Process for Usability Evaluation for Selecting an Information System in Public Procurement

Measuring Usability of Health and Social Welfare IS

Mari Tyllinen
Process for Usability Evaluation for Selecting an Information System in Public Procurement

Measuring Usability of Health and Social Welfare IS

Mari Tyllinen

A doctoral thesis 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 AS1 of the school on 18 August 2023 at 12:00.

Aalto University
School of Science
Department of Computer Science
Human-Centred Health Informatics (HCHI)
Abstract

Information systems (IS) that support work efficiency could partly solve the challenges facing public healthcare. However, poor usability and users' negative experiences of IS used have been discussed both in academic research and media extensively over the last decade. The social welfare sector seems to suffer similarly. IS used for patient and client care is acquired mainly as further configured products. Considering usability already during the selection has been deemed necessary.

Including comparative usability evaluation (UE) at this stage does not have a ready solution especially in the regulated environment of public procurement, and in the case of IS with a multitude of use contexts and user groups. The research questions are: RQ1) What are effective evaluation methods and metrics for measuring usability during the selection phase in public procurement of IS? RQ2) How to construct the evaluation procedure supporting the quantification and comparison of usability for this use?

The research project was identified in a public procurement project for a patient and client IS in Finland. An action design research approach was used. Artifacts (UE methods quantifying usability) and an evaluation procedure were developed iteratively and collaboratively. The procedure was implemented in the project and its use is evaluated. Also, national surveys on different professional groups’ experiences of IS usability were analysed together.

The usability attributes, methods and measures for a two-phase procedure are introduced. Perceived usability questionnaire, heuristic evaluation for demonstrations (HED), and paired-user testing are presented in detail along with scoring. The survey analysis provided complementing context insights for different user groups.

All three methods are suitable for public procurement. HED was found most efficient and comprehensive, but with paired-user testing more attributes can be evaluated. They differ also in user involvement, level of usability expertise and produced information. The evaluated attributes are determined based on the goals of the procurement. When planning the evaluation details, the user groups’ differences should be considered. This dissertation presents one way of scoring overall usability. Additionally, the first version of a process model for including usability measurement of IS in the public procurement selection is presented.

In conclusion, there are multiple viewpoints emerging from the requirements of the regulatory environment and the goals of each procurement that affect evaluation planning and implementation. This adds complexity compared to regular summative UE, which is also usually more confined in scope. Based on the results, there is no reason not to measure usability as a selection criterion, even when procuring a large-scale IS. Using a few measures can reveal the differences, but the number of measures should depend on the goals.

Keywords usability evaluation, measuring usability, public procurement, information system
Acknowledgements

Working on my doctoral thesis has been a long and winding journey, perhaps longer than I ever expected but nevertheless enjoyable. I wish to extend my thanks to all those who have traveled with me and beside me during all or parts of this journey.

The research for this thesis was carried out while working at the Department of Computer Science at Aalto University in two different research groups: first with Prof. Marko Nieminen and later with Assistant Prof. Johanna Viitanen. I thank you both for your contributions in supervising and advising this work in different roles from the very beginning and throughout the years. Thank you for all the constructive discussions, valuable feedback and providing a safe and supportive academic environment. Marko, thank you for seeing that collaborating with Johanna would be fruitful. Johanna, thank you for making it a possibility to work on a meaningful and interesting real-life project, guiding me to this topic and also for your friendship. I am also grateful to the preliminary examiners of this thesis, Prof. Åsa Cajander and Prof. Morten Hertzum – with their insights I was able to make final improvements to the manuscript. My gratitude also goes to Prof. Christian Nøhr who has agreed to act as the opponent in the upcoming public defence.

I wish to thank all my other co-authors, Dr. Tinja Lääveri, Elina Tynkkynen, Janne Pitkänen, Matti Pitkäranta and Antti K. Haapala for collaboration. A special thank you to Tinja for giving a different perspective from another discipline on scientific work and always reminding to keep the writing brief.

The empirical research for this thesis was collected while also working outside academia. Without all those working in the procurement project, this research work would not have been possible (or nearly as enjoyable). I’ve learned so much about public procurement, healthcare and social welfare. Thank you to Heikki Onnela for taking me on the project, and Sirkka Häkkinen, Raija Huttunen, Piritta Wartiainen, Veli-Pekka Lehtonen, Anna Eskola and Leena Keskisaari-Kajaste for your collaboration as well as all the other
professionals working with the product comparisons. I am also grateful for the
group of people that had a passion for the topic and brought into existence the
separate questionnaire study, especially Samuel Salovaara and Katri Ylönen.

Last year, after many years of working on the thesis with limited time, I was
fortunate to be able to take a study leave and devote six months to writing this
compilation full time. Thank you to my competent and wonderful team for
making it easy to take a much-needed break. During this time, I was able to
spend four months on beautiful Vancouver Island visiting University of Victo-
ria. I am grateful especially to Prof. Andre Kushniruk but also Prof. Elizabeth
Borycki and Assistant Prof. Helen Monkman for making it possible and for the
insightful conversations during my time there.

During my studies I’ve also had the chance to work with many great people
at the university. Thank you to STRATUS research group members during the
early years; especially to Dr. Mika P. Nieminen for coaxing me to start the
post-graduate study journey with my first thesis topic so many years ago and
Dr. Sirpa Riihialho for giving a solid foundation to usability evaluation from
which to build on. Thank you to all colleagues in HCHI research group for the
academic discussions and community during the last years. Especially
throughout the pandemic I’ve felt very blessed to have such caring and smart
fellow doctoral students as Paula Valkonen and Nina Karisalmi to share the
ups and downs with. Thank you, Paula, also for reading the first full draft of
this thesis – your comments were invaluable!

Working in the project has also brought so many people into my life whom I
would have never met otherwise. Former and current colleagues who I’ve
shared time with at work and outside of it – thank you for all the fun times!
Special thank you goes to Adam for all the interesting conversations and your
encouragement during the past year.

A thank you goes also to all my dear childhood and university friends who
have given me much needed distractions and support during these years. I
appreciate you so much! And finally, a heartfelt thank you to all my family for
always being there for me.

This research work was partly funded by the HANKI project (funded by the
Finnish Work Environment Fund) and with travel grants from the Finnish
Work Environment Fund and Tietotekniikan tutkimussäätiö.

Vantaa, June 25, 2023,

Mari Tyllinen
Contents

Acknowledgements .............................................................................................................. I
List of Abbreviations .......................................................................................................... V
List of Publications .............................................................................................................. VII
Author’s Contribution ........................................................................................................ IX
1. Introduction ..................................................................................................................... 1
   1.1 Motivations for the Dissertation ................................................................................. 2
   1.2 Research Environment .............................................................................................. 4
   1.3 Key Concepts and Domain of the Research .............................................................. 8
   1.4 Objectives and Scope ............................................................................................... 12
   1.5 Research Questions .................................................................................................. 14
   1.6 Anticipated Contributions of the Dissertation .......................................................... 16
   1.7 Dissertation Structure ............................................................................................. 16
2. Theoretical Foundation .................................................................................................... 19
   2.1 Public Information System (IS) Procurement ............................................................. 19
   2.2 Usability in IS Procurement ...................................................................................... 29
   2.3 Usability Evaluation .................................................................................................. 42
   2.4 Synthesis .................................................................................................................. 55
3. Action Design Research: Developing Usability Assessment ......................................... 63
   3.1 Research Paradigm and Approach ............................................................................. 63
   3.2 Research Context ..................................................................................................... 66
   3.3 Research Design ....................................................................................................... 68
   3.4 Building the Artifacts: Iterative Methods Development ............................................. 70
   3.5 Intervening and Evaluating: Implementing the Methods .......................................... 72
List of Abbreviations

ADR           Action Design Research
AR            Action Research
BIE           Building, Intervention, Evaluation
CIS           Client Information System
CLIPS         Clinical Information Processing Scenarios
COTS          Commercial-off-the-shelf (Software)
CSUQ          Computer System Usability Questionnaire
CW            Cognitive Walkthrough (Method)
DPUQ          Demonstration based Perceived Usability Questionnaire (Method)
DSR           Design Science Research
EHR           Electronic Health Record (System)
EMR           Electronic Medical Record (System)
EPR           Electronic Patient Record (System)
EMAT          Economically Most Advantageous Tender (in procurement)
ESCR          Electronic Social Care Record (System)
HCI           Human-Computer Interaction
HFE           Human Factors Engineering
HF&E          Human Factors & Ergonomics
<table>
<thead>
<tr>
<th>HE</th>
<th>Heuristic Evaluation (Method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HED</td>
<td>Heuristic Evaluation during Demonstrations (Method)</td>
</tr>
<tr>
<td>HIS</td>
<td>Health Information System</td>
</tr>
<tr>
<td>ICC</td>
<td>Intra-class Correlation Coefficient</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communications Technology</td>
</tr>
<tr>
<td>IRA</td>
<td>Inter-Rater Agreement</td>
</tr>
<tr>
<td>IRR</td>
<td>Inter-Rater Reliability</td>
</tr>
<tr>
<td>IS</td>
<td>Information System</td>
</tr>
<tr>
<td>PACS</td>
<td>Picture Archiving and Communication System</td>
</tr>
<tr>
<td>PIS</td>
<td>Patient Information System</td>
</tr>
<tr>
<td>QUIS</td>
<td>Questionnaire for Interaction Satisfaction</td>
</tr>
<tr>
<td>RFP</td>
<td>Request for Proposals (in procurement)</td>
</tr>
<tr>
<td>SUMI</td>
<td>Software Usability Measurement Inventory</td>
</tr>
<tr>
<td>SUS</td>
<td>System Usability Scale (Method)</td>
</tr>
<tr>
<td>UCD</td>
<td>User-Centred Design (Process)</td>
</tr>
<tr>
<td>UI</td>
<td>User Interface</td>
</tr>
<tr>
<td>USco</td>
<td>Usability Scoring (Model)</td>
</tr>
<tr>
<td>UX</td>
<td>User Experience</td>
</tr>
</tbody>
</table>
List of Publications

This doctoral dissertation consists of a summary and of the following original publications which are referred to in the text as Papers I-VI. The publications are reproduced with kind permission from the publishers.


VI Viitanen, Johanna; Tyllinen, Mari; Tynkkynen, Elina; Lääveri, Tinja. 2022. Usability of information systems: Experiences of outpatient physicians, outpatient nurses, and open care social welfare professionals from three large cross-sectional surveys in Finland. Elsevier. International Journal of Medical Informatics, volume 165, issue September 2022, article 104836. ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2022.104836.
Author’s Contribution

**Publication I** “A Framework for Usability Evaluation in EHR Procurement”

Tyllinen is the main author. Kaipio and Lääveri came up with the basis for the usability evaluation framework presented in the publication. Tyllinen and Kaipio planned the framework in detail, and the methods included in the framework. All authors contributed to the writing of the publication, while Tyllinen was responsible for finalizing the publication.

**Publication II** “We Need Numbers! - Heuristic Evaluation during Demonstrations (HED) for Measuring Usability in IT System Procurement”

Tyllinen is the main author. Kaipio and Lääveri planned the first version of the method, Tyllinen and Kaipio the second version presented in the publication. Tyllinen and Kaipio implemented the method in the case study and conducted data collection. Tyllinen planned and conducted the statistical data analysis for the publication. Tyllinen and Nieminen wrote the introduction together. Tyllinen wrote most of the publication. Kaipio and Lääveri provided comments and suggestions for writing and shortening the publication. Conclusions were written in collaboration with all the authors.

**Publication III** “End-Users’ Voice in EHR Selection: Development of a Usability Questionnaire for Demonstrations in Procurement (DPUQ)”

Tyllinen is the main author. Kaipio and Lääveri planned the first version of the questionnaire, Tyllinen and Kaipio prepared the second version (including scoring) presented in the publication together. Tyllinen and Kaipio conducted data collection. Tyllinen planned and conducted the statistical analysis and wrote most of the publication. Kaipio wrote the first version of background. Lääveri and Nieminen provided comments and suggestions for writing and shortening the publication.

**Publication IV** “Usability Analysis of Contending Electronic Health Record Systems”
Tyllinen is the main author. Tyllinen and Kaipio planned and implemented the usability testing procedure described in the publication. Tyllinen, Kaipio, and Lääveri planned the usability test scenarios and test tasks. Tyllinen planned and conducted the statistical analysis and wrote most of the publication. Kaipio and Lääveri provided comments and suggestions for writing and shortening the publication.

**Publication V** “UXtract – Extraction of Usability Test Results for Scoring Healthcare IT Systems in Procurement”

Tyllinen is the fifth author. Pitkänen is the first author and had main responsibility for the publication. Tyllinen and Kaipio together were responsible for the usability testing procedure described in the article, including planning, running, and reporting the tests as well as defining the scoring formulas. Pitkänen, Pitkäranta and Haapala were responsible for the technical solution described in the publication. Pitkänen was responsible for the statistical analyses for the publication. Tyllinen provided comments and suggestions on the whole publication.

**Publication VI** “Usability of information systems: Experiences of outpatient physicians, outpatient nurses, and open care social welfare professionals from three large cross-sectional surveys in Finland”

Tyllinen is the second author. Viitanen, Tyllinen and Lääveri had the original idea for the publication. Tyllinen, along with Viitanen and Lääveri, was one of the main contributors to plan and implement the social welfare professional survey study and Tynkkynen contributed to the design of the questionnaire. Tyllinen was responsible for applying ethical approval for the survey. Viitanen and Lääveri contributed to the design of the physician and nurse surveys. Lääveri was responsible for the statistical analysis for the publication. Viitanen had main responsibility for the publication. The whole publication was written together with all authors.
1. Introduction

Public healthcare systems are facing challenges of increasing patient-to-professional ratios with the aging of the population and more recently due to shortage of healthcare employees around the world (YLE News, 2022; Triggle, 2023). Implementing clinical information systems to support efficiency and reallocation of work has been seen as one solution to solve these issues (Finnish Government, 2019; Kraus et al., 2021; World Health Organization, 2021), and the significance of these systems is growing in this sector. However, the poor usability of information systems (ISs) used in the healthcare sector and negative experiences voiced by their users, especially doctors, is a problem. The usability issues have been discussed both in academic research as well as in other forums and media extensively during the last decade (e.g. Schumacher et al., 2010; Viitanen et al., 2011; Hallamaa, 2012; Kellogg, Fairbanks and Ratwani, 2017; Hyppönen et al., 2018; Lareau, 2020). All the above is true both internationally and in Finland, where this dissertation research has taken place.

In Europe, there are additional plans to integrate healthcare and social welfare services in several countries, including in Finland (Baxter et al., 2018; Keskimäki, Sinervo and Koivisto, 2018). The need for services in one is linked in many instances to the use of services in the other (Byrne et al., 2003; Vedsted and Christensen, 2005). “High utilizers” of these services also account for almost 80% of the costs and supporting the coordination of services with sharing of data in both sectors is seen as one solution (Malmström, Peltokorpi and Lappalainen, 2015). The healthcare and social welfare context is very varied and includes several medical specialties both in outpatient and inpatient as well as social welfare and social care services both in open care and institutional settings. In all these settings there are different types of information systems in use, such as general systems for mainly documenting patient and client data and more specialized systems used in for example operating rooms. The procurement of these ISs in the public sector is regulated by both the
European Union with Directive 2014/24/EU (European Union, 2014) and the related national legislation in each member state.

1.1 Motivations for the Dissertation

The poor usability of the systems used in healthcare has been shown to increase errors and decrease patient safety (Koppel et al., 2005; Kushniruk et al., 2005). It also influences the efficiency of use by definition. Furthermore, one of the most common usability problems of IS in healthcare, minimizing cognitive load for its users (Zahabi, Kaber and Swangnetr, 2015), has been linked to burnout (Iskander, 2019; Harry et al., 2021). In the social welfare sector this discussion on usability of IS has only been emerging in recent years. The situation does not seem to be much better than in healthcare (Ylönen et al., 2020), although it has not been studied much (Lagsten and Andersson, 2018). A similar thought process than in healthcare could be followed to view the consequences in social welfare. Corresponding to patient safety, poor usability in social welfare can have the effect of prolonging processing times and compromising the care of the client. Information systems should support the professionals using them by being a tool to manage and help in their everyday work. However, IS with poor usability is usually less a technological companion, and more a hindrance and nuisance that takes time away from patients and clients.

The solution for making IS more usable has long been seen to be applying user-centred design (UCD) and human factors engineering (HFE) during the development process. These practices in the field of health IS have been promoted and researched for at least two decades (e.g. Kushniruk and Patel, 2004; Peute and Jaspers, 2007; Belden, Grayson and Barnes, 2009). Early on, IS for healthcare was developed mainly in-house or as dedicated development projects for certain functionalities and contexts (Hammond, 1994), but the shift has been made to commercially available clinical solutions (Kuhn and Giuse, 2001) which organizations acquire predominantly as ready products that are further adapted and configured to fit the needs of the purchasing organization. While the early systems were technology-oriented in their design without much consideration for end-users, the change to understanding the relationship between the systems and user workflows happened in 1990s (Hammond, 1994; Kuhn and Giuse, 2001).

