality	2003	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, paired-user testing, post-test questionnaire, post-test interview	X	7
101	System for optimising road building	2003	SW	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test questionnaire, post-test interview	X	5
102	Web based system for information services for engineering	2003	SW	Pro	CA	Heuristic evaluation, pluralistic usability walkthrough, use diary	-	5
103	Application for planning and monitoring physical training	2008	SW	Re	CA	Heuristic evaluation, cognitive walkthrough, pre-test questionnaire, usability test, SUS questionnaire, post-test interview	X	6
104	IRC gallery	2008	SW	Re	CA	Heuristic evaluation, pre-test questionnaire, informal walkthrough, post-test interview	X	6
105	Mobile map application	2008	SW	Re	CA	Heuristic evaluation, cognitive walkthrough, usability test, post-test interview	X	6
106	Patient monitoring through web service	2008	SW	Pro	CA	Heuristic evaluation, pre-test interview, usability test, peer tutoring, post-test interview, SUS questionnaire	X	6
107	Payment terminal for a gaming slot machine	2008	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, usability test, paired-user testing, SUS questionnaire, post-test interview	X	7
108	Removable user interface for elevator installation and service	2008	SP	Pro	CA	Heuristic evaluation, cognitive walkthrough, usability test with an opportunity to call to a "colleague"	X	5
109	Enterprise recourse planning system	2009	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, usability test, SUS questionnaire, post-test interview	X	6
11	Mobile banking application	2009	SW	Re	CA	Expert review, interview, usability test	X	4
111	Mobile music store	2009	SW	Re	RP	Pre-test questionnaire, usability test, post-test interview	X	30
112	Mobile ticketing service	2009	SW	Re	RP	Cognitive walkthrough, action analysis	-	0
113	Multi-user interactive game board	2009	SW	Re	CA	Heuristic evaluation, informal walkthrough, multi-user testing, post-test questionnaire, post-test interview	X	16
114	Personal training application	2009	SW	Re	CA	Heuristic evaluation, cognitive walkthrough, informal walkthrough, post-test interview	X	8

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
115	Portal for building permissions	2009	SW	Pro, Re	CA	Heuristic evaluation, cognitive walkthrough, pre-test questionnaire, contextual walkthrough, usability test, post-test questionnaire, post-test interview	X	6
116	System for monitoring group exercises	2009	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, pre-test questionnaire, contextual walkthrough, usability test, post-test interview	X	8
117	Video surveillance system	2009	SW	Pro	CA	Heuristic evaluation, usability test, post-test interview	X	6
118	Web pages of an open university	2009	SW	Re	CA	Heuristic evaluation, cognitive walkthrough, usability test, card sorting, post-test questionnaire	X	15
119	Enterprise recourse planning system	2010	SW	Pro	RP	Heuristic evaluation, expert review	-	0
120	Enterprise recourse planning system	2010	SW	Pro	CA	Contextual interview, usability test, paired-user testing, post-task interview	X	16
121	Heart rate monitor	2010	SP	Re	CA	Heuristic evaluation, cognitive walkthrough, focus group, participatory design, pre-test questionnaire, usability test	X	11
122	Mobile banking application	2010	SW	Re	RP	Usability test, post-test interview	X	5
123	Mobile ticketing service	2010	SW	Re	RP	Pre-test questionnaire, usability test, post-test interview, post-test questionnaire	X	39
124	Mobile ticketing service	2010	SW	Re	CA	Heuristic evaluation, cognitive walkthrough, pre-test questionnaire, usability test, SUS questionnaire, post-test interview	X	7
125	Patient monitor	2010	SP	Pro	CA	Heuristic evaluation, pre-test questionnaire, usability test, post-test interview	X	9
126	System for commenting course assignments	2010	SW	Pro	CA	Heuristic evaluation, usability test	X	4
127	Travel and expense management system	2010	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, usability test, paired-user testing	X	7
128	Building information system	2011	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, design guidelines review, contextual interview, informal walkthrough, pre-test questionnaire, paired-user testing, usability test, SUS questionnaire, post-test interview	X	8
129	Class room technical equipment	2011	SP	Pro	RP	Observation (N=20), interview (N=3), use diary (N=3), survey (N=very small), activity logging, cognitive walkthrough, heuristic evaluation	-	26