When information systems are not developed for in-house use, but acquired, there exists a contractual situation between a vendor and a procurer. This applies to both situations where new IS is acquired as a development project or
as a product. The distinction between “product” and “project” environments in software engineering was established by Fox in early 1980’s. The difference between the two is the former is built for the market for more than one client while the latter is custom-built for one client (Fox, 1984; according to Carey, 1991). In the development of these products, to ensure usability, it is important to incorporate the user-centred practices as mentioned above. However, the role of the purchasing organization has also been promoted (Jimison et al., 1999) and taking into account the quality aspects of the system during the selection of the purchased system was deemed necessary already in the 1990s (Carey, 1991; Scown, 1998).

Among these quality aspects is also usability. This surfaces the question of how can the purchasing organization include usability considerations into the procurement process? And furthermore, how can this be done within the highly regulated environment of public procurement and in the case of a complex information system, such as those in healthcare, with a multitude of use contexts, user groups and tasks?

In this dissertation, I research how to integrate usability evaluation into public IS procurement to affect the selection process. The research presented in the dissertation was done primarily in one procurement project in Finland. In this project the above questions were faced, in its early stages, in the years 2012-2013. The project was initiated to solve the problem of several municipalities and a hospital district in the capital region having hundreds of separate and outdated ISs in use in its healthcare and social welfare services. The procured system was meant for primary, secondary, and tertiary as well as specialized healthcare, and open and institutionalized care in social welfare.

More specifically, the topic of this dissertation is researching the suitable methods, process, and considerations for including usability measurement as a selection criterion in public IS procurement. This is accomplished through a study combining action research and design science research of a procurement project and an analysis of survey results on users’ experiences of their IS in healthcare and social welfare. This research was initiated by an identified real-world problem of comparing the usability of competing systems in the procurement project. No readily applicable solutions were identified to the problem from usability practice and the possible need to develop new methods and an evaluation procedure within the procurement surfaced. During the project, a need for insights comparing the experiences of different user groups in healthcare and social welfare was recognized. The survey study addresses this. Finally, based on the research this dissertation presents a process model
for including usability measurement of IS in the public procurement selection phase.

1.2 Research Environment

Digitalization has long been a running theme in organizing and developing public services and supporting the structural reforms in Finnish society (Ministry of Finance Finland, n.d.a). It is said that the number of public electronic services is one of the highest in the world, and in the European Union (EU) Finland is leading the eGovernment movement with top rankings for years in the overall digital economy and society index (European Commission, 2022). To understand the research environment for this dissertation work, I will present the following key areas: the organization of healthcare and social welfare in Finland, the utilization of and terminology used for IS in these sectors, definition and legislation for public procurement and the researched procurement project.

Healthcare and Social Welfare in Finland

The organization of healthcare and social welfare services is the responsibility of the public sector in Finland. Until the beginning of 2023, the responsibility was on the municipalities. The basic services were provided alone by the municipalities or by forming joint municipal authorities. The services could be purchased by the municipalities from other municipalities, organizations, or private companies. Specialized medical care is provided by hospital districts. The social services required to be provided by the public organizer include a wide range of services for families with children, adults, those with disabilities and older people (Ministry of Social Affairs and Health Finland, n.d.).

During the writing of this dissertation a health and social services reform has been underway in Finland. In the new model, the responsibility of organizing the health and social services were transferred from municipalities to the newly established wellbeing services counties, which are still public entities, beginning in January 2023 (Ministry of Finance Finland, n.d.b). This reform affects the organization, producing and funding of the services. There are 21 wellbeing services counties in Finland; in the metropolitan area four new counties and the city of Helsinki are the responsible organizers of the services. Additionally, the joint county authority for the Hospital District of Helsinki and Uusimaa is still responsible for the specialized healthcare services in the area.
Information Systems in Healthcare and Social Welfare

The utilization of IS in the healthcare context has been reality since the turn of 1950s to 1960s in some form, and in clinical contexts increasingly by the 1990s (Hammond, 1994). The situation was similar for Finland early on, but the digitalization in public healthcare was rapid and by 2007 health IS coverage had reached almost 100% in specialized care and by 2010 also in primary healthcare (Reponen et al., 2019). In social welfare, the emergence of information technology as a prevailing practice was noted in the early 2000s (Garett, 2005) although early systems have been in use for longer. However, use of IS in the field as an essential tool has reached a high percentage only in recent years in Finland (Kuusisto-Niemi, Hyppönen and Ruotsalainen, 2019).

The technology and information systems in use in the healthcare and social welfare sectors are extremely varied, from specialized systems used for a clearly defined purpose to those covering a wide range of contexts and tasks.

The terminology and definitions for these systems is also diverse. According to Rantanen and Heimo (2014) there is a lack of clear definitions for the terms and terminology both in Finnish and English, and the translations between the two are not established. They list problematic concepts to be patient information system (PIS; direct translation from the Finnish term ‘potilastietojärjestelmä’), electronic health record (EHR), electronic patient record (EPR), electronic medical record (EMR; closest term in Finnish is ‘potilaskertomusjärjestelmä’), health information system (HIS) and eHealth. Rantanen and Heimo (2014) discovered that even the main national actors (such as legislators, authorities and technology vendors) in Finland do not have a commonly shared definition for PIS (‘potilastietojärjestelmä’), the main term used in Finnish.

EHR and EPR are the commonly used terms for translation of the Finnish term PIS (Rantanen and Heimo, 2014). These terms, however, do not convey the same amount of breadth and purpose because of their use of the word ‘record’ instead of ‘information system’. Also internationally EHR and EMR are commonly used interchangeably even though they are completely different (Garets and Davis, 2005). Rantanen and Heimo (2014) define eHealth to include all electronic healthcare systems, and HIS as a subset of it that includes all information systems in healthcare sector. While EPRs, EHRs, EMRs and the Finnish PIS can all be considered HIS, they are not equal (Rantanen and Heimo, 2014). This discrepancy with some of the terms has also been identified by other researchers (Oh et al., 2005; Häyrinen, Saranto and Nykänen, 2008). The multitude of definitions makes discussion and research harder. Rantanen and Heimo (2014) argue that this problem contributes to a lack of
common understanding and confusion when discussing these systems, their procurement and implementation as well as challenges with their use, even on a national level.

Based on the general definition of information system, HIS can be defined as a system combining software and other technology, paper documentation, healthcare professionals and patients, their stored data, knowledge and communication (Rantanen and Heimo, 2014). And based on this, Rantanen and Heimo (2014) argue, that PIS “are not only technical problems nor solutions and they should not be viewed as such, but instead as a complex combination of workers and their tools flavoured and tied together with the communication between them”. They describe several ways through which to define PIS: functionalities, content, structure and examples and negations.

The technology and terminology in use in the social welfare sector is also varied. The Finnish term commonly used that is comparable to PIS in healthcare, can be directly translated to ‘client information system’. A similar problem with defining and translating the term is possible also in this sector than described previously. In social welfare literature, varied terms are used in English, such as electronic social care record (ESCR) (Wilson, Walsh and Vaughan, 2007), client information system (CIS), professional case management system (Jørgensen et al., 2022), management information system and client management system (Gillingham, 2016).

In the publications included in this dissertation (Papers I-VI), a variety of terms have been used either to describe the information system or technical parts of it, which is the object of procurement, or the other systems used by professionals. In this compilation part of the dissertation, I will primarily use the terms HIS and CIS to describe separate systems and health and social welfare IS to describe the procured system.

Public Procurement
“Public procurement refers to the acquisition (through buying or purchasing) of goods and services by government or public sector organizations” (Hommen and Rolfstam, 2009). When it comes to public procurement there are strict legal requirements that need to be followed, which results in a more complex process than in a private setting. In the EU there have been directives in place since the early 1990s (Moe, 2014), which the member states’ national legislation should follow. During the procurement project the Directive 2004/18/EC (European Union, 2004) and national legislation in Finland, Act 348/2007 (Finland, 2007), was followed. Currently, newer legislation is in
The legislation outlines different types of procurement processes and the detailed requirements that need to be followed by the purchasing organization. There are four types of procurement that are relevant for the acquisition of IS: open procedure, restricted procedure, negotiated procedure and competitive negotiated procedure (European Union, 2014; Finland, 2016). The legislation in place during the procurement project defines the same four types (Finland, 2007), with some differences in the details. The negotiated procedure was used in the procurement project (which is described in the next subsection). The principles are the same in both versions of legislation: The vendors must be treated equally and fairly. When following the negotiated procedure, for the selection the minimum requirements and criteria, along with weights, need to be published in the very beginning. The negotiations can be done in stages, limiting the number of vendors based on the predetermined criteria for selection (European Union, 2014; Finland, 2016).

The Researched Procurement Project

One of the purposes of the project was to procure and build the first integrated health and social welfare information system in the world for the several participating municipalities and hospital district in the Helsinki metropolitan area, in Finland. Prior to the project, there were hundreds of different electronic systems, which were largely not integrated, used for the same functions in the region. The goal was to reduce the number of systems in use and thus integrate information flow and simplify use. This purpose was part of a bigger goal of developing and unifying the service and care delivery processes in the area to combat the increasing costs. The procured system was visioned to not only act as an information storage for client and patient data, but also to guide the professionals in providing the care. The scope of the procured information system included functionalities for the professionals working in primary, secondary, and tertiary as well as specialized healthcare, and open and institutionalized care in social welfare. The health and service web portal for citizens was also included. In the procurement project, the computerized parts of the information system were divided into core and integrated systems, so the definition was through structure. Core systems, which were the object of procurement, include for example client and patient records and the information on treatments and services (Rantanen and Heimo, 2014).

The project was initiated in 2012 and ended with the last major go-lives of the procured system in 2021. The project has defined the phases to have been
the following: planning (2012-2013), procurement (2013-2016), definition (2013-2019), implementation (2016-2020) and deployments (2018-). In the project, both quality and price of the offered solutions affected the selection. During the procurement, the project utilized negotiations to determine the exact requirements and product comparisons to first determine that minimum quality for the products was met and then influence the final selection. In the final selection, the offered price and different quality criteria were weighted with predetermined rules.

1.3 Key Concepts and Domain of the Research

In the following, I briefly introduce key concepts and research domains that are relevant for this dissertation work. These include the public IS procurement process and its key terminology, the concept of usability, its evaluation, and their use in relation to health and social welfare IS. Also, utilizing usability evaluation in the selection phase of procurement is introduced.

Public IS Procurement

The public procurement process has not been academically researched nearly as much as the private process (Bergman and Lundberg, 2013). Moe (2014) defines a public IS procurement project to include the following phases: development of requirements specification, tendering, selection (and negotiations), contracting, implementation and completion. The selection phase can include negotiations on price, training, implementation schedule and scope. The selection of the vendor can be based on price alone or a combination of price and quality (Moe, 2014). This definition of the phases differs from those used by the researched procurement project in this dissertation. Using the definition above the whole duration (2012-2021) of the project was procurement. Tendering, selection and contracting lasted from 2013 to 2016. And implementation and completion from 2016 to 2021.

Key concepts essential to public procurement that should be understood are: “supplier selection method”, “minimum quality standards”, “scoring rule”, “weighting function” (Bergman and Lundberg, 2013) and “tender evaluation model” (Mateus, Ferreira and Carreira, 2010). In the EU the “supplier selection method” is the economically most advantageous tender (EMAT) which during the old legislation meant price and quality combined (European Union, 2004; Finland, 2007) but could also mean price alone according to the terminology in the newer legislation (European Union, 2014; Finland, 2016). The “scoring rule” is a function that sets numeric values to different levels of
quality or converts a measured value on one scale to another scale (price to quality or quality to price). The “weighting function” then combines these different values to one single score with different weights, but it also refers to combining different quality scores to one overall score (Bergman and Lundberg, 2013). The “tender evaluation model” then describes all the above in detail, but it is not mandated to be published beforehand by EU regulations (Mateus, Ferreira and Carreira, 2010).

The Definition of Usability and Usability Evaluation
There are several ways that usability is defined in literature. Two commonly used definitions are by the International Organization for Standardization (ISO, 2018) and by Jakob Nielsen (Nielsen, 1993). In these definitions usability has several attributes: effectiveness (ISO, 2018), efficiency, satisfaction (Nielsen, 1993; ISO, 2018), learnability, memorability and lack of errors (Nielsen, 1993). A definition of usability by McGee, Rich and Dumas (2004) addresses usability from a user’s perspective: “Usability is your perception of how consistent, efficient, productive, organized, easy to use, intuitive, and straightforward it is to accomplish tasks within a system”. Interestingly, their study proposes that while efficient, intuitive and effective are usability components - satisfaction is not a usability characteristic at all. While they are moderately related, they are not integral to usability (McGee, Rich and Dumas, 2004). The context of use is an integral concept related to usability. The ISO standard defines users, their goals and tasks as well as resources and the environment to be the components for context of use (ISO, 2018). The ISO standard highlights, that these are important aspects to consider when trying to assess the usability attributes. Usability of an IS is the outcome of use in the context of use (ISO, 2018).

Hertzum (2010) has proposed to any one definition of usability can only offer a partial explanation to the concept and there are actually six different “images of usability” that are partially overlapping and complementary. These six images are “universal usability”, “situational usability”, “perceived usability”, “hedonic usability”, “organizational usability” and “cultural usability”. These different perspectives view usability through different lenses. The ISO (2018) and Nielsen (1993) definitions of usability use the lens of situational usability (Hertzum, 2010) while the McGee, Rich and Dumas (2004) definition has that of perceived usability. Hedonic usability focuses on the role of positive emotions, organizational usability incorporates the organizational setting and collaborative aspects and cultural usability emphasizes the relevance of culture on usability beyond merely user interface elements (Hertzum,
The different images have differing relation to the different facets of usability: objective and perceived, process and outcome, performance and pleasure, individual and collaborative use, and short-term and long-term use (Hertzum, 2010).

Designing IS with excellent usability and developing usability evaluation methodology has been studied since the 1980’s in the fields of human-computer interaction (HCI) and human factors and ergonomics (HF&E) (for example, see Nielsen, 1993; Dumas and Salzman, 2006). As an outcome of this work several established methods for usability evaluation exist, such as heuristic evaluation and usability testing. Usability evaluation is used primarily to gather qualitative data during iterative product development, but also quantitatively when benchmarking, comparing design versions and validating the end result (Sauro and Lewis, 2012). The context of evaluation where the purpose is to develop the product further is referred to as formative evaluation and assessing the level of usability at the end of the product development is referred to as summative evaluation (Nielsen, 1993; Hartson, Andre and Williges, 2003).

Usability of Health & Social Welfare IS

In the United States, the National Institute of Standards and Technology has provided guidance for vendors on how to utilize these practices during EHR development (Schumacher and Lowry, 2010; Lowry et al., 2015). However, usability of healthcare IS still remains unsatisfactory (Kaipio et al., 2017; Ratwani et al., 2018), the efforts by vendors have only improved in the recent few years (Hettinger, Melnick and Ratwani, 2021) and the majority of evaluation studies published seem to be focused on implementation or post-implementation phase (Ellsworth et al., 2017). Healthcare has also been identified as a context for complex information systems (Mirel, 2004), and the lack of research on usability evaluation methods for these types of systems has been pointed out (Redish, 2007; Chilana, Wobbrock and Ko, 2010).

In the field of social welfare research, the role, meaning and use of information and communications technology (ICT), including software, is viewed more from a socio-technical point of view (Jørgensen et al., 2022). However, there is some research, especially during the last decade, on users’ experiences with IS (for example, see Gillingham, 2016; Salovaara and Ylönen, 2021), user participation in the design (Martikainen et al., 2021) and collaborative design of these tools (Gillingham, 2014). However, consistent efforts to evaluate usability and apply user-centred practices are not reported in literature the same
way as in healthcare, and this has been identified as a future need for research (Lagsten and Andersson, 2018).

**Usability Evaluation in the Selection Phase of IS Procurement**

The first possibility for an organization taking already existing IS into use to affect its usability is during the procurement (Lif, Göransson and Sandbäck, 2005). Research has focused on how to define usability, user and process requirements in procurement (Forsgren and Rahkonen, 1995; Lauesen, 1998; Lauesen and Younessi, 1998; Artman, 2002; Bevan *et al.*, 2002; Lehtonen *et al.*, 2010). User input has been identified as an essential part of successful procurement (Howcroft and Light, 2002; Artman and Markensten, 2005), while also emphasizing the need for more research on this topic Moe (2014).

The focus in previous research of the overall procurement process has been on the tendering and especially selection phases (Moe, 2014). The research has been mostly theoretical on the decision criteria and few papers actually seem to rely on empirical data (Moe, 2014). The selection phase is also relevant from a usability point of view: it can include conducting usability evaluations on the candidate systems and including the results in the selection criteria.

Including such evaluations in the procurement of IS have been reported in the healthcare sector in the last two decades (Beuscart-Zéphir *et al.*, 2002, 2005; Corrao *et al.*, 2010; Jensen, Rasmussen and Lyng, 2014; Jorritsma, CNossen and van Ooijen, 2014), but also in other contexts of use (Larson, Hocko and Bye, 2010; Riihiaho *et al.*, 2015). Suggestions on how to select suitable methods (Kushniruk *et al.*, 2010) and how to organize the procedure in stages (Schumacher, Webb and Johnson, 2009) also exist. However, in many cases the methodologies used are described separately from the whole and they are not detailed on how the measurements on usability are combined and compared. Researchers have argued that there is a need for more knowledge on how to include quality aspects, such as usability, into the procurement process of any IS, including also HIS (Holbrook *et al.*, 2003). Moreover, the role of usability and methods and metrics for evaluating and comparing usability during procurement of IS should be researched further, particularly for health (and social welfare) IS (Beuscart-Zéphir *et al.*, 2002, 2005; Jensen, Rasmussen and Lyng, 2013) but also in general.

As introduced earlier, there are many established methods for usability evaluation during product development. However, these methods are primarily for formative purposes and the procurement selection context requires a summative lens. There are still many issues and problems related to measuring
usability (Hornbæk, 2006; Gillan and Bias, 2014). These challenges are even further complicated by the demands and constraints of a regulated public IS procurement process. Indeed, few of the studies on usability evaluation during selection have focused on the regulated public procurement context.

1.4 Objectives and Scope

This dissertation focuses on the following research problem:

*How to measure and compare usability during the selection of a client and patient information system for public social welfare and healthcare?*

Based on the considerations and restrictions that come from the law governing public procurement and the goals of implementing usability evaluation as part of the selection as presented earlier in sections 1.2 and 1.3, the aspects of procurement that are relevant for the research problem are:

- the focus is on comparing the systems and not identifying their problems,
- there is a need to quantify the entire system’s usability instead of isolated use cases,
- the evaluation process needs to be clear and equal,
- the criteria for measurement need to be predetermined in detail,
- there are possibly several candidates to evaluate in a limited timeframe, and
- the systems that are evaluated are not available during the planning of the procedure.

There is a need to understand how these aspects together with a complex domain combining both healthcare and social welfare, with numerous different use contexts, goals, user groups and highly specialized user tasks can be covered during the usability evaluation. Figure 1 (next page) visualizes the scope of this dissertation.
This dissertation has three main objectives:

1. First, to understand the requirements of a public procurement process, the needs for the planned evaluation procedure within the selection phase, and how these affect the planning and execution of the summative usability evaluation.

2. Second, to plan and execute a comparison procedure of usability of several complex systems so that
   a. it covers the IS, use context, user groups and their tasks and goals, as well as usability attributes sufficiently, and
   b. produces numeric results that are comparable and can be utilized for the selection.

3. Third, to validate the procedure and methods that were developed during a real-life procurement project.

Other selection criteria for usability besides evaluation of prospective systems exist, such as scoring vendors agreement to meeting usability requirements (Lauesen, 1998), evaluating the development process (Jimison et al., 1999) and assessing the usability expertise of the vendors. The inclusion of usability as a selection criterion during procurement is also not enough, as the process of adapting and configuring the selected system before taking it in use will have a significant effect on the end result, therefore also on its usability (Møller-Jensen, Lund Pedersen and Simonsen, 2006; Lee, Williams and Sheikh, 2016; Ratwani et al., 2016). This process can be guided by including
usability aspects also in the request for proposals (RFP) and the contract between the vendor and customer organizations. There are also stakeholders, such as the organization within which the IS is used, that affect the use of IS and the experiences of end-users. These aspects are to be considered when implementing and configuring the selected IS. The majority of previous evaluation studies in the healthcare context seem to focus on these latter parts of the process (Ellsworth et al., 2017). While these are all important parts of the overall procurement process and ensuring usability of the procured IS, they are outside the scope of this dissertation.

A closely related concept to usability, is user experience (UX). UX, to this date, does not have a commonly shared definition within the HCI community despite attempts at defining it for over a decade (Law et al., 2009; Mirnig et al., 2015; Gómez-López, Simarro and Bonal, 2019). Some definitions include usability or some aspects of it within UX or as a precondition to it, and a shared aspect of the definitions seems to be the focus on the user’s experiences, emotions and inner state (Law et al., 2009). While the concept of UX relates to the use of IS and user’s perceptions of it, evaluating UX during IS procurement is not within the scope of this dissertation.

1.5 Research Questions

The research questions that aim to answer the research problem “How to measure and compare usability during the selection of a client and patient information system for public social welfare and healthcare?” presented in section 1.4 Objectives and Scope are as follows:

- **RQ1**: What are effective evaluation methods and metrics for measuring usability during the selection phase in public procurement of IS?
- **RQ2**: How to construct the evaluation procedure supporting the quantification and comparison of usability for this use?

The research questions are defined in detail in the following.

**RQ1**: What are effective evaluation methods and metrics for measuring usability during the selection phase in public procurement of IS?

The first research question addresses the issue of understanding the methodological perspective: first, what is the range of suitable usability evaluation methods and metrics for measuring it presented in HCI literature; second, does the comparison aspect have an effect on method selection; third, what
methods and metrics have been previously used in procurement selection and why; fourth, what demands does the public procurement legislation present on the applicability of these methods and metrics; and finally, to evaluate the usability of complex IS during procurement is there a need to modify existing methods or develop new ones, and if so how and what is needed.

**RQ2: How to construct the evaluation procedure supporting the quantification and comparison of usability for this use?**

The second research question addresses the practical issue of planning the entire evaluation procedure for a public procurement project’s selection phase: first, what is the overall structure of the procedure and what are its components; second, what details are needed to be planned from a legislative perspective; and finally, how to quantify and score the usability of vendors’ solutions to support the selection phase.

The research to answer these questions aims to both explore the existing literature and theory, and to construct and apply suitable methods and a usability evaluation framework as part of a real-life procurement project. The existing theory on usability evaluation, including usability evaluation in procurement, on public procurement and quantifying qualitative criteria for public procurement as well as quantifying usability is explored. The research design follows the action design research (ADR) approach.

An additional theme related to these research questions is, how the contexts of use (user groups and their tasks and goals as well as the environment) affect the construction and planning of the usability evaluation procedure for this use? This theme addresses the issue of understanding the context into which the information system is being procured and how these contextual aspects of usability can be addressed appropriately in the usability evaluation procedure, and how they might affect the decisions for the procedure. Aspects related to this theme are presented when found in the related literature on including usability evaluation in the selection phase of procurement and when presenting the research results of this dissertation. Additional usability focused survey study results outside the procurement context are included in the dissertation. This study contributes to the understanding of use contexts and various end user groups needed prior to procurement. The selected research approaches and reasoning behind them as well as the study design are presented in Chapter 3.
1.6 Anticipated Contributions of the Dissertation

This dissertation contributes by both offering practical solutions to the research problem and conducting scientific research on a public IS procurement project that includes usability evaluation in the selection phase. According to previous research, the public procurement process is understudied overall from a practical perspective and a need for developing and studying applicable usability evaluation methodology exists. The research presented in this dissertation has been done within the healthcare and social welfare context.

The most central contributions of the dissertation are developing and validating two new evaluation methods for measuring usability for the selection phase in procurement, presenting the metrics and scoring rules used with usability testing and presenting an overall methodological framework for usability evaluation in this context. The dissertation also creates knowledge on how measuring usability fits into the public procurement process and what considerations are needed for scoring usability in this context. This is examined in more detail for the two developed methods and usability testing.

The practical contributions of this dissertation are providing procuring organizations, especially within the public sector, practical tools and knowledge on how to include usability evaluation as one of the selection criteria. This includes information on what methods to use, how the evaluation procedure should be planned, what aspects should be considered while planning and what kind of expertise is needed for the planning and execution.

Evaluating usability as one of the selection criteria when procuring an existing information system can be considered essential if the system has end users. However, it is not straightforward, especially, when a complex information system with numerous user groups is in question. This dissertation also proposes a model for defining and planning the usability scoring process in practice. The model benefits both practitioners and procuring organizations in understanding what is involved in the process.

1.7 Dissertation Structure

This dissertation is a compilation thesis, composed of a summary including six chapters and six individual research articles. Following this introductory chapter that outlines the motivation for this dissertation, related background, objectives, research problem and research process, the structure of this summary is as follows.
Chapter 2 presents related research on the topics of public IS procurement, including usability in IS procurement and especially the selection phase, including usability evaluation in the procurement of HIS as well as general literature on usability evaluation methodology and measuring usability.

Chapter 3 presents the research paradigm and approach. After that, it describes the research context and design in the conducted research. The iterative method development as well as the process of implementing and evaluating the developed methods during a real-life procurement project are presented. Also, relevant quality criteria for the research are introduced.

Chapter 4 presents the main results of the research work by organizing the research in the papers under themes. All methods, metrics and measures are introduced, the method development and final version of three methods, quantifying usability with the methods, method implementation and evaluation, and construction of the whole evaluation procedure for scoring usability including considerations of use context are presented.

Chapter 5 is the discussion in which the research questions are answered, a new model for defining and planning the usability scoring process is presented, the implications, as well as the research quality are evaluated, and limitations of the research presented. Finally, recommendations for future research are discussed.

In Chapter 6 the answers to research questions and the model are summarized. Finally, conclusions and practical recommendations based on the research are given.
2. Theoretical Foundation

This chapter presents the core literature related to the topic of this dissertation. The research presented here is based on literature presented in the research papers included in this dissertation but extends beyond this by going into more detail. In the context of this dissertation three tracks of related research are of interest: context of public information system (IS) procurement; usability evaluation in IS procurement; and quantitative usability evaluation.

The chapter starts by introducing the topic of public IS procurement, and the most important specifics related to the topic of this dissertation in public procurement processes in the European Union and Finnish legislation are presented. In the following section, previous research on including usability, and especially usability evaluation, in the context of IS procurement is introduced. Also, the research on evaluating usability during selection in the domain of healthcare and social welfare IS is presented. Next, research on usability evaluation methods, specifically measuring usability including metrics and combining them, as well as developing usability evaluation methods is presented. Finally, the key takeaways from this research are discussed.

2.1 Public Information System (IS) Procurement

Public procurement, also referred to as governmental procurement, is the process around acquiring new products or services by a public, or governmental, organization (Hommen and Rolfstam, 2009). Compared to purchasing in the private sector, public procurement is strictly governed with regulation. Based on literature, Moe (2014) defines public procurement to include everything from the requirements elicitation and formulation, the actual purchasing (including tendering, selection, contract signing), implementation of the selected product or service to validating it meets the requirements. Also, other aspects,
for example management of different stakeholder views is included in his definition.

The purchasing stage of public procurement is very regulated in the European Union and has been since the early 90s (Moe, 2014). This stage of public procurement which includes the selection phase is of particular interest when it comes to the topic of this dissertation (see Figure 1 in section 1.4) and will be presented in the next section.

2.1.1 Requirements for the Purchasing Process

The EU directive regarding public procurement was updated most recently in 2014 with Directive 2014/24/EU replacing the previous directive, Directive 2004/18/EC (European Union, 2004, 2014). In all member states there is national legislation in place that is in alignment with the directive. As is relevant for the context of this dissertation research, in Finland, the previous act, Act 348/2007, was replaced by newer legislation, Act 1397/2016 (Finland, 2007, 2016). At the time of the research done in this dissertation the previous directive and legislation was in place.

This regulation makes the process far more complicated than acquisition decisions in the private sector. In the tendering phase (start of the purchasing stage as described above) the purchasing organization announces the procurement and invites vendors to submit their interest within a certain timeline (European Union, 2004, 2014; Finland, 2007, 2016). The selection phase starts when the vendors’ bids are received. Depending on the type of procurement this phase can include negotiations on some aspects of the procurement and evaluating the bids to reduce the number of vendors (European Union, 2004, 2014; Finland, 2007, 2016; Moe, 2014). The selection of the winning vendor is based on either price or price and quality combined, these are referred to as determining the “economically most advantageous tender” (EMAT) (European Union, 2014; Finland, 2016). After signing the contract there is a specified time for appealing the decision (Moe, 2014). The complex regulation creates the risk in public procurement of the losing vendor(s) appealing to the market court because of obscurities in the process, even if the procurer has tried to follow the process according to the best knowledge that is available to them (Alanne et al., 2015). Thus, understanding and following the legislation is integral to the success of the procurement.

Special conditions related to requirements in public procurements universally identified by Carey (1991) are: “in ideal public procurements, special care must be taken to avoid requirements which discriminate for or against
potential contractors”, and “there is also less flexibility for amendment to a contract, in order to help ensure that the original tendering process was fair to all concerned”. Equality, fairness and declaring criteria upfront are indeed three principles that are essential to public procurement. These principles show in the regulation in different ways which will be discussed in the following.

Four types of procurement specified in the EU directive are most relevant when it comes to purchasing already existing software: open procedure, restricted procedure, competitive procedure with negotiation and competitive dialogue (European Union, 2014). In the previous directive a similar procedure than competitive procedure with negotiation was called negotiated procedure (European Union, 2004). The regulations define the following types of requirements for the process (European Union, 2004, 2014; Finland, 2007, 2016): In order for the process to be fair and ensure competition, a minimum numbers of vendors is required to be included in the different stages of procurement. In an open procedure, all requirements and needs are to be stated upfront, and vendors submit their final proposals based on the combined contract notice and call for tender. In the other three types of procurement, vendors request to participate in the process based on a contract notice.

The contents of the contract notice differ based on the type of procurement (European Union, 2004, 2014; Finland, 2007, 2016): For the **restricted procedure and the competitive procedure with negotiation/negotiated procedure**, minimum suitability requirements and all evaluation criteria are both declared in the contract notice. A call for tenders is issued and tenders requested when vendors are selected based on the requests to participate. In the competitive procedure with negotiation/negotiated procedure this tender is preliminary and basis for the following negotiations. The call for tenders must include the description of needs, minimum requirements, and criteria for the selection. The procurer can modify the call for tenders or invitation to negotiate once negotiations have started. (However, the minimum requirements and criteria for the selection cannot be changed.) This means that negotiations can proceed in stages, and number of vendors can be reduced. The final tenders are not negotiable. For the **competitive dialogue**, the contract notice includes specifying needs and requirements as well as comparison criteria. However, all requirements of the procurement can still be negotiated. The purpose of the negotiations is to determine the best way to fulfill the requirements with the vendors. The negotiations can include stages, but only to limit the number of ways of fulfilling the requirements based on the comparison criteria. The final tenders can still be clarified with the vendors, and once
the best candidate is selected the tender can be further negotiated. Equal treatment of all vendors must still be guaranteed.

The regulations (European Union, 2004, 2014; Finland, 2007, 2016) also specify the criteria based on which the contracts can be awarded: In the competitive dialogue to determine EMAT, price-quality evaluation criteria must be selected. In the three other types of procurement procedure, it is an option, together with lowest price or most affordable cost. The EU regulation (European Union, 2014) also permits restricting the use of price only even further in national legislation. According to the regulations, the criteria for assessing EMAT must be measurable, related to the object of the contract, and comply with the principles of transparency, non-discrimination, and equal treatment. The comparison criteria and relative weights for each criterion must be specified in advance. However, it should be noted that neither of the EU regulations specify that the actual measurements and scoring rules must be declared with the comparison criteria.

Using those procurement types that include negotiations, competitive procedure with negotiation/negotiated procedure and competitive dialogue, before selection seem to be suggested as a best process in general and these types might be most suitable for IS procurement, specifically (Moe, 2014). The specific considerations related to EMAT in research literature will be presented in the next subsection.

2.1.2 Weighting Functions and Scoring Rules with EMAT

The EMAT principle was widely used in EU already in early 2000s (Verdeaux, 2003). According to researchers (Bergman and Lundberg, 2013) the previous EU directives were looser in requiring the definition of supplier selection methods and scoring rules in advance. However, the award of public contracts in the EU should be done with equal treatment and non-discrimination of vendors as well as with transparency, among other principles (European Union, 2004, 2014). The EU directive states that for equal treatment to be possible, the criteria used should be such that the tenders can be compared objectively.

Also, to ensure transparency, the vendors participating should be aware of the criteria and arrangements with which this comparison is done. Thus, already the previous directive (2004/18/EC) required publishing the criteria and the weights in advance. Researchers, however, argue that this requirement still permitted the procurer to have unrestricted freedom of choice in practice, and did not make the comparison objective (Mateus, Ferreira and Carreira, 2010). The newer EU directive (2014/24/EU) specifies that it should be ensured that
the award criteria does not permit such unrestricted freedom and must include “specifications that allow the information provided by the tenderers to be effectively verified in order to assess how well the tenders meet the award criteria” (European Union, 2014).