ID	System	Year	SW/ SP	Pro/ Re	RP/ CA	Methods	TA	Users
130	Database for designing accessible buildings	2011	SW	Pro	RP	Heuristic evaluation, cognitive walkthrough, contextual interview (N=4), survey (N=16)	-	20
131	Enterprise recourse planning system	2011	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, design guidelines review, pre-test interview, paired-user testing, post-test interview, SUS questionnaire	-	6
132	Entertainment system for big crowds	2011	SW	Pro	CA	Heuristic evaluation, design guidelines review, contextual interview, usability test, SUS questionnaire, post-test interview, paired-user testing, informal walkthrough	X	10
133	Mobile service supporting nature and culture heritage tourism	2011	SW	Re	RP	Interview (N=13), survey (N=3), activity logging, focus group (N=8)	-	24
134	Web based activity monitoring service	2011	SW	Pro	CA	Heuristic evaluation, cognitive walkthrough, usability test, pre-test interview, post-task interview, post-test interview, SUS questionnaire	X	6
135	Database for designing accessible buildings	2013	SW	Pro	RP	Pre-test interview, visual walkthrough, usability test, post-test interview	X	9
136	Dental software	2013	SW	Pro	CA	Heuristic evaluation, action analysis, design guidelines review, contextual interview, pre-test interview, informal walkthrough, SUS questionnaire, post-test interview	X	11
137	Information service for healthcare professionals	2013	SW	Pro	CA	Heuristic evaluation, contextual interview, pre-test interview, usability test, visual walkthrough with heat maps, SUS questionnaire, post-test interview	X	9
138	Mobile financial application	2013	SW	Re	CA	Heuristic evaluation, design guidelines review, usability test, SUS questionnaire	X	10
139	Web interface for the blind	2013	SW	Re	CA	Heuristic evaluation, interview, contextual interview, usability test, SUS questionnaire, AXE with objects	X	9
140	2 financial management systems for SMEs	2014	SW	Pro	CA	Heuristic walkthrough, pre-test interview, informal walkthrough, usability test, SUS questionnaire, post-test interview	X	8
141	2 research information systems	2014	SW	Pro	CA	Heuristic walkthrough, SUS questionnaire, SAM questionnaire, post-test interview, user-interface walkthrough	-	17
142	3 mobile conference program applications	2014	SW	Pro, Re	CA	Heuristic evaluation, contextual interview, usability test, paired-user testing, SUS questionnaire, post-test interview	X	12
143	3 mobile person-to-person payment applications	2014	SW	Re	CA	Contextual interview, heuristic evaluation, cognitive walkthrough, design guidelines review, usability test, SUS questionnaire, trust questionnaire, post-test interview	X	9

Appendix B: Background questionnaire (in Finnish)

T-121.3110-kurssin taustatietokysely

Tämä kyselylomake on osa T-121.3110-kurssin tehtävää 3: Osallistuminen käyttäjätestiin. Kyselyn vastausten perusteella opiskelijat jaetaan eri testeihin. Kyselyyn tulee vastata ti 26.1. mennessä. Valinnoista kerrotaan pe 29.1. mennessä, jolloin tulee myös tarkemmat ohjeet testeihin ilmoittautumisesta. Testit ovat viikoilla 6-9.

1. Opiskelijanumero: _____
2. Tutkinto-ohjelma:
   - Tik / IV / TLT / TKK:n muu perustutkinto-ohjelma /
   - jatko-opiskelija / muu
3. Tämänhetkinen opintopisteiden määrä
   - alle 60
   - 60-99
   - 100-139
   - 140-179
   - 180-220
   - yli 220
4. Sukupuoli
   - mies
   - nainen
5. Ikä
   - alle 20
   - 20-25
   - 26-30
   - yli 30
6. Millaisen kännykän omistat?
   - En omista kännykkää
   - Peruspuhelin ilman lisäominaisuuksia, kuten kamera tai sähköposti
   - Multimediapuhelin
   - Puhelin, jolla voin lukea sähköpostini ja selata web-sivuja
   - Älypuhelin (mahdollisuus asentaa itse lisää ohjelmia)

7. Millä perusteella olet valinnut nykyisen puhelimesi?

_____

_____

8. Mikä on nykyisen puhelimesi merkki ja malli ? (esim. Nokia N96, Sony Ericsson XPERIA. Jos käytät useampaa, voit listata ne kaikki. Laita ensimmäiseksi se, jota käytät eniten.)