To conquer these challenges, the legislation in place for public procurement already during the previous EU directive in Portugal, an EU member state, required procurers to publish the entire tender evaluation model in advance (Mateus, Ferreira and Carreira, 2010). According to Mateus et al. (2010) this included the criteria, their weights and the scoring rules for each criterion which must be defined either with a mathematical function or with an ordered set of performance levels. It was also a requirement that the scoring rules were not dependent on features of other tenders (Mateus, Ferreira and Carreira, 2010), a practice that is used sometimes by applying relative scoring functions that are dependent on for example the highest or average value among tenders for a particular criterion (Dini, Pacini and Valletti, 2006; Ho Chen, 2008; according to Mateus, Ferreira and Carreira, 2010).

**Importance of Scoring Rules in Defining Weights**

Using EMAT that utilizes both price and quality makes public procurement more efficient, but also makes the procedure more complicated. In practice, this complexity often translates to scoring rules that are defective (Bergman and Lundberg, 2013). Researchers (Mateus, Ferreira and Carreira, 2010) explain why this is by the following reasoning: Requiring the evaluation criteria and their weights to be published beforehand in public procurement is not enough, because without actually determining the scoring rules for each criterion first, these weights will be arbitrary and most likely not reflect the purchaser’s real values accurately. According to Mateus et al. (2010) this is both due to the common misconception of assigning these weights simply by comparing the assumed relative importance of the different criteria, and the fact that the accurate determination of correct weights requires mathematical analysis with the actual scoring rules. They state that the former faulty approach seems “common sense” but is in fact erroneous and the result is not aligned with the actual values of the procurer. According to them this is the most common mistake procurers make with the weighting procedures (Mateus, Ferreira and Carreira, 2010).

**Designing the Overall Scoring Rule**

Determining how the tenders will be evaluated and designing the scoring rules appropriately is essential in public procurement especially because the selected vendor should be the one receiving the highest score (Bergman and
Lundberg, 2013). Bergman and Lundberg (2013) categorize scoring rules into three types: those that only consider quality, those that translate price to quality (numerical value) and those that translate quality to price (monetary value). This means a procurer can select the tender purely on quality with the price fixed or the tender that has the highest overall score with the price and quality combined. A scoring rule is needed for comparing the different offers, when either there is uncertainty about the cost and “both the cost and the value of quality vary significantly with the level of quality”, or there is only one version of the product available from each vendor (Bergman and Lundberg, 2013).

Designing a scoring rule means that an equation is needed that produces one score from a combination of price and one or more quality criteria. With a quality-to-price scoring rule the difference in quality is evaluated, and the quality-adjusted lowest price will be selected (Bergman and Lundberg, 2013). With a price-to-quality scoring rule the price is transformed to be comparable with quality, and the price-adjusted highest-quality will be selected (Bergman and Lundberg, 2013). While Bergman and Lundberg (2013) argue that using quality-to-price scoring would be preferable, according to them using price-to-quality scoring is more common.

The weighted sum is the most used method for combining the results from different criteria into a score in public procurement when price-to-quality scoring is used (Pictet and Bollinger, 2003; according to Mateus, Ferreira and Carreira, 2010). This method is even required by the Portuguese legislation (Mateus, Ferreira and Carreira, 2010), while no such requirement exists in Finland. Mateus et al. (2010) discuss in detail how the quality criteria, weights and scoring rules should be identified and defined based on their own extensive involvement with drafting the Portuguese legislation. The key points are presented in the following.

**Defining and Selecting the Criteria and their Measures**

When all aspects that are aligned with the objectives of the procurement have been identified as evaluation criteria, some aspects can be added as e.g. minimum quality standards (Mateus, Ferreira and Carreira, 2010). Acceptable evaluation criteria should have the following properties: “isolability (mutual independence preference), non-redundancy, completeness, conciseness, operationality and measurability, non-ambiguity and consensuality” (Keeney, 1992; Dogson et al., 2000; according to Mateus, Ferreira and Carreira, 2010). Dividing an identified criterion into sub criteria is sometimes also necessary to make distinct aspects of the criterion measurable. Then a performance
measure for each criterion must be determined, which describes and clarifies
the meaning of the criterion as objectively as possible (Mateus, Ferreira and
Carreira, 2010). It is essential that all performance measures can be measured
from the tenders, and thus the information required from the vendors must
be defined in detail (Mateus, Ferreira and Carreira, 2010). What contents to
include and how the information should be presented are to be defined, to
ensure comparability between tenders (Mateus, Ferreira and Carreira, 2010).

Mateus et al. (2010) state that either quantitative (numbers only) or qualit-
itive (verbal description for each level) scales can be used with performance
measures. The proper scale depends on the characteristics of the measure. In
case of a qualitative scale the levels should be defined as objectively as possible
and include at least two aspects, in order to eliminate as much subjectivity as
possible from the analysis (Mateus, Ferreira and Carreira, 2010). Performance
measures can be divided into three different types (Keeney, 1992; according
to Mateus, Ferreira and Carreira, 2010): natural (levels directly reflect the ef-
ect/consequence or end), proxy ("levels mostly reflect causes or means") and
constructed ("levels defined through holistic combination of plausible tenders
features"). Natural performance measures are to be preferred for their objec-
tivity according to Mateus et al. (2010).

Designing the Scoring Rule for a Criterion

Mateus et al. (2010) describe that the purpose of scoring functions is to con-
vert the measures collected in various units into partial scores that are on a
numeric scale and describe the level of performance on that measure. They
specify that: 1) Set levels of performance should reflect how desired each level
is to the procurer and how well it helps to achieve the procurer’s objectives.
The numeric difference between the scores of any two levels is not necessarily
the same if the improvement between any two levels is not valued the same.
2) Scoring functions should be absolute and not depend on another tender’s
performance (Mateus, Ferreira and Carreira, 2010).

Defining the scoring functions depends on which type of performance meas-
ure is in question (natural, proxy or constructed) (Mateus, Ferreira and Car-
reira, 2010). The scoring functions when using constructed measures are typ-
ically defined with continuous or discrete mathematical expressions while
natural/proxy measure scores are directly rated to predefined levels. Numer-
ous methods for constructing these scoring functions exist. For a continuous
performance measure, one suitable method is the bisection method (Goodwin
and Wright, 1997) which Mateus et al. (2010) use to go through an example of
determining a scoring function. Using the method results in a linear, piecewise
linear or exponential function depending on how the differences between performance levels are valued by the procuring organization (Mateus, Ferreira and Carreira, 2010).

In the case of qualitative or discrete measures, the scoring functions will not cover all possible performance levels. “Direct rating” (Von Winterfeldt and Edwards, 1986) is the most common method to determine these ratings (Mateus, Ferreira and Carreira, 2010). Mateus et al. (2010) give an example using this method, where the performance levels are ranked and given scores based on their desirability.

With the first example described above, it is shown that there can be different types of scoring functions for continuous measures. The key is to use a few concrete performance levels to determine suitable scores with all necessary experts by comparing their desirability and then use mathematics with the rest (Mateus, Ferreira and Carreira, 2010). With the latter example, it is important to understand that the decided starting scores are arbitrary and the differences in scores between levels is what can be compared - not the actual scores (Mateus, Ferreira and Carreira, 2010). These both examples illustrate how there are things that should be considered and procedures to follow that help in determining scoring functions, where common sense might fail.

**Determining Weights**

When using the weighted sum method for determining the overall score, the defined weights imply a trade-off relationship between the criteria (Mateus, Ferreira and Carreira, 2010). That is, the performance on the criteria compensates for each other. This trade-off can be seen by calculating what the loss of actual points and thus overall points on one criterion mean in actual points on another criterion (Mateus, Ferreira and Carreira, 2010).

According to Mateus et al. (2010) the weights should be determined by considering the trade-off relationships of the criteria relating to the procurer’s desires. This can be achieved by posing specific questions to those deciding about the procurement and requiring them to establish which trade-offs on the actual scoring units of the performance measures they are willing to make. This requires comparing two actual plausible examples for tenders, with significantly different performance levels on each criterion (Mateus, Ferreira and Carreira, 2010).

There are several methods available for composing technically sound weighting, of which Mateus et al. (2010) highlight with an example the trade-off procedure (Keeney and Raiffa, 1976) when two criteria are used. Utilizing this procedure means solving a linear equation system by using the scoring
functions and several fictitious tenders (as many as there are criteria) with different performance levels on each criterion (Mateus, Ferreira and Carreira, 2010). As can be imagined, the more criteria, the more complex the linear equation system.

Mateus et al. (2010) argue that in addition to needing to determine the scoring rules first for the procedure to be sound, the comprehensive tender evaluation model should be published in advance because this would benefit both parties. The procurer would get offers that better suit their values and the vendors would be able to evaluate what internal trade-offs they should consider to make the best possible tender they are willing to offer (Mateus, Ferreira and Carreira, 2010).

2.1.3 Challenges Identified in Public IS Procurement

Based on literature, there are a multitude of challenges and gaps in knowledge related to public procurement of IS. These challenges and knowledge gaps are found in all phases of public procurement (Moe, 2014). The challenges seem to stem mostly from legal requirements, complexity of information system environments in the public sector and resourcing of procuring organizations. From the research perspective, a need for more empirical data on actual procurement projects has been identified (Moe, 2014).

Overall, the purchasing phase is costly to the procurer (Moe, 2014), and requires knowledge from numerous different areas. However, public organizations seem to lack in-house experience on procurement and they do not have enough knowledge on IS or technologies (Moe, Risvand and Sein, 2006).

Public organisations typically have a complex environment where the procurement needs to take place. The existing IS environment is usually technically challenging with varied existing systems and a need to integrate the new system to them (Moe and Päivärinta, 2013). This environment often includes multiple stakeholders with each having their own goals and needs for the procurement that are often conflicting (Boyne, 2002; Flak, Nordheim and Munkvold, 2008; Moe, 2014) and unique to each project (Alanne et al., 2015). With a joint procurement project, the complexity of the procurement and contractual decisions have been found to affect the customization phase and resulting IS usability (Lee, Williams and Sheikh, 2016). Low user involvement in public procurements might also be a problem, despite this being identified as a success factor (Moe, 2014).

The above complexity makes the specification phase demanding, but unfortunately this phase seems to often lack resources (Moe, Risvand and Sein,
The law demands specifying the requirements (at least to some extent) before the purchasing process is started, however requirements always evolve as understanding of the project grows (Moe, 2014). It has been noted, that non-functional requirements, such as usability, typically suffer from being indistinct (Moe, 2014), while technical and functional requirements, can be too restrictive and system specific (Moe, Risvand and Sein, 2006). This is problematic, as the possible lack of expertise, both in understanding the available options (Alanne et al., 2015) and expressing the technical requirements (Lemmetti and Pekkola, 2012) can result in too rigid requirements that can also limit vendor interest in participating in the first place (Moe, Risvand and Sein, 2006). Furthermore, due to the legal requirement of publishing the selection criteria and weights in advance, and the resulting need to define the scoring rules before this as evidenced by literature (Mateus, Ferreira and Carreira, 2010), these all should also be established already during the specification phase (Moe, 2014).

However, the definition of criteria, weights and scoring is also full of challenges. Researchers (Bergman and Lundberg, 2013) have argued, that designing a good price-to-quality scoring rule is more difficult than a quality-to-price scoring rule, and in practice they are often faulty. Often, the resulting ranking of tenders becomes arbitrary as a result. When determining the weights for the criteria, the way the score of each criterion is determined (measures and performance levels) should be considered. Not understanding this, is the most common mistake in weighting procedures and produces arbitrary weights that are misinterpreted as reflecting the relative importance of the criteria (Mateus, Ferreira and Carreira, 2010). Also, procurers rarely focus on determining if they value a similar difference in increase of quality at distinct points of the quality scale differently (Bergman and Lundberg, 2013).

There can also be problems with defining the quality scale: for example, if there are minimum requirements in place for the level of quality which have already affected the participating tenders, but the quality scale measures this quality criteria over a wider range of values this affects the importance of the quality differences in the scoring. Thus, despite specified exact weights, the evaluation cannot be foreseen (Bergman and Lundberg, 2013).

Furthermore, without knowing the exact scoring rules, as presented by Mateus et al. (2010), it is impossible to predict what levels of performance in each criteria will give the best overall score. This means that the vendor won’t be able to predict how to offer economically the most advantageous tender, that means usually the best value for money, according to the procurer’s preferences (Mateus, Ferreira and Carreira, 2010).
2.2 Usability in IS Procurement

Usability is one of the quality criteria referred to in the previous section that can be identified and included in the procurement. In the literature there are several ways discussed in which usability should be incorporated in the process, as I go through in this chapter. In the early 1990s usability requirements and their importance in public IS procurement was highlighted (Carey, 1991), and a few years later also the importance of understanding human factors during procurement in general (Scown, 1998). During this time it was also identified, that the origins of usability engineering or UCD is in a collaborative setting with users’, clients’ and developers’ interests in alignment and the contractual setting that exists in procurement might require changes to the process (Carey, 1991). Years later, researchers still argued that as the research focus has been on the process and methods on designing for good usability in the organizations developing the software, the procuring organizations have trouble implementing similar practices in the procurement setting because good practices do not exist (Artman and Zällh, 2005).

However, there have also been arguments to the contrary, stating that public organizations are not ready to take responsibility for usability and thus avoid including it in procurement (Jokela and Buie, 2012). Against the perhaps more generalized opinion of not seeing its value, these organizations do seem to find usability important (Artman and Zällh, 2005). According to researchers, the problem lies in their understanding of how to require, develop and monitor it during a procurement. They see that the solution is in addition to requiring usability to also taking on a more active role in working towards it during the entire process. The challenge however is in the procurers’ limited understanding of how to do it properly and how complex it will be to integrate these practices successfully (Artman and Zällh, 2005).

When shifting the focus from system development projects to purchasing existing software, the challenges become somewhat different. For an existing product, applying the iterative UCD approach might be impossible from a contractual standpoint (Lauesen, 1998). Although it could theoretically be used for enhancements, vendor-customer agreement on terms and price might be difficult. Unrealistic expectations might mean no system to acquire (Lauesen, 1998). Thus, the key difference between processes of UCD during product development and when implementing commercial-off-the-shelf (COTS) software are less (or no) possibility for developing the software further based on user needs or identified usability issues and iterative evaluation (Beuscart-Zéphir et al., 2002). However, when purchasing configurable systems, there
might be opportunities to implement user-centred processes in this implementation work (Ratwani et al., 2018).

This means, that focusing on the contractual requirements and the selection phase are key, if the IS can possibly not be developed further later in the procurement process. Research has proven, that usability criteria and user evaluation can be included in public IS procurement for selection of COTS software (Riihiaho et al., 2015), however, the process had not been established yet in the early 2010s (Jokela, Laine and Nieminen, 2013). Results (Beuscart-Zéphir et al., 2005) suggest that a procedure of usability evaluation during procurement can prevent unusable software from being selected and support the further re-engineering (if possible) to make the system acceptable for users. This approach was also seen to support the goal of involving users with the selection. The role and input of users in selecting the software has indeed been identified in earlier research as key to a successful procurement, since it will affect their work in a significant way (Howcroft and Light, 2002). It has also been identified to be essential that users from different functional areas are involved in the process (Moe, 2014). When it comes to selecting existing software, users’ input is seen essential because the question is not only on understanding user requirements but also comparing them against the existing possibilities. However, at least in the early 2000s, the practices of including user perspectives were not in place despite recommendations (Howcroft and Light, 2002).

Previous research on including usability in the requirements, evaluating usability during the selection of IS in healthcare and suggestions for usability evaluation practices during procurement will be presented in the next subsections.

2.2.1 Usability Requirements in IS Procurement

As stated earlier, unrealistic expectations might mean no system to acquire, and this dilemma applies also to usability requirements. In the late 1990s, it was recognized that there was not enough literature on usability requirements and the existing literature was lacking especially on concrete examples of how to write them (Lauesen, 1998). The earliest attempts to include usability requirements in the procurement requirements documents suffered from similar problems than with other types of requirements (Carey, 1991): they were not written so they could be validated with testing, they were too easy and systems that could fulfill them were not actually acceptable, they were too strict and the requirements could rule out acceptable systems, and they were not
complete with providing needed information on meeting them. Examples of these are “the system must be easy to learn” or demanding certain features for the UI (Carey, 1991).

According to Carey (1991) the biggest challenge comes from the incomplete requirements, as usability exists in relation to users - and the users can evolve, even with the new system itself. Carey states that the need for testability of requirements can be addressed with measurable usability objectives, however the incompleteness could still be an issue from the vendor point of view. And while trying to give more information the requirement might become overstated. The testability of the requirements beyond the initial learning stage can also be a problem according to Carey. While the timeline could be extended to cover longer periods of use for measurable requirements, defining controllable procedures for measuring them is challenging (Carey, 1991).

Usability requirements can be categorized into six different styles (Lauesen, 1998; Lauesen and Younessi, 1998): performance, defect, process, subjective, design and guideline. Lauesen and Younessi (1998) elaborate on the definition of these different styles of requirements in the following way: Performance style requirements specify the users’ performance on certain tasks, usually with time. Defect style requirements specify how many usability problems users can encounter during task completion. This can also give feedback on correcting the system. Process style requirements describe how usability should be ensured during the development of the software. Subjective style requirements rely on users for their opinions. Design style requirements describe the user interface (UI) in detail and are closer to functional requirements. Guideline style requirements specify general appearance and functioning for the whole system by relying on existing internal or external guidelines.

All these different styles have also their drawbacks (Lauesen, 1998; Lauesen and Younessi, 1998) and not all are suitable for COTS software procurements. According to Lauesen and Younessi (1998) the challenge with performance style is it requires experience for choosing the right tasks and it can also be challenging to apply when selecting a standard system. They also state that process style requirements can give too much freedom to developers on users, tasks, and corrections. The challenges with subjective style are related to the inconsistencies between user preferences and actual efficiency (Lauesen and Younessi, 1998), and satisfaction being influenced also by organizational factors and possibly out of control of the vendor (Carey, 1991). With design style requirements, Lauesen and Younessi (1998) elaborate that they prevent iteration during development and leave all responsibility for the usability on the one writing the requirements. This can obviously result in an adequate system
if a user-centred approach was used for the requirements process. They are also not suitable for existing systems. And finally, the guideline requirements are difficult to verify, but useful in addition to other styles of requirements (Lauesen and Younessi, 1998).

One of the main challenges with using certain types of usability requirements in IS procurement of existing systems, is also whether the vendors will accept and commit to them. This applies especially to performance and defect styles (Lauesen, 1998). With performance style requirements setting target values can be risky when selecting from standard systems - they can easily be either too restrictive or too relaxed. According to Lauesen one suggested solution is letting vendors set the target values and then including these values when comparing the proposals. Another proposed solution is using alternative styles of requirements that the vendor can choose from to commit to, depending on their solution (Lauesen, 1998).

Finally, the key elements in eliciting usability requirements according to Lauesen (1998) are defining critical tasks, user profiles, system goals and previous usability problems to identify key usability issues for the new system. He emphasizes that selection of critical tasks is essential, because in complex systems, covering all possible tasks and their usability is impossible. To do this appropriately, domain knowledge is required (Lauesen, 1998).

2.2.2 Usability in Selection Phase of HIS Procurement

Healthcare has been recognized as a complex IS domain (Mirel, 2004). Perhaps consequently, most literature found on including usability, and especially usability evaluation, in the selection phase of existing IS is written in the healthcare context. Already in the late 1990s, the importance of UCD and usability considerations of health applications during purchasing was promoted (Jimison et al., 1999). At that time, the suggestion was to base purchasing decisions on existing evaluation results of the systems. It was seen that this requirement would also influence the development of such applications to include these practices which were not utilized by the industry (Jimison et al., 1999). However, it has taken two decades to reach a point now where US based EHR vendors seem to actually utilize UCD practices reliably during development (Hettinger, Melnick and Ratwani, 2021), while still little over five years ago this was not the case (Ratwani et al., 2015).

While usability evaluation as a methodology has become more common in the healthcare IS domain, this does not directly extend to the situation of procurement where a more quantitative approach is needed in order to compare
the systems (Jorritsma, Cnossen and van Ooijen, 2014). Indeed, researchers (Jensen, Rasmussen and Lyng, 2013) state that there is a lack of descriptions on how usability evaluation methods can be used in procurement, and as presented in previous sections the requirements of public procurement makes the situation even more challenging. The key challenges they have identified with usability assessments in public procurement are: results must be comparable; the assessment must be uniform; process has to be transparent; there is limited time to conduct and analyze; as well as data must be easily collected and available (Jensen, Rasmussen and Lyng, 2013).

It has been also criticized that when acquiring EHR systems, the purchasing organizations presume usability of the system to be the responsibility of the vendor (Edwards et al., 2008). However, these systems are also considered highly configurable which could be used to improve their usability during the implementation phase after selection (Møller-Jensen, Lund Pedersen and Simonsen, 2006; Ratwani et al., 2018). This would imply the purchasing organization needs to take responsibility for the usability in later stages of procurement (Edwards et al., 2008).

Traditional procurement practices, also in the healthcare context, include asking vendors for demonstrations (Beuscart-Zéphir et al., 2002). These demonstrations can give insight on both the suitability of the software and the vendors (Howcroft and Light, 2002). The risk with demonstrations seems to be that the art of sales easily becomes more important than the actual system in making the purchasing decision (Howcroft and Light, 2002). Also, traditional vendor demonstrations only provide weak evidence on system usability (Kushniruk et al., 2010). Another common practice, also for selection of software in hospitals, is relying on comparing the fulfilment of functional requirement lists (Jorritsma, Cnossen and van Ooijen, 2014). This approach does not take into consideration how that functionality is actually implemented, which has been proven to result in difference in usability between systems that fulfil identically the functional requirements (Jorritsma, Cnossen and van Ooijen, 2014).

Researchers also in the healthcare domain emphasize that the user needs analysis (Beuscart-Zéphir et al., 2005), planning the evaluations carefully because of a specialized system environment and using experts in the work domain in this planning is important during procurement (Jorritsma, Cnossen and van Ooijen, 2014). This ensures that the tasks are representative of the actual work outside the test setting. However, in the case of large-scale healthcare procurements, there are additional challenges related to this: how to cover all aspects of the system, consider all specialties, and take into account
all types of users (and possibly also include them) in the evaluation (Jensen, Rasmussen and Lyng, 2013).

In Finland, it has been revealed by research that there is often variance in processes, conventions and requirements just within one organization in one field of healthcare (and social welfare), and even more differences between municipalities - even though the regulations by law are the same (Alanne et al., 2015). Other researchers concur, that healthcare is a very fragmented field in which different specialties have their own requirements for their IS, which must in turn support work processes and tasks (Rantanen and Heimo, 2014).

In the following, I will go through, based on examples presented in literature, how usability evaluation has been or is proposed to be included in the selection phase of IS procurement in the healthcare domain. These examples include both those in the private sector (where practices are not regulated) and those in the public sector. It seems typical that multiple methods are used in the same case (see Table 1).

<table>
<thead>
<tr>
<th>Case Example</th>
<th>Questionnaire (when used)</th>
<th>Usability Inspection (method used)</th>
<th>User Testing (method used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beuscart-Zéphir et al. (2002)</td>
<td>Demonstrations</td>
<td>Heuristic evaluation</td>
<td>Usability testing in real context</td>
</tr>
<tr>
<td>Beuscart-Zéphir et al. (2005)</td>
<td></td>
<td>Heuristic evaluation</td>
<td>Usability testing in real context with real patients</td>
</tr>
<tr>
<td>Holbrook et al. (2003)</td>
<td>(Demonstrations), User testing</td>
<td></td>
<td>Usability testing</td>
</tr>
<tr>
<td>Jensen et al. (2013)</td>
<td>User testing</td>
<td></td>
<td>Clinical simulation</td>
</tr>
<tr>
<td>Jorritsma et al. (2014)</td>
<td>User testing</td>
<td></td>
<td>Usability testing</td>
</tr>
<tr>
<td>Kannry et al. (2006)</td>
<td>Demonstrations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liljegren and Osvalder (2004)</td>
<td>Clinical trial</td>
<td>Cognitive walkthrough</td>
<td>Usability testing in real context</td>
</tr>
<tr>
<td>Marcilly et al. (2021)</td>
<td>Usability inspection</td>
<td>Usability walkthrough</td>
<td></td>
</tr>
</tbody>
</table>
Utilization of User Questionnaires

Different types of questionnaires seem to be utilized commonly in the literature in different setups for evaluating with users during procurement of HIS. In a procurement, questionnaires cannot be a standalone method but need to be used with some other method to evaluate the system. These include questionnaires applied in relation to demonstrations (Beuscart-Zéphir et al., 2002; Holbrook et al., 2003; Kannry, Mukani and Myers, 2006) as well as after using the system (Beuscart-Zéphir et al., 2002; Holbrook et al., 2003; Liljegren and Osvalder, 2004; Jensen, Rasmussen and Lyng, 2013). The purpose and goal of using questionnaires also differs. From the standardized usability questionnaires, System Usability Scale (SUS), has been recommended for EHR system procurement as a post-test questionnaire (HIMSS EHR Usability Task Force, 2010).

Holbrook et al. (2003) described their procurement process of an electronic medical record (EMR) where all input was collected with questionnaires based on demonstrations and actual test setups. During the demonstrations they were used for subjective evaluation of specific functionalities rather than usability with 101 questions in 14 domains, while after a user test setup the questionnaire included rating of the system with five questions regarding UI, use with training, improvement of care and potential for actual use (Holbrook et al., 2003). Kannry et al. (2006; according to Kushniruk et al., 2010) describe using questionnaires during demonstrations that were scripted to follow clinical workflows. In these demonstrations users graded the scenarios and the grades were mapped to functionalities. Beuscart-Zéphir et al. (2002) developed a rapid usability assessment methodology to use in the late stage of a procurement. Their methodology included questionnaires during demonstrations and after user testing. During demonstrations three dimensions related to usability were addressed: “rapidity” (quick to use), “ease of use” (for data reading and data entering) and “exhaustiveness”, while after testing users assessed the usability for tested functionalities (Beuscart-Zéphir et al., 2002). The questions used Likert scales for evaluation, however, it is not clear what questions were used exactly.

Liljegren and Osvalder (2004) incorporated usability evaluation methods into selection of a patient monitoring system for a hospital. They applied a user questionnaire among other methods to evaluate the usability of three systems after their use in a clinical trial. The questionnaire was extensive with 112 questions, rating 13 functionalities from three perspectives: “presentation of
information”, “physical properties” and “difficulty of tasks” compared to the current system (Liljegren and Osvalder, 2004). Similarly, in their study, Jorritsma et al. (2014), simulated the selection phase of procurement by comparing three picture archiving and communication systems (PACS) with both usability testing and fulfilling of functional requirements and comparing their results. A usability questionnaire with 19 statements was used with other measures to evaluate the usability. The entire questionnaire is included in the article and many of the questions seem general and similar to established usability questionnaires but also include more feature specific questions (Jorritsma, Cnossen and van Ooijen, 2014).

In their procurement process, Jensen et al. (2013) developed a clinical simulation method for the selection of an EHR system. The method included using questionnaires to collect user satisfaction ratings after the simulated use of the systems. They developed questionnaires based on existing usability questionnaires (Davis, 1989) to use as part of an observation guide for future end users to assess usability during the clinical simulations where other users performed the tasks (Jensen, Rasmussen and Lyng, 2013).

Schumacher et al. (2009) suggest using usability walkthroughs to evaluate usability with a group of users who after review of each task rate the different usability attributes for that task: effectiveness, efficiency, and subjective satisfaction. In this method, the group consists preferably of a usability expert who guides the process and several users. This method has been used recently by Marcilly et al. (2021) in a real-life procurement case. In their case, the users filled in the SUS questionnaire after the usability walkthrough session.

Using Usability Inspection Methods
There are not many examples of using expert evaluation methods during procurement. Beuscart-Zéphir et al. (2002) developed a model of usability evaluation for procurement that applied heuristic evaluation (HE) to uncover usability problems which would inform the decision of selecting between clinical information systems for a hospital. The evaluation was done by three usability specialists and took one day per system. A severity rating on a five-point scale was given to all discovered problems, and they were categorized based on the heuristic violated. This method was also used in another procurement case by Beuscart-Zéphir et al. (2005) when selecting a CIS for another hospital, this time a four-point scale for severity was used.

Liljegren and Osvalder (2004) applied also cognitive walkthrough to evaluate the usability of patient monitoring systems. In this case, five typical and
common tasks were selected for evaluation with the method. The evaluation was performed by one usability expert on the three candidate systems and resulted in a list of potential problems.

Marcilly et al. (2021) used the usability walkthrough method suggested by Schumacher et al. (2009) when selecting an EHR for a hospital. In the case, nine users from different user groups (three nurses and physicians, pharmacist, secretary, and admissions officer) evaluated five different EHRs during three-hour walkthrough sessions. Realistic user scenarios were prepared and provided to the vendors beforehand. The problems were documented by a usability specialist during the session and gone through afterwards with the users.

Performing User Testing
Different types of user testing methodologies have been used in procurement cases. In the model developed by Beuscart-Zéphir et al. (2002, 2005), usability testing was one of the methods. In the first case (Beuscart-Zéphir et al., 2002), the usability testing was described to take three days per vendor and include 67 different users from three different user groups testing the three candidate systems during one hour sessions. The study design was between-subjects, with the exception of three users (Beuscart-Zéphir et al., 2002). In both these cases, the usability tests were conducted in the real future work context. In the second case, the test were performed with three users testing both candidate systems with real patients (Beuscart-Zéphir et al., 2005). In this case, both qualitative and quantitative data was collected and compared to the current situation with paper-charting.

Liljegren and Osvalder (2004) used also usability testing in their case for selecting a patient monitoring system. A between-subjects design was used in their study, with 18 users divided into three groups (one per evaluated system). The tests used the same tasks as in the cognitive walkthrough with two new added. The documentation from the tests in the study by Liljegren and Osvalder was completion of tasks, use of system-specific features and whether usability problems from cognitive walkthrough were also discovered in usability testing.

Jorritsma et al. (2014), utilized a within-subjects design in their simulated study for using usability testing in procurement. In their study it was, however, not possible to counterbalance the order of the systems. Prior to the usability testing the vendors demonstrated the systems to the users in relation to the tasks they would then perform. There was no interaction between the
system and the users at this stage. From the tests, they collected quantitative data of performance along with user’s answers to questionnaire.

Holbrook et al. (2003) describe how they used user testing as part of the overall evaluation framework of selecting an EMR for primary care physicians. In their case 12 users attempted charting two different patient cases with four candidate systems. The evaluation took half a day per user. The collected data, however, appears to be completely questionnaire based.

The clinical simulation method developed by Jensen et al. (2013) could also be considered as a type of user testing. In their case, the usability evaluation was done after initial negotiations and reduction of number of vendors took place. The users performed the simulation in teams of six professionals, and there were three teams in total. There were also two observers. All teams used a total of nine days to evaluate the three systems, which makes three days per system. Observations from the simulation were gathered in addition to the user’s answers to questionnaires (Jensen, Rasmussen and Lyng, 2013).

**Used Metrics, Measures and Scoring Rules**

In procurement the handling of usability metrics and measures to be used for comparing the systems becomes an essential issue. However, the cases described in literature usually describe the measures at least in general, but the details of measures, rankings and scoring rules are largely absent, with the exception of Liljegren and Osvalder (2004). Usually it seems that either no unified scores are used or at least they are not described (Beuscart-Zéphir et al., 2002, 2005; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014).

The studies utilizing the model developed by Beuscart-Zéphir et al. (2002, 2005) did not produce a unified score for comparing the usability of the systems based on evaluations, but rather the analysis of the numeric results used qualitative analysis to come to a recommendation between the systems. The HE results were quantified in a simplistic way by calculating a mean score for each heuristic from the severity scores of reported violations by individual experts. The authors claim that this identifies flaws and strengths, and helps comparison between vendors solutions (Beuscart-Zéphir et al., 2002). From the usability testing procedure, in the first case qualitative data on problems as well as the usability rating of functionalities by participants (Beuscart-Zéphir et al., 2002), and in the second case time for execution of tasks and qualitative data on issues with documentation workflow (Beuscart-Zéphir et al., 2005) were collected. This process aimed also at supporting the implementation and re-engineering following the selection by making qualitative
analysis of the usability problems and workflows (Beuscart-Zéphir et al., 2002, 2005). This process does not seem possible for a governmental procurement, as qualitative analysis of numeric results is in such a key role.

In the study by Liljegren and Osvalder (2004) the tested systems were ranked based on the results from each of the three methods (user questionnaires, cognitive walkthrough and usability test) and the rankings were then combined into one ranking. The used metrics in their study were: for the questionnaire “number of median ratings higher than 3”, for cognitive walkthrough “number of potential problems” and for usability testing “number of encountered problems from CW”. The highest overall ranking was awarded to the system that was ranked first the most times. This ranking, however, had only an informational value in the final purchasing decision (Liljegren and Osvalder, 2004).

In the usability walkthrough suggested by Schumacher et al. (2009) the user ratings for effectiveness, efficiency, and subjective satisfaction for each task are averaged for all of the tasks for each evaluated system. They further suggest a three-dimensional visualization to plot average ratings for each product for effectiveness (x-axis), efficiency (y-axis) and subjective satisfaction (bubble diameter). This visualization can be used to visually pick the vendors that will then be evaluated in more detail (Schumacher, Webb and Johnson, 2009). The actual procurement case described by Marcilly et al. (2021) utilizing this method does not describe in detail how the results were quantified and compared. They only mention collecting the SUS score and determining severity ratings for the identified problems.