_____

9. Käytätkö kännykkää viikoittain

Puhumiseen
Tekstiviesteihin
Multimediaviesteihin
Valokuvien ottamiseen
Musiikin kuunteluun
Videoiden katseluun
Kalenterina
Sähköpostien lukemiseen ja lähettämiseen
www-selailuun
Muuhun: _____

10. Oletko maksanut puhelimellasi

HKL:n kertalipun
Juoma- tai välipala-automaatin ostoksen
Pysäköinnin
Jotain muuta: _____

11. Arvioi suhtautumistasi uusiin teknologioihin asteikolla 1-5, jossa

1= täysin eri mieltä

5= täysin samaa mieltä

Olen erittäin kiinnostunut uusista teknologioista ja niitä hyödyntävistä sovelluksista
Otan uudet teknologiset sovellukset ja tuotteet mahdollisimman pian omaan käyttööni

12. Oletko valmis osallistumaan noin tunnin kestävään käyttäjätestiin, joka videoidaan analysointeja varten?

Kyllä
Ei (ota yhteyttä kurssin henkilökuntaan korvataksesi tehtävän)

13. Sähköpostiosoite, johon voi lähettää tietoa testijärjestelyistä

Appendix C: Experiment introduction (in Finnish)

Testitilanteen alustus (10-15 min):

- esittely: kuka olen, onko muita paikalla, tilat, pientä tarjoilua (5-7 min)
- eTicket (1-2 min)
  - Tutkimushanke, jossa kehitellään palveluita kännykkään. Eräänä palveluna on työn alla bussilippujen tai muiden kertaluonteisten lippujen hankinta kännykällä. Tampereella saa esimerkiksi myös uimahalliin lippuja vastaavan tyyppisellä palvelulla.
  - Tähän on kehitetty eTicket-niminen sovellus. Sitä on kehitetty hyvin paljon tekniikan ehdoin ja nyt halutaankin arvioida, mille tasolle käytettävyys on siinä saatu.
- Testin kuvaus (4-6 min)
  - Järjestelmä on eka versio eikä kovin hiottu, joten erilaisia ongelmia tulee luultavasti vastaan. Kannattaa siis muistaa, että arvioinnin kohteena on eTicket etkä sinä.
  - Testiin kuuluu kuusi tehtävää ja kaksi jälkikyselyä. Jos jokin tehtävä tuntuu liian vaikealta etkä saa sitä suosiolla tehtyä, voit jättää sen kesken. Jos koko testi alkaa ahdistaa liikaa, senkin voi keskeyttää. Saat silti suoritusmerkinnän kurssille, kunhan täytät molemmat jälkikyselykaavakkeet verkossa.
  - Analysointia ja muuta jatkokäsittelyä varten tallennamme testitilanteen tuolla kameralla. Nauhoitetussa kuvassa näkyy näytön kuva, kännykkä ja hieman sormia, ei kasvoja. Olisi hyvä, että pitäisit kännykkää mahdollisimman paljon alustan kohdalla, jotta kännykkä näkyisi kuvassa.
  - Materiaali käsitellään niin, että kenenkään henkilöllisyys ei käy ilmi tehtävissä raporteissa.
- Testiasetelman esittely sopivan vaihtoehdon mukaisesti

  1. Testailemme samalla hiukan erilaisia testiasetelmia ja tällä kertaa emme käytä ääneen ajattelua niin kuin yleensä. Lisäksi ohjaaja eli minä siirryn ensimmäisen tehtävän jälkeen tuonne takahuoneeseen. Tee siis tehtäviä omaan tahtiisi ja kerro kunkin tehtävän jälkeen, koska olet valmis. Tuppisuuna ei tarvitse muutenkaan olla, jos tekee mieli kommentoida jotain ääneen, mutta erikseen ei tarvitse ajatella ääneen. Olen tosi-

aan testin alun tässä vieressä, mutta sitten siirryn takahuoneeseen. Tarpeen tullen tulen apuun, mutta toimi mahdollisimman paljon niin kuin olisit “omillasi”.