In their case, Jensen et al. (2013) discuss the fact that to be able to compare the vendors’ solutions in a uniform way, the assessment needs to produce quantitative results. The data collected was task completion rate and number of difficulties in using the system collected by observation, in addition to the user ratings from questionnaires (Jensen, Rasmussen and Lyng, 2013). However, they do not describe how these measures are exactly quantified and combined for the final scoring.

The study presented by Jorritsma et al. (2014) did not combine the measures collected into a unified score which the paper acknowledges would be mandatory in a real-life procurement case. They provide two suggestions for the combination following the process of assigning scores to requirements on a list and then calculating a weighted average score: either combine the usability measures first and then treat the usability evaluation as one additional requirement with its own weight; or treat each usability measure as a separate requirement. In the first case they suggest following the method presented by
Sauro and Kindlund (2005). The measures used in this study were execution time for task and number of interactions with the system (mouse and key presses) to evaluate efficiency, and user answers to questionnaire items (Jorritsma, Cnossen and van Ooijen, 2014). The study also suggests using multiple measures to get a more accurate and reliable view of the system. This is due to their finding of the subjective evaluations of users on usability differing from the objective task performance measures.

2.2.3 Suggestions for Usability Evaluation Procedure

The four primary steps in the first two phases of procurement in which usability should be included are eliciting usability requirements, comparative evaluation of multiple systems to reduce the number of vendors, deep evaluation of few systems for selection and establishing usability requirements in the final RFP (Schumacher, Webb and Johnson, 2009). Practical suggestions for the two evaluation steps based on literature in the field of healthcare IS procurement are presented in the following subsections.

Two-phase Evaluation Process

The suggestion on a two-phase procedure for narrowing down the number of vendors is mentioned by several references (Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014). The preliminary assessment can be based on scoring the fulfilment of functional requirements (Jorritsma, Cnossen and van Ooijen, 2014), establishing minimum usability requirements (Beuscart-Zéphir et al., 2005) or estimating relative usability with expert evaluation methods (such as usability walkthrough or heuristic evaluation) (Schumacher, Webb and Johnson, 2009). The shortlist evaluation where usability is measured with usability testing should be based on established numeric goals (Schumacher, Webb and Johnson, 2009). For the measures to be suitable for the selection, they should be combined into a single score (Jorritsma, Cnossen and van Ooijen, 2014).

Selecting Appropriate Methods

The approach of two-phases of usability evaluation with slightly differing goals for the evaluation is supported by recommendations on the use of traditional usability evaluation methods in procurement. Heuristic evaluation is considered as a suitable method for preliminary assessment (primarily because of
cost-effectiveness and quickness) (Carvalho, Borycki and Kushniruk, 2009; Kushniruk et al., 2010) and usability testing is recommended by several sources (Beuscart-Zéphir et al., 2005; Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014).

The selection of suitable methods for usability assessment in healthcare IS procurement is aided by a framework that considers different methods based on their given level of evidence on usability (Kushniruk et al., 2010). The framework places different methods traditionally used and suggested for use in procurement on a continuum of evidence. Based on the framework (see Figure 2) the least evidence is provided by traditional vendor demonstrations (focusing on functionalities), followed by demonstrations following clinical information processing scenarios (CLIPS). Utilizing CLIPS with heuristic evaluation and usability testing are considered to give strong evidence, with on-site analysis giving the strongest evidence (Kushniruk et al., 2010).

The planned procurement usability evaluation procedure should support both the selection of a vendor and ensure that the usability of the selected vendor is at an acceptable level (Beuscart-Zéphir et al., 2005). However, there are also other purposes that should be considered depending on the procurement case: supporting the configuration process during implementation and identifying needs for further development (Beuscart-Zéphir et al., 2002). Other considerations for selecting the methods are time and effort required compared to the complexity of the object of procurement: more complex procurements where the procurer is unsure whether suitable solutions exist or the future acceptance of the system is known to be difficult should utilize a more elaborate evaluation procedure (Beuscart-Zéphir et al., 2005). Outside the healthcare context, it has been identified that planning and execution of usability testing during procurement is resource intensive: a test setup of five tasks per two different user groups and two vendors took one month of effort (Riihiaho et al., 2015).

**Figure 2.** Methodological continuum of evidence on usability. Adapted from Kushniruk et al. Increasing the safety of healthcare information systems through improved procurement: toward a framework for selection of safe healthcare systems. Healthcare Quarterly. (Kushniruk et al., 2010)
Selecting the Objects of Evaluation and Planning the Procedure

The planning of the evaluations is as important as the execution, especially when public procurement is in question. The equal treatment of vendors means a strictly followed procedure must be used (Jensen, Rasmussen and Lyng, 2013). Selecting and understanding the key user groups and use contexts when developing the scenarios used in the evaluation is the first step (Schumacher, Webb and Johnson, 2009; Jorritsma, Cnossen and van Ooijen, 2014). In complex procurements it is impossible to include all different user groups in hands-on evaluations due to time restrictions, so prioritization needs to be done (Jensen, Rasmussen and Lyng, 2013). Identifying critical and frequent tasks should be the second step (Schumacher, Webb and Johnson, 2009). Selecting the right scenarios is important because only a fraction of all functionalities can be tested and thus they must be representative and diverse enough to include several user groups with multitude of tasks and work contexts (Jensen, Rasmussen and Lyng, 2013). It is also suggested to use both tasks that are specific and those that are more loose to evaluate different aspects of usability (Jorritsma, Cnossen and van Ooijen, 2014).

In public procurement, the scenarios used need to be described in detail beforehand and performed almost identically in order to treat the vendors equally (Jensen, Rasmussen and Lyng, 2013). This could be seen to present a challenge for the design and writing of the tasks for unknown systems. There is also a need for at least vendor trainings for users and users practicing before the evaluation should be considered (Jensen, Rasmussen and Lyng, 2013).

2.3 Usability Evaluation

While the previous section presented literature on how usability evaluation methodology is applied into IS procurement and specifically in the field of HIS, this section takes a more generic view to the topic. I will go through this methodology as it has been established in the context of the UCD development cycle. As presented in the introduction, in section 1.3, several definitions for usability exist in literature that include slightly different components to them (Nielsen, 1993; McGee, Rich and Dumas, 2004; ISO, 2018). These definitions are used as guiding principles when evaluating usability. Usability evaluation is a practice used primarily during development to gather qualitative evidence on improving the product’s usability. Additionally, it is also used for benchmarking, comparing design versions and validating the end result in a quantitative manner (Sauro and Lewis, 2012). When it comes to complex information systems and their evaluation, researchers have pointed out the
absence of suitable usability evaluation methods (Redish, 2007; Chilana, Wobbrock and Ko, 2010). In the following subsections, I present the most established usability evaluation methods, practices and issues related to quantifying usability, and finally, how usability evaluation methods are developed.

2.3.1 Existing Usability Evaluation Methods

Usability evaluation methods can be divided to different categories in several ways based on different aspects of the methods. One categorization can be done by using the aspect of user involvement and way of doing the evaluation: inspection methods which don’t include users (also called expert evaluation methods), inspection methods with users (group inspection methods) and user testing methods (e.g. Nielsen, 1993; Dumas and Salzman, 2006). A similar, but slightly different way of categorizing the methods is to divide them into subjective and objective methods based on the perspective (e.g. Hornbæk, 2006). The practice and purpose of usability evaluation can also be used for the categorization: formative and summative evaluation (Hartson, Andre and Williges, 2003). The former is performed during product development and aims to identify usability problems and develop the evaluated product or UI further, while the latter aims at assessing the usability level of the product (usually at the end of the development process) and its performance against usability requirements (Nielsen, 1993; Hartson, Andre and Williges, 2003). A distinction between these two approaches is also the type of data they focus on: formative evaluation focuses on qualitative data while summative evaluation focuses on quantitative data (Nielsen, 1993; Hartson, Andre and Williges, 2003). The challenge with all categorizations is that some methods are difficult to fit in them at all, and others might fit into multiple categories with only slight modifications.

Inspection Methods

One of the most established methods for expert evaluation of usability is **Heuristic Evaluation (HE)** (Nielsen and Molich, 1990; Gulliksen and Göransson, 2001). Especially in formative development contexts where the design can be adjusted or functionality changed based on the evaluation results, it was in wide use already in the 1990s (Dumas and Redish, 1999; Gediga, Hammond and Düntsch, 1999). During the years, HE has also been used as a basis for modifying it to different purposes (e.g. Sears, 1997; Sutcliffe, 2002). The drawbacks of traditional HE are none or limited incorporation of use context, a restricted set of UIs that can be evaluated, a requirement of selecting the UIs
prior to evaluation and no capability to identify major missing functionalities during the evaluation (Chin, Diehl and Norman, 1988; Nielsen, 1993).

The procedure and principles of the HE discussed above can be described shortly as follows. The evaluation focuses on the UI and is based on established design guidelines, heuristics (Nielsen and Molich, 1990; Nielsen, 1993). The evaluation is done in two stages (Nielsen and Molich, 1990; Nielsen, 1993): First, evaluators go through the interface independently, noting down all violations of the heuristics and scoring them on severity. Second, the evaluators aggregate their findings into one list of violations and agree on their severity. Severity can be scored either as one dimension or with impact and frequency separately (Nielsen, 1993). Traditionally, the evaluation is done on the interface as a whole and not task-based (Nielsen, 1993).

Another well-known expert review method is Cognitive Walkthrough (CW) which went through several iterations in the early 1990s to arrive at its most known version (Lewis et al., 1990; Polson et al., 1992; Wharton et al., 1994). Like HE, also it has been modified further (e.g. Sears, 1997). The premise of CW is to evaluate the learnability of a system from the perspective of a novice user. The procedure of performing CW is to examine the interface by going through the steps of accomplishing a task, and asking four questions at each step (Wharton et al., 1994): 1) “Will the user try to achieve right effect?” 2) “Will the user notice that the correct action is available?” 3) Will the user associate the action with the effect? 4) After the action, will the user understand that progress was made? The answers to the questions and any issues uncovered during the analysis are documented and an overall judgement is made for each of the steps in the task (Wharton et al., 1994). Task efficiency can also be modelled by experts, and the most used method is GOMS analysis which estimates used time by counting key and mouse interactions (Kieras, 2012).

Another usability inspection method mentioned in subsection 2.2.2 when discussing methods used in procurement of HIS was usability walkthrough (Schumacher, Webb and Johnson, 2009; Marcilly et al., 2021). It is not clear what methodological procedure exactly is referenced in this instance, as there are several usability evaluation methods that use walkthrough in their name. The description by Schumacher et al. (2009) include the idea of a group of users from different roles led by a usability specialist going through tasks with the system and documenting the problems. Karat et al. (1992) on the other hand described usability walkthrough as a two-stage process including first a “self-guided exploration” of the UI and then “use of prescribed scenarios” that could be either done individually or as pairs.
User Testing Methods
When it comes to analyzing the usability of IS with users, usability testing is considered an established method (Nielsen, 1993). It has been regarded as the “gold standard” for both determining the usability of an individual system as well as comparing it between systems (Dumas and Fox, 2012). In a traditional usability test, the user is guided by a facilitator in a controlled laboratory setting to attempt completing tasks with the system that have been prepared beforehand, usually accompanied with a think-aloud methodology (Nielsen, 1993). When the purpose of usability testing is qualitative discovery of usability issues, usually five users is considered to be sufficient (Nielsen, 1993). For quantitative testing, it is generally regarded that a significantly larger number of users is needed but this is necessarily not the case (Sauro and Lewis, 2012). Sauro and Lewis (2012, chaps 6 & 7) go through extensively in their book how the appropriate sample size depends on the goals of the study and the expectations for the results with both qualitative and quantitative testing and can be calculated for each case with appropriate formulas.

A variation of traditional usability testing is doing it in teams of two or more users (Hackman and Biers, 1992). Hackman and Biers (1992) concluded that in such a test the discussion between users affects the time on task and perceived learnability compared to a single user, while the overall performance or subjective evaluation is not affected.

Usability Questionnaires
For post-use usability evaluation, especially when products are developed, it is a common practice to use usability questionnaires. Different types of questionnaires were also used in the procurement cases in subsection 2.2.2. The use of standardized questionnaires is encouraged in the usability literature. There are several such questionnaires, and these focus mainly on either perceived usability or user satisfaction. Sauro and Lewis (2012, chap. 8) present altogether 19 different standardized usability questionnaires. In the following, I will go through some of these, namely QUIS (Questionnaire for Interaction Satisfaction) (Chin, Diehl and Norman, 1988), SUMI (Software Usability Measurement Inventory) (Kirakowski and Corbett, 1993; Kirakowski, 1996), CSUQ (Computer System Usability Questionnaire) (Lewis, 1995, 2002) and SUS (System Usability Scale) (Brooke, 1996).

For QUIS (Chin, Diehl and Norman, 1988), the version 5.0 of the questionnaire consists of 27 items in five themes. Previous versions had different number of items, with the original version at 90 items. The themes incorporated into QUIS 5.0 are: overall questions, screen factors, terminology, and system feedback, learning factors and system capabilities. The scale that is used in
QUIS has two opposing concepts, such as “confusing” and “clear” (bipolar scale) and the 5.0 version has a range of 0 to 9. The resulting scores for the items are viewed individually (Chin, Diehl and Norman, 1988). SUMI (Kirakowski and Corbett, 1993; Kirakowski, 1996) has 50 items. These are divided into a global scale of 25 items and five subscales which each have 10 items: efficiency, affect, helpfulness, control, and learnability. The answers to SUMI are rated on a three-point scale: “agree”, “undecided” and “disagree” and results for each scale are viewed on a scale whose mean is 50 and standard deviation 10. The long-used CSUQ version 2 (Lewis, 1995), a variant of the PSSUQ questionnaire, had 19 items. The most recent version of CSUQ, however, has 16 items (Lewis, 2002). The questions are on ease of use, learnability, clear and relevant information, and user interface. CSUQ has a seven-point scale, from “strongly disagree” to “strongly agree” and provides four scores that are averaged over the responses to the relevant questions: overall score, system usefulness, information quality and interface quality (Lewis, 2002). And finally, SUS (Brooke, 1996) includes ten items and generates an overall summative score between 0-100. The statements are rated on a five-point scale of “strongly disagree” to “strongly agree”. A distinctive feature of SUS is its statements alternating between negative and positive. However, a positive version has also been devised (Sauro and Lewis, 2011).

Examples of statements from these questionnaires are: “Terminology relates well to the work you are doing” (QUIS) (Chin, Diehl and Norman, 1988), “Tasks can be performed in a straightforward manner using this software” (SUMI) (Kirakowski, 1996), “The information provided with the system is effective in helping me complete my work” (CSUQ) (Lewis, 2018) and “I think that I would like to use this system frequently” (SUS) (Brooke, 1996). The most used questionnaires for post-test studies are CSUQ and SUS which have been proven to correlate in their measure of perceived usability (Lewis, 2018).

In the context of healthcare, when evaluating usability from the end-users’ perspective the contextual aspects of work are important which these standardized usability questionnaires lack in including (Viitanen et al., 2011). Such aspects in this context are the concurrent use of several different information systems, work practice variability, and the different specialties and work contexts. To conquer this shortcoming, a new usability questionnaire was developed first for physicians (Viitanen et al., 2011; Kaipio et al., 2017), and subsequently it was adapted also for nurses (Hyppönen et al., 2018) and social welfare professionals (Ylönen et al., 2020).
Limitation of Traditional Usability Evaluation Methods

However, these established methods do also have their issues. Both inspection methods and qualitative usability testing have been criticized for the evaluator effect i.e. lack of agreement on usability problems and also their severity by several studies (Hertzum and Jacobsen, 2003; Dumas and Fox, 2012). Hertzum (2010) has also criticized the widely used usability evaluation methods (usability testing and heuristic evaluation) for focusing on a very limited view of usability. These methods consider specifically the individual and short-term perspectives of use and their established position in the field might lead to unintentional omission of other perspectives to usability when planning its evaluation (Hertzum, 2010). According to Hertzum (2010), especially the collaborative and long-term perspectives are lacking in evaluation methodology.

Usability evaluation can also be practiced outside the controlled laboratory setting, called “field usability” or “naturalistic usability”, in the actual environment of system use with a real setting (Siegel, 2012). This approach was used in some of the procurement cases presented in subsection 2.2.2. Naturalistic studies attempt to mitigate the limitations of traditional usability testing, such as unrealistic motivation for use, bias for the initial stages of using a system, focusing on individuals using the system compared to in an actual organizational context and a limited range of tasks (Siegel, 2012). While usability testing in the laboratory setting focuses mainly on the perspective of situational usability, evaluation in the field can also reveal aspects of organizational usability (Hertzum, 2010, 2018). The different levels of evaluation for organizational usability are user and the system, organization and the system, and environment and system of which only the first can be evaluated in laboratory settings (Hertzum, 2018). Another differentiating aspect of evaluating organizational usability, is the focus on benefits realization as opposed to usability problems (Hertzum, 2018).