2. Testailemme samalla hiukan erilaisia testiasetelmia ja tällä kertaa emme käytä ääneen ajattelua niin kuin yleensä. Tee siis tehtäviä omaan tahtiisi ja kerro kunkin tehtävän jälkeen, koska olet valmis. Tuppisuuna ei tarvitse muutenkaan olla, jos tekee mieli kommentoida jotain ääneen, mutta erikseen ei tarvitse ajatella ääneen. Olen testin ajan tässä vieressä ja annan tehtäviä suullisesti ja muistin tueksi paperillakin. En juurikaan vastaa kysymyksiin testin aikana, mutta autan toki tarpeen tullen.

3. Jotta saamme mahdollisimman paljon irti testistä, olisi hyvä, että ajattelisit ääneen testin aikana eli kertoisit mitä etsit, kaipaat tai odotat järjestelmältä ja mitä esimerkiksi ajattelet eri vaihtoehtojen tarkoittavan. Voin näyttää esimerkin. (Säädä soittoääntä)
Testailemme tässä samalla hiukan erilaisia testiasetelmia. Tällä kertaa ohjaaja eli minä siirryn ensimmäisen tehtävän jälkeen tuonne takahuoneeseen. Tee siis tehtäviä omaan tahtiisi ja kerro kunkin tehtävän jälkeen, koska olet valmis, ja siirry seuraavaan tehtävään. Tarpeen tullen tulen apuun, mutta toimi mahdollisimman paljon niin kuin olisit “omillasi”.

4. Jotta saamme mahdollisimman paljon irti testistä, olisi hyvä, että ajattelisit ääneen testin aikana eli kertoisit mitä etsit, kaipaat tai odotat järjestelmältä ja mitä esimerkiksi ajattelet eri vaihtoehtojen tarkoittavan. Voin näyttää esimerkin. (Säädä soittoääntä)
Olen testin ajan tässä vieressä ja annan tehtäviä suullisesti ja muistin tueksi paperillakin. En juurikaan vastaa kysymyksiin testin aikana, mutta autan toki tarpeen tullen.

- Testin aluksi kerron kuvauksen tilanteesta, jossa tehtäviä tehdään, eli skenaarion. Annan tehtäviä (yksi kerrallaan sekä suullisesti että paperilla/paperilla). Tee tehtävät niin kuin koet itsellesi luontevaksi ja kerro, kun olet mielestäsi saanut kunkin tehtävän loppuun.
- Kysyttävää?

Appendix D: Test scenario and tasks (in Finnish)

Skenaario (2 min)
Tampereella järjestetään mielenkiintoinen kahden päivän seminaari käytettävyydestä. Olet majoittunut paikallisen ystäväsi luo jo edellisenä iltana. Selvität ystäväsi kanssa, mikä on paras tapa hankkiutua seminaaripaikkaan. Hän suosittelee sopivaa bussia. Samalla hän muistaa, että hänellä on yksi ylimääräinen kännykkä, jossa on matkalippusovellus, jolla lipun saa halvemmalla kuin kuskilta. Puhelin ei ole hänellä aktiivikäytössä, koska sen näppäimistö on hieman outo, joten saat sen lainaan, kunhan et käytä hänen maksutietojaan vaan omiasi. Ystäväsi on pian kiiruhtamassa vielä juhliin, mutta ehtii vielä ennen lähtöään katsomaan, että löydät tarvitun ohjelman kännykästä.

Tehtävät (20-30 min)
Tarvitun ohjelman nimi on eTicket. Etsi se kännykästä. *(5 min)*
(*Tässä vaiheessa ystävä vielä auttaa kännykän käytössä tarvittaessa)*
Nyt ystävä joutuu lähtemään. Lähtiessään hän vielä muistuttaa, että käytäthän omaa luottokorttiasi.

Ohje, kun jää yksin:
Nyt minä siirryn takahuoneeseen. Tee tehtäviä omaan tahtiisi ja kerro aina jokaisen tehtävän päätteeksi, kun olet mielestäsi valmis. Siirry sitten seuraavaan tehtävään.

Ohje, kun ohjaaja jää viereen:
Nyt ystävä siis lähti. Minä jään vielä ohjaamaan testiä, mutta, kuten mainitsin tuossa aiemmin, en juurikaan vastaa kysymyksiin. Tee tehtäviä omaan tahtiisi ja kerro aina jokaisen tehtävän päätteeksi, kun olet mielestäsi valmis. Annan sitten seuraavan tehtävän.