2.3.2 Measuring Usability

As described earlier, certain scenarios of usability evaluation require the results to be in a quantitative form. This means that a qualitative attribute of a system, usability, needs to be dissected into aspects that can be measured and thus presented numerically. A fundamental issue of measurement is how the construct and the numbers used to measure it are related (Gillan and Bias, 2014). Gillan and Bias (2014) state that a measurement model describes how an unobservable construct (what is ultimately being evaluated/tested) and observable behaviour (a dependent variable in the experiment) are related and
the first transformed into the latter. They describe typical measures in psychological research, where much of usability research has its roots, to be response time and scaling. With scales, the respondents assign a numeric value “to a property of a psychological state” (Gillan and Bias, 2014).

There are certain distinct measures, that are widely used to represent the ISO (ISO, 2018) aspects of usability (Hornbæk, 2006; Sauro and Lewis, 2012). Hornbæk (2006) conducted an extensive literature review on metrics used in usability studies and identified a large number of metrics that could be categorized under the usual aspects of efficiency, effectiveness and satisfaction. Examples of these metrics include counting the percentage of successfully completed tasks in a usability test to represent effectiveness; comparing the ratio of steps between actual task completion and the optimal solution to represent efficiency (Nielsen, 1993); and using questionnaire responses of users or user rankings of systems according to preference to represent satisfaction (Hornbæk, 2006). According to Hornbæk (2006), subjective and objective methods can both be used to measure usability. The relationship between subjective and objective measures is however unclear, previous research has both found correlation between user evaluations and objective metrics (Nielsen and Levy, 1994; Sauro and Kindlund, 2005; Sauro and Lewis, 2009) and cases where these two different metrics do not correlate (Bailey, 1993; Kissel, 1995; Hornbæk and Law, 2007). There is also evidence that even objective measures for different usability attributes do not correlate (Frøkjær, Hertzum and Hornbæk, 2000). And while there might be a correlation, the information that these metrics provide on the system usability is usually different (Sauro and Kindlund, 2005). Thus, it is recommended that several different metrics are collected to get a better general understanding of system’s usability (Frøkjær, Hertzum and Hornbæk, 2000; Sauro and Lewis, 2012).

As described earlier, usability testing is commonly used for collecting quantitative data to compare systems. Some typical objective measures collected are task completion rate, task time and errors made during a task (Sauro and Lewis, 2012). These are commonly analyzed per task in literature. Sauro and Lewis (2012, chap. 5) give an elaborate presentation of the calculations needed for evaluating whether there is a statistical significance in the difference of these and other metrics when comparisons between designs are made.

While usability testing, and questionnaires, are used commonly for summative purposes, the same does not apply to inspection or expert evaluation methods. These methods, like heuristic evaluation (HE), have been developed for use in formative analysis and their focus is to provide qualitative data on usability problems (Dumas and Salzman, 2006). Even though HE includes
numeric evaluation of the individual identified issues, it does not provide an overall score for usability. Attempts to further quantify HE have been presented (Beuscart-Zéphir et al., 2005; González et al., 2009) for different purposes. As previously presented in subsection 2.2.2, the model of Beuscart-Zéphir et al. (2005) results in an average score for each heuristic, but does not provide an overall score for usability. The basis of the model by Gonzáles et al. (2009) is placing heuristics into categories, calculating weights to each category depending on the number of heuristics and then categorizing the found issues accordingly. This results in one calculated usability score.

There are also concerns related to measuring usability. For example, it has been found that while conducting usability testing properly requires attention to details and the whole procedure, this seems to be lacking in practice in many studies according to research (Dumas and Fox, 2012). In relation to questionnaires, Gillan and Bias (2014) point out that the scales used in usability questionnaires are subject to common systematic sources of distortion in the responses. They outline these to include acquiescence bias (more likely to agree with statements), social desirability bias (inclination to try to look good in relation to others), and central tendency bias (using or avoiding particular values, usually the extreme values of the scale). Unfortunately, these same problems apply also to magnitude estimation which is an alternative to scales (Gillan and Bias, 2014). Gillan and Bias (2014) also state that a typical problem with the use of rating scales, which also applies to using them in usability, is that there are no measurement models provided to explain how people translate their views into numbers. According to them, it has been typical to treat rating scale data as if it perfectly reflects the construct that it measures. They caution not to accept this thinking without further research into how ratings relate to usability (Gillan and Bias, 2014).

Gillan and Bias (2014) also argue that “the definitional approach to measuring usability seems dissatisfying because it appears to place usability in the technological artifact”. They raise the question of whether the definition could vary based on the product, users, or the environment - or is it possible that usability exists solely in the interaction between the user and technology. This leads to viewing usability as dependent on the context in which the interaction occurs. They conclude that user, task and use context should all be considered in an acceptable measurement model (Gillan and Bias, 2014).

Combining Usability Metrics
While most researchers suggest handling the selected usability metrics separate from each other (Hornbæk, 2006), there have also been efforts in
combining usability metrics into a single score (e.g., Sauro and Kindlund, 2005). The combined score of Sauro and Kindlund (2005) includes task completion, error rate, satisfaction, and task time by converting them to z-scores. Z-scores are calculated differently for continuous and discrete data (Sauro and Kindlund, 2005). These metrics were also found to correlate. Sauro and Lewis (2009) also found correlation among task time, completion rate, errors, post-task satisfaction, and poststudy satisfaction. Correlation is important for increasing the reliability of the combined score (Sauro and Lewis, 2012, p. 255).

Another approach to combining the usability metrics is using a rank-based method presented by Lewis (1991). His method allows combining through rank averaging, by putting the different metrics on a common ordinal ranking scale. This approach works for comparing similar systems with similar tasks, and can’t be regarded as a universal usability measure (Lewis, 1991).

It seems that there exists no definitive answer to the problem of combining usability metrics and there are differing opinions among researchers. Sauro and Lewis (2012) conclude that there are situations where the practice of usability evaluation requires one score for usability and combining the measures in a meaningful way makes sense. Gillan and Bias (2014) argue that the many definitions of usability present us with a multitude of questions related to the measurement of usability: Can the attributes of usability measured in a system simply be summed together to get an overall score for usability or should they perhaps be weighted? Are all the attributes relevant for all systems, or can something be declared usable without looking at all of them? Can one attribute compensate for lack of another in some cases? Is it possible that two or more attributes can be multiplied or combined in some other way to get another attribute of usability? The ISO 9241-11 standard (ISO, 2018) gives guidance that the different usability components have relative value depending on the context of use and purpose for usability considerations, and also that at least one measure for each of the three usability attributes in the standard should be included in an evaluation.

2.3.3 Development of Usability Evaluation Methods

Methodological development has not been a very popular type of research in HCI in the recent years (Wobbrock and Kientz, 2016) but it is an important contribution of usability science (Gillan and Bias, 2001). To understand how usability evaluation methods have been developed in the past, this subsection will go through the development of a few well-known methods, heuristic evaluation (Molich and Nielsen, 1990; Nielsen and Molich, 1990) and cognitive
walkthrough (Lewis et al., 1990; Polson et al., 1992), as well as a method combining them, heuristic walkthrough (Sears, 1997). Also, the practices of developing usability questionnaires will be presented.

**Heuristic Evaluation, Cognitive Walkthrough and Heuristic Walkthrough**

The method we know as heuristic evaluation today was first presented in 1990 (Nielsen and Molich, 1990). In their paper Nielsen and Molich recount the development efforts that lead to the method suggestions. Before this development HE was considered an informal analysis made by a specialist to find usability problems. They made four experiments where tens of evaluators who were not usability specialists were asked to list all usability problems in a UI. In three of the experiments the evaluators were coached beforehand on a list of nine usability principles, heuristics, for the evaluation. Molich and Nielsen had developed the heuristics they used in the experiments in a separate research (Molich and Nielsen, 1990). The listed problems were then analyzed against a list of “known” problems identified by usability specialists beforehand to determine the percentage of problems each evaluator identified as well as the average. These lists of known problems were realized to be lacking and were augmented based on the evaluator’s reports before percentages were calculated. Based on quantitative analysis of the evaluators performance, they suggested that HE should be done by three to five evaluators first working individually and that an aggregation phase should be added to the process to combine the found problems (Nielsen and Molich, 1990). The method was developed further to add some details to the procedure, such as how lack of domain expertise might be solved, how to go through the interfaces with the heuristics and who should do the evaluation (Nielsen, 1993). Severity ratings were also introduced later into the aggregation stage (Nielsen, 1993). The original heuristics were based on personal experience of authors on good interface design while also corresponding to principles described elsewhere (Molich and Nielsen, 1990). These specific heuristics have since also been through several iterations (Cockton et al., 2012), for example in Nielsen (1994) by analyzing usability problems to identify what principles could have predicted most of them.

The first version of Cognitive walkthrough (CW) was introduced likewise in 1990 by Lewis et al. (1990). The method was developed to bring a realistically applicable way of utilizing cognitive theory in evaluating UIs. The questions asked in CW were based on a theory of learning and subsequent design guidelines developed by Polson and Lewis (1990) and the procedure on review practices of software engineering (Lewis et al., 1990). The reasoning for developing
CW as stated by the authors was: “a guideline’s applicability or the trade-offs between conflicting guidelines are resolved by evaluating the design using the theory that motivated the guidelines” (Lewis et al., 1990). This version of CW includes 18 questions to ask for each step of the interaction sequence. CW was assessed by comparing the results of an evaluation with those from user testing (Lewis et al., 1990) to determine how effective it was in predicting the problems. In the experiment four versions of the same interface were evaluated with two tasks by four experts who also numerically rated the extent of the problems. These results were compared to previously done user testing with at least 15 users for each interface option. The comparison study did not result in any modifications to the method; however, it was observed that deep knowledge of the underlying theory affected the ability for predicting problems and training was seen necessary for evaluators (Lewis et al., 1990). A later paper (Polson et al., 1992) extends the theory with other models, modifies and elaborates on the details of how to conduct the evaluation and apply the theory. In this paper CW was also theoretically compared to other evaluation methods, such as GOMS, guidelines and user testing. This version was identified as complex both by the original paper and others (Mahatody, Sagar and Kolski, 2010). The complexity concerns resulted in a later version of CW where the questions were reduced down to four (Wharton et al., 1994); this is considered the current version.

Heuristic walkthrough (Sears, 1997) is a method first presented in 1995 that combines HE, CW and a third method, usability walkthrough by Karat et al. (1992). The method was developed by identifying the known shortcomings of each method and then combining parts from each to a new method; unstructured evaluation and heuristics from HE, using tasks and questions from CW and the idea of two evaluation rounds from usability walkthrough. A simulation experiment was conducted to compare three of these methods for their performance in real-like conditions when evaluating a paper design to results from previously conducted user testing on a live system (Sears, 1997). The study was done as a between-subjects design, where twenty evaluators were trained on all three techniques but each used only one in a time-limited session (Sears, 1997). The results of the study were not used to modify the HW method but rather to give guidance on the selection of each method and more insight into their pros and cons.

Cockton et al. (2012) synthesize usability inspection methods to have four stages: “analyst preparation, candidate problem discovery, confirmation or elimination of candidate problems, and problem reporting”. Based on this, it can be deduced that the development and method description should address
all these stages. However, the preparation stage has not generally been the focus of research (Cockton et al., 2012).

**Evaluating Developed Methods**

Generally, in usability literature it has been thought that any evaluation method is only as valuable as its quality (Woolrych et al., 2011). Hartson, Andre and Williges (2003) have elaborated on this topic in the context of assessing the quality of formative usability evaluation methods. To assess the quality of a method an ultimate criterion for its goodness needs to be identified based on the purpose. However, as the ultimate criterion can’t be measured, instead one or several actual criteria to represent it needs to be chosen (Hartson, Andre and Williges, 2003). The purpose of formative usability evaluation methods is to identify usability problems and produce descriptions of them and they should be evaluated against this goal (Hartson, Andre and Williges, 2003). Hartson, Andre and Williges (2003) have identified the following actual criteria as the performance measures for these methods: thoroughness, validity and reliability from (Bastien and Scapin, 1995), as well as effectiveness (combination of thoroughness and validity), cost effectiveness and downstream utility. Interestingly, they argue, that in the context of formative usability evaluation methods, the ability of these methods to actually evaluate usability should not be compared, because they generally don’t produce user performance data (Hartson, Andre and Williges, 2003). It’s clear they consider evaluating usability to mean evaluating its attributes.

The ability of the evaluators to find problems was analyzed in the original paper (Nielsen and Molich, 1990) with percentages, distribution, and how many evaluators found individual problems. Later, also the impact of the evaluators’ experience on finding problems was analyzed in a separate experiment (Nielsen, 1992). When evaluating the quality of HE, Nielsen (1994) used Kendall’s coefficient to measure interrater reliability on severity ratings. Later, also other researchers have evaluated HE from different perspectives (Cockton et al., 2012).

The original evaluation of CW (Lewis et al., 1990) analyzed the ability of the evaluators in predicting problems, similarly to HE. This was done by evaluating the consistency of the evaluators in predicting the same problems and comparing the identified problems in CW to action trees of the taken paths from user testing by counting percentages. The different versions of CW have been evaluated also by several other studies (Mahatody, Sagar and Kolski, 2010).
Sears (1997) proposed specific metrics for the three criteria of evaluating method quality, reliability, validity and thoroughness (Bastien and Scapin, 1995), and performed an experiment to compare HE, CW and heuristic walkthrough with these metrics. Reliability was measured by “the ratio of the standard deviation of the number of problems found to the average number of problems found”, validity by “the ratio of ‘real’ usability problems identified to all issues identified as usability problems” and thoroughness by “the ratio of real problems that are identified to the number of problems that exist in the system” (Sears, 1997).

Woolrych et al. (2011), however, argue that this tradition of evaluating the individual methods’ quality is actually not appropriate, especially when it comes to usability inspection methods, and instead the resources (such as procedural, reporting, task selection) that different methods use and their content and role in the performance should be assessed individually. This is because according to their view these methods are merely approaches and do not and cannot prescribe every detail related to the evaluation. Every approach is a collection of resources and requires adaptation to the circumstances (Cockton et al., 2012). Also, it has been proven that when using HE usability specialists rely on different kinds of knowledge resources, distributed cognitive resources, beyond what the method description of HE entails (Woolrych, Cockton and Hindmarch, 2005). Additionally, Cockton et al. (2012) view that focusing on isolated methods in research does not reveal the outcomes of usability work, and research should also focus on how the methods are selected and used together.

Usability Survey Development and Evaluation

User evaluation surveys or questionnaires are one type of survey typically used in HCI and Ozok (2012) has outlined the practices and considerations for their formulation and development. The goal of these surveys is evaluating usability and user satisfaction from the user perspective. These types of surveys are used immediately after the use of a system in experimental settings to ensure that the participants recollection of the situation is still vivid (Ozok, 2012).

The design of the questions and their rating scales is an important part of the process (Ozok, 2012). The selection of questions should be based on previous literature, but can also be based on experts and individual opinions (Ozok, 2012). Both qualitative and quantitative data can be collected with the questions. Pilot testing is recommended to develop the survey before larger implementation (Ozok, 2012). It is typical, that the survey needs modifications after this stage where the understandability of the survey is assessed.
After the first version is designed, it should be evaluated to ensure the validity and reliability. Survey validity in HCI is evaluated usually by construct validity and predictive validity (Ozok, 2012). Construct validity is proven by providing the reasoning for the questions and their basis in previous research. Predictive validity is addressed by demonstrating that the survey tool reflects usability accurately both in the case it was developed for as for other participants and evaluation subjects. Both these evaluations are qualitative in nature. For evaluation surveys, reliability is determined by internal reliability while the common measure of inter-rater reliability is not as useful in this case as participants views on usability can differ (Ozok, 2012). Internal reliability is evaluated usually with Cronbach’s alpha which determines if the same questions are answered in a similar way (Ozok, 2012).

Other important questions in survey implementations are the environments in which they are administered (controlled or open) and the sample representativeness (Ozok, 2012).

2.4 Synthesis

Public procurement has several stages and one of them is the purchasing stage which includes the selection phase (Moe, 2014). The purchasing process is strictly regulated within the European Union (2014) and its member states’ national legislation. The key components of this legislation are equal treatment, fairness, and non-discrimination of vendors, as well as transparency. This is realised with upfront declaration of criteria so that vendors are aware of criteria and how comparisons are done. Both price and quality are typically considered for selecting economically the most advantageous tender (EMAT). The criteria must be such that objective comparison is possible, and for EMAT the criteria must be measurable (Finland, 2007). There are different types of procurement processes defined in the legislation (Finland, 2007, 2016) of which the negotiated procedure is of interest in relation to this thesis.