Lisää ohjelmaan omat maksutietosi, jottet vahingossa käytä ystäväsi rahoja. Ohessa on Visa Pahvi, jota voit käyttää. *(5 min)*

Osta yhden päivän lippu Tampereen paikallisbussiin huomista varten.

Sitten siirryt valmistelemaan muita asioita huomisen varalle ja suljet sovelluksen. *(3 min)*

Yö meni hyvin ja olet aamulla matkalla bussipysäkille. Bussi tulee muutaman minuutin päästä, joten on sopiva hetki laittaa illalla hankkimasi lippu siihen kuntoon, että sillä voi matkustaa.

Bussin sisällä kännykkä pitää viedä lähelle lukulaitetta, joka tarkistaa, että siinä on voimassa oleva matkalippu. Sovellus voi silloin olla suljettuna, mutta jätä se vielä auki matkan ajaksi. *(2 min)*

Matka seminaaripaikkaan vie jonkin aikaa, joten päätät muistaessa hankkia matkalipun myös seuraavalle päivälle. Uusit siis käytössäsi olevan lipun.

Varmista vielä, että sait lipun kännykkääsi. *(5 min)*

Bussin pitäisi olla seminaaripaikassa noin puolelta. Jos sinulla on vielä muutama minuutti aikaa, poista omat maksutietosi sovelluksesta. Jos bussi on jo melkein perillä, jätä tuo toiseen kertaan. *(2 min)*

Appendix E: Post-task questionnaire on system

The questionnaire used to collect the users' experiences with the eTicket system is presented in the following table. On the left, there is the original Finnish version, and on the right, the English translation. The questionnaire was provided as a web survey.

Original Finnish version	English translation
eTicketin käyttökokemus: arvioi asteikolla 1-5 (1=täysin eri mieltä, 5=täysin samaa mieltä) • löysin helposti eTicketistä tarvitsemani toiminnot • lipun ostaminen oli yksinkertaista • lipun käyttö bussissa oli helppoa • lipun uusiminen oli helppoa • jos kertalippu eTicketillä ostettuna olisi halvempi (esim. 10%) kuin kuskilta ostettu, ostaisin lipun eTicketillä • jos kertalippu eTicketistä ja kuskilta ostettuna olisivat saman hintaisia, ostaisin lipun eTicketillä • suosittelisin eTicketiä myös ystävilleni • haluaisin ostaa muitakin lippuja (esim. uimahalliin tai museoon) eTicketin avulla • eTicketin käyttö oli kaiken kaikkiaan miellyttävää	Use experience with eTicket: Assess on the scale from 1 to 5 (1=totally disagree, 5=totally agree) • I found easily the functions that I needed from the eTicket • Buying a ticket was easy • Using the ticket in the bus was easy • Re-buying a ticket was easy • If the single ticket through eTicket was cheaper (*e.g.* 10%) than buying in the bus, I would buy the ticket with eTicket • If the single ticket through eTicket was of same price as bought inside the bus, I would buy the ticket with eTicket • I would recommend eTicket also to my friends • I would like to buy also other tickets (*e.g.* to a swimming hall or a museum) with eTicket • Overall, the use of eTicket was pleasant
Kysymykset, joissa vapaa tekstikenttä: • Mitä muita kertaluonteisia lippuja haluaisit ostaa kännykälläsi? • Kommentteja sovelluksesta tai terveisiä kehittäjille?	Questions with freeform text field: • What other single time used tickets would you like to buy with your mobile phone? • Comments about the application or feedback to the developers?

The feedback text that the users received from the first part of the questionnaire also included a link to the second part with questions on the test setting. The second part of the questionnaire is presented in Appendix F.

Appendix F: Post-task questionnaire on test setting

The questionnaire used to collect the users' experiences about the test setting is presented in the following table. On the left, there is the original Finnish version, and on the right, the English translation. The questionnaire was provided as a web survey following the questionnaire on the experiences with the use of the eTicket application. The first questions were the same for everyone, but the remainder of the 1-5 scale questions depended on the test setting that the user belonged to (1st digit T/0: Thinking aloud used or not; 2nd digit M/0: Test moderator present in the test room or not).