Including Usability Evaluation in Procurement

In the context of usability of procured existing products, the development of the system further might not be an option, so focusing on the selection phase seems key. Traditional procurement practices, such as demonstrations or reviewing functional requirements, do not give a good enough understanding of the usability of the product (Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014). Usability evaluation is seen to prevent unusable systems from being selected and support user involvement (Schumacher, Webb and
Johnson, 2009; Kushniruk et al., 2010) which is an important success factor for procurements (Moe, 2014). The legislation for the negotiated procedure in public procurement allows the selection to proceed in stages (European Union, 2014; Finland, 2016). This is also the recommended procedure for usability evaluations based on research on IS procurement in the field of healthcare (Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014).

The two-stage procedure includes a preliminary assessment phase to reduce number of vendors and a short-list evaluation phase to score the products based on numeric goals (Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014). The literature on procurement acknowledges that some identified criteria can be used also as minimum quality standards (Mateus, Ferreira and Carreira, 2010; Bergman and Lundberg, 2013). This is similar to the argument of including usability as a minimum requirement, which can be used in the first preliminary assessment phase. This phase could also include assessing relative usability of the product (Schumacher, Webb and Johnson, 2009). The planned procedure should support both these goals (Beuscart-Zéphir et al., 2005), but other factors should also be considered (such as previously mentioned user involvement).

Kushniruk et al. (2010) have presented a framework to aid in choosing evaluation methods to utilise in this process, so that strong enough evidence on usability is received. These include using heuristic evaluation and usability testing with clinical information processing scenarios (CLIPS). The first is suggested for use in the preliminary assessment phase by Kushniruk et al., while the latter is recommended in the more specific short-list evaluation. Standardized questionnaires are also recommended (Kushniruk et al., 2010). Case examples of usability evaluation in health IS procurement typically utilize multiple methods and both subjective and objective measures. It should be noted, however, that user preference can be in conflict with the actual efficiency and not necessarily entirely related to the product itself (Lauesen and Younessi, 1998).

In this context the purpose of the evaluation is clearly summative and not formative. The suggested methods are established in the general literature on usability evaluation but were developed and are mainly used in a formative evaluation context for gathering qualitative data. While user testing and questionnaires are also used for summative purposes, this discrepancy was not clearly outlined in the literature. These methods also have other drawbacks. Traditional heuristic evaluation doesn’t include context and has a limited scope of evaluation (Chin, Diehl and Norman, 1988; Nielsen, 1993),
questionnaires suffer from biases in the scales (Gillan and Bias, 2014) and usability testing in a controlled setting might not be representative of actual use (Siegel, 2012) and it is also resource intensive (Riihiaho et al., 2015). There are established measures for measuring the ISO definition attributes of usability, usually utilizing usability testing (Hornbæk, 2006; Sauro and Lewis, 2012). In usability testing, the metrics are usually analysed per task in the literature (Sauro and Lewis, 2012) but idea of averaging over all tasks per system is presented in relation to procurement (Schumacher, Webb and Johnson, 2009). Heuristic evaluation is not usually quantified, as it has been developed for a formative qualitative evaluation purpose, although some attempts exist. For example, Beuscart-Zéphir et al. (2002, 2005) quantified heuristic evaluation in a simple way by calculating a mean score for each heuristic. This way of calculation treats all heuristics as equal and dissolves the possible differences in the number of problems between systems. It also lacks a way to create one unified score for each system.

Research on procurement also highlights that to ensure comparability of tenders while measuring, the information required from the vendors must be detailed carefully (Mateus, Ferreira and Carreira, 2010). When it comes to usability evaluation specifically, the planning phase is essential to define a strict repeatable procedure that is vendor neutral in how the tasks are written (Jensen, Rasmussen and Lyng, 2013). Also complex information systems, such as those in healthcare, require careful selection of scenarios, critical tasks and user groups for the evaluation as not nearly all can be covered (Schumacher, Webb and Johnson, 2009; Jensen, Rasmussen and Lyng, 2013). This selection requires domain knowledge, and the procedure needs to be representative enough (Jensen, Rasmussen and Lyng, 2013). The adequate training of the users should also be considered before evaluation (Jensen, Rasmussen and Lyng, 2013).

Scoring Rules in Public Procurement and Quantification of Usability as a Criterion

Legislation requires that the selection criteria and their relative weights are disclosed in advance but declaring the actual measures and scoring rules is not required in the beginning (European Union, 2014). However, researchers argue that determining correct weighting requires starting with planning the detailed scoring rules of each criterion (Mateus, Ferreira and Carreira, 2010). Unfortunately, EMAT with combining both price and quality makes public procurement complex and according to Mateus et al. (2010) this often results in defectively constructed overall scoring rules. They state that this is the result from not understanding the mathematical complexity and using the
“common sense” approach of determining the weights by their seeming relative importance without considering the underlying scoring for each criterion. Using price-to-quality scoring rules (as opposed to quality-to-price scoring), where price is transformed into numeric value to be assessed in one equation with quality, is the common approach in public procurement cases (Bergman and Lundberg, 2013).

The weighted sum is the most used method in price-to-quality scoring rules for combining the results from the different criteria and price to one overall score (Pictet and Bollinger, 2003; according to Mateus, Ferreira and Carreira, 2010). Discussion of using the weighted sum is found in the literature on quantifying usability and combining measures from different attributes (Gil-lan and Bias, 2014), however the general recommendation is to deal with the measures as separate (Hornbæk, 2006). There is not much concrete research on combining the measures outside the unified score of Sauro and Kindlund (2005) and the rank-based approach of Lewis (1991). In the procurement literature, similarly than different usability attributes, dividing a criterion to sub criteria is discussed for it to be measurable (Mateus, Ferreira and Carreira, 2010). For each criterion, a performance measure is defined, and a scoring function/rule is used to convert the collected measure into a partial score on a numeric scale that describes the level of performance on that measure (Mateus, Ferreira and Carreira, 2010).

The definition of the set performance levels is a key part of the planning of scoring and should reflect how desired each level of performance is to the procurer in the specific case and what is the difference compared to the other levels (Mateus, Ferreira and Carreira, 2010). The numeric difference between all levels is not necessarily equal when the planning is done from the desirability perspective. The resulting scoring function, depending on the type of measure, can be either directly related to the predefined levels or defined with continuous or discrete mathematical expressions (Mateus, Ferreira and Carreira, 2010). Mateus et al. (2010) present available methods to aid in composing these expressions to reflect the procuring organisation’s preferences (which in mathematical terms can be for example linear, exponential, or piecewise linear with continuous measures).

When using the weighted sum method for the overall score, the defined weights imply a trade-off relationship between the criteria and the weights should be determined by considering this related to the procurer’s preferences (Mateus, Ferreira and Carreira, 2010). In the literature on measuring usability, the idea of whether all usability attributes are equal or whether their importance depends on the system, its users and its context is raised (Gillan and
Bias, 2014). An adequate measurement model for usability should account for all these aspects. This implies that similar methods to determine the weighted sum of usability attributes “inside” the usability criterion could be used than are presented in the general procurement literature. Determining the weights requires using likely examples of possible tenders and their performance levels for each criterion and then deciding where these tenders should fall on the overall score scale (Mateus, Ferreira and Carreira, 2010). Again, there are methods available to aid in composing technically sound weighting which requires more mathematics to solve the problem (Mateus, Ferreira and Carreira, 2010). However, it should be noted that there is no consensus on whether the weighted sum is a suitable representation of usability and whether there exist trade-off relationships between the attributes.

Challenges and Knowledge Gaps
There are several challenges identified with public procurement in general and usability evaluation in particular, especially within the healthcare context. Some challenges include:

- the requiring of knowledge and expertise from numerous areas (Moe, 2014) (even within the selection phase and one criterion, usability, as can be seen from the literature);
- the lack of required expertise in the projects (Moe, Risvand and Sein, 2006) which also applies to including usability (Artman and Zällh, 2005);
- need to specify many details upfront in the process (Moe, 2014);
- the usability evaluation procedure needs to be comparable, uniform and transparent which is not a common scenario (Jensen, Rasmussen and Lyng, 2013);
- there is a very limited time available for the planning, conducting and analysis of results (Jensen, Rasmussen and Lyng, 2013);
- problems with scoring rule definition, which can result in arbitrary results (Mateus, Ferreira and Carreira, 2010);
- public organisations typically having complex IS environments (Moe, 2014);
- multiple stakeholders with their possibly contradicting goals and needs which are always unique to the case (Moe, 2014; Alanne et al., 2015); and
- withing healthcare the above contradiction is typically found even between several user groups of the same software (Rantanen and Heimo, 2014; Alanne et al., 2015).
There are also identified knowledge gaps in the research on the topic of public procurement. Overall, there is a lack of research on the topic and especially not enough research based on empirical data (Moe, 2014). Based on the related research, it seems clear that there is a need for more detailed descriptions of usability evaluation during public IS procurement (Jensen, Rasmussen and Lyng, 2013) and new methods specifically targeting measuring usability, also without involving users. While there are several examples in literature on including usability evaluation in healthcare IS procurement (e.g. Beuscart-Zéphir et al., 2005; Carvalho, Borycki and Kushniruk, 2009; Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnos-sen and van Ooijen, 2014), only few include the specific context of public procurement. Overall, it also seems that the HCI community needs to put more emphasis on understanding the procurer perspectives and develop usability methods suitable for them (Artman and Zällh, 2005).

**Developing Usability Evaluation Methods**

The previous method development in HCI can inform the development of new usability evaluation methods that procurement projects need. HE and CW were both developed based on existing resources, design heuristics and theory on learning, respectively, to address shortcomings of previous approaches to evaluation (Lewis *et al.*, 1990; Molich and Nielsen, 1990; Nielsen and Molich, 1990; Polson *et al.*, 1992). Heuristic walkthrough was developed to combine the positives of both these methods to perform better (Sears, 1997). After the initial method development in all these cases, experiments were conducted and reported on applying the methods and used to evaluate their performance. Usability inspection methods seem to consist of similar stages as identified by Cockton *et al.* (2012): “analyst preparation, candidate problem discovery, confirmation or elimination of candidate problems, and problem reporting”. Thus, when developing inspection methods all these aspects should be described. For surveys, the questions for user satisfaction from the user’s perspective are developed based on previous literature and expert opinions (Ozok, 2012).

Survey design has been identified to follow a design, redesign and revision cycle (Ozok, 2012). A similar pattern does not seem to be reported for individual inspection method’s development in the literature, but the overall arch of the evolution of HE and CW to their currently established form as presented earlier does seem to follow this pattern of iteration.
After developing a method, its quality should be evaluated. HE and CW have been evaluated with quality criteria such as reliability, validity and thoroughness (Nielsen, 1994; Sears, 1997). Evaluation surveys on the other hand are usually evaluated in HCI by their validity and reliability (Ozok, 2012). It has also been argued that usability evaluation methods should not be evaluated in isolation for their performance, but studies on the individual resources used by them and of combining them in real work contexts should be done (Woolrych, Cockton and Hindmarch, 2005; Cockton et al., 2012).
3. Action Design Research: Developing Usability Assessment

In this chapter, I present the research approach and study design. The appropriate research paradigm and approach for this dissertation work is discussed in section 3.1. In section 3.2, I present the overall research design. Sections 3.4 and 3.5 present the action design research study with an iterative process for method development and their implementation in the procurement project. These sections present the methods and materials for answering research questions. Section 3.6 concludes the chapter by discussing the quality criteria for this type of research.

3.1 Research Paradigm and Approach

The research in this dissertation falls under the disciplines of human-computer-interaction (HCI) and/or usability science (Gillan and Bias, 2001), but is also closely related to the research fields of information systems (IS) research and health informatics. HCI is a relatively new field with its roots in other fields of study, such as software engineering, psychology and human factors (Lazar, Feng and Hochheiser, 2017, chap. 1). The different research orientations within HCI can be thought of as communities with their own theories, methods, practices, and interests that they mostly draw upon, such as computer science, sociology, human factors engineering and management information systems (Lazar, Feng and Hochheiser, 2017, chap. 1). Researchers have also suggested usability science as a separate field from HCI to bridge the gap between cognitive science and human factors from practical usability engineering (Gillan and Bias, 2001). I have received my formal education in computer science and engineering with a focus on UCD. I believe this background has influenced my interests and way of approaching research as a practical effort.
As discussed previously in Chapter 1 *Introduction*, this research is motivated by the practical problem of measuring usability in a comparative way during IS procurement. A research philosophy and paradigm that is concerned with practical problems and actions to solve them is pragmatism. According to Goldkuhl (2004) "pragmatism means an interest for actions in their practice context" and also "an interest for practical consequences of knowledge". But it also means "an interest in what works and what does not work" (Goldkuhl, 2004). This is a good fit for the way I approach science as a researcher. Gillan and Bias (2001) identified usability science as a pragmatist science.

Wobbrock and Kienz (2016) have divided the research contributions of HCI research into seven categories of which “methodological contributions” is one, while not the most popular type of research in the field. Gillan and Bias (2001) proposed that there are five types of issues that usability science is concerned with, one of which is also development and adaptation of methods. They suggest that methods can be considered to fit in a multidimensional space, where the dimensions are the origins of the method, the purpose of the method and the point in which its used in the design cycle. Accordingly, these dimensions can be used to identify opportunities for new methods and develop them. And once developed, determining their validity is a key contribution of usability science (Gillan and Bias, 2001). Usability engineering methods research was dominant in the late 1980s and 1990s and while the current focus is no longer on efficiency of use, it is however on the experiences and emotions of people towards technology and the context they will use it in (Lazar, Feng and Hochheiser, 2017, chap. 1). The research problem of this dissertation also encompasses these themes in relation to the context of use of the IS being procured.

Goldkuhl (2004) states that the pragmatist view is a possible explanation and justification for application of action research (AR) strategies and methods development in IS research. AR is focused “on research in action rather than research about action” (Coughlan and Coghlan, 2002). Kock (2014) also sees AR as an approach that could be used more in HCI. The objective in AR is foremost understanding an organizational issue and second developing a solution to this context instead of generalizing knowledge (Hayes, 2011). The goal is not to attain the best solution, but to gain better understanding through the action and learning from the outcomes to contribute to scientific knowledge (Coughlan and Coghlan, 2002; Hayes, 2011), but also to guide practice (Järvinen, 2007). In actuality, AR is an umbrella term for a variety of research method used in different disciplines with common characteristics (Herr and Anderson, 2005, chap. 1; Kock, 2014). Depending on the tradition the positionality of the researcher (insider-outsider) varies (Coughlan and
Coghlan, 2002; Herr and Anderson, 2005, chap. 1), but most traditions agree that “action research is inquiry that is done by or with insiders to an organization or community, but never to or on them” (Herr and Anderson, 2005, chap. 1). This collaborative role of the researchers, along with the focus on changing action and practice are the key characteristics of AR (commentary by Munkvold in Kock, 2014). The description of the AR research process that is most referenced is by Gerald Susman and Roger Evered devised in 1978 (Kock, 2014). This is a cyclical process of five phases, “diagnosing, action planning, action taking, evaluating and specifying learning”, with iterative cycles either in the same context or another (Kock, 2014). These iterations can produce better solutions over long periods (Hayes, 2011).

Another paradigm used in IS research is design science which ”supports design research” (Collatto et al., 2018). The research method used with this paradigm is usually design science research (DSR) in which artifact design, building and application to gain knowledge of a problem is used (Hevner et al., 2004; Collatto et al., 2018). While AR in IS has its background in the socio-technical design movement, DSR has its based in engineering (Iivari, 2007). In DSR artifacts are constructs, models, methods or instantiations that rely on “research rigor” in utilizing existing knowledge in both developing and evaluating them (Hevner et al., 2004). The problems DSR tries to solve are typically formulated by the researchers (Järvinen, 2007), although solving practical problems inspire DSR (Hevner et al., 2004). However, developing and testing of the artifacts is typically done in laboratory settings without connection to practice (Iivari, 2007). The artifacts are evaluated by their utility, quality and efficacy (Hevner et al., 2004). While developing and testing have been identified as the main activities of DSR (Järvinen, 2007), sometimes the evaluation is even postponed and separated (Sein et al., 2011). This results from putting more emphasis on “technological rigor at the cost of organizational relevance” and not understanding the importance of the interaction of the artifact and context even if the origins of the artifact are researcher driven (Sein et al., 2011).

The objectives, evaluation, researcher’s role and need for an empirical base are the substantial difference of AR and DSR (Collatto et al., 2018). Although some insist that design science and action research should be treated as completely separate as their epistemological basis are different with design science having a positivistic view (Iivari, 2007; Iivari and Venable, 2009), others have viewed them as similar (