Original Finnish version	English translation
Testiasetelmaan liittyvät kysymykset (kaikille yhteiset): arvioi asteikolla 1-5 (1=täysin eri mieltä, 5=täysin samaa mieltä) • testitilanne tuntui luontevalta enkä jännittänyt suoritustani • testitehtävien tekeminen oli miellyttävää • skenaario eli tilannekuvaus testin alussa auttoi minua ymmärtämään tehtäviä • ymmärsin tehtävistä hyvin, mitä minun odotettiin tekevän • pyrin tekemään tehtävät mahdollisimman nopeasti • pyrin tekemään tehtävät ensi yrittämällä mahdollisimman oikein ja välttämään virheitä • pyrin tutustumaan järjestelmään kokonaisuutena ennen tehtävien suoritusta • pyrin kartoittamaan mahdollisimman paljon erilaisia tapoja suorittaa kukin tehtävä ennen kuin valitsin oman toimintatapani • tein tehtäviä kärsivällisemmin kuin normaalitilanteissa	Questions related to the test setting (common for all): Assess on the scale from 1 to 5 (1=totally disagree, 5=totally agree) • The test situation felt natural and I did not stress about my performance • Doing the tasks was pleasant • The scenario in the beginning of the test helped me to understand the tasks • I understood easily what I was expected to do • I tried to do the tasks as quickly as possible • I tried to do the tasks as correctly as possible in the first try and to avoid mistakes • I tried to get familiar with the whole system before starting the tasks • I tried to find as many solutions as possible for each task before selecting my own solution • I did the tasks more patiently than usual
Testiasetelmaan liittyvät kysymykset (ryhmäkohtaiset): arvioi asteikolla 1-5 (1=täysin eri mieltä, 5=täysin samaa mieltä) 00 • tehtävien tekeminen vaiti tuntui luontevalta • kaipasin testin ohjaajaa testi-huoneeseen kanssani 0M • tehtävien tekeminen vaiti tuntui luontevalta • ohjaajan läsnäolo häiritsi tehtävien suoritustani • ohjaajan läsnäolo kannusti minua suorittamaan tehtävät loppuun T0	Questions related to the test setting (according to the test group): Assess on the scale from 1 to 5 (1=totally disagree, 5=totally agree) 00 • Doing the tasks silently felt natural • I would have preferred to have the test moderator in the same room with me 0M • Doing the tasks silently felt natural • The presence of the test moderator disturbed my performance • The presence of the test moderator encouraged me to finish all the tasks T0 • Thinking aloud felt natural • The thinking aloud affected to the extent

• ääneen ajattelu tuntui luontevalta • ääneen ajattelu vaikutti siihen, miten paljon pohdin niin sanottua oikeaa ratkaisua • kaipasin testin ohjaajaa testihuoneeseen kanssani TM • ääneen ajattelu tuntui luontevalta • ääneen ajattelu vaikutti siihen, miten paljon pohdin niin sanottua oikeaa ratkaisua • ohjaajan läsnäolo häiritsi tehtävien suoritustani • ohjaajan läsnäolo kannusti minua suorittamaan tehtävät loppuun	to which I thought about the right way of doing the tasks • I would have preferred to have the test moderator in the same room with me TM • Thinking aloud felt natural • The thinking aloud affected to the extent to which I thought about the right way of doing the tasks • The presence of the test moderator disturbed my performance • The presence of the test moderator encouraged me to finish all the tasks
Yhteisenä kaikille loppuun vapaa-muotoinen kysymys: • Kommentteja testitilanteesta ja testin sujumisesta?	A common question to all users with a freeform text field: • Comments about the test setting and the flow of the test?
Kuittauksena vastauksista: "Kiitos arvokkaasta panoksestasi tässä tutkimuksessa! Ethän kerro muille testitilanteesta ennen kuin he ovat itse osallistuneet testiin.”	Feedback text after submission: ”Thank you for your valuable time and effort in this study! Please, do not tell about the test situation to other participants before their own tests.”

Usability testing has become a standard method when evaluating the usability of various systems with real users. Despite its general use and wide acceptance, the factors of usability testing have been given little attention in academic forums. This thesis studies methods and factors of usability testing focusing on the effects of thinking aloud and the presence of a test moderator. It combines an extensive literature review with experiences of usability testing from 22 years covering 143 usability studies. It also includes an experimental part with relaxed thinking aloud and the presence of a test moderator as independent variables.

Aalto-DD 75/2015

ISBN 978-952-60-6226-6 (printed)
ISBN 978-952-60-6227-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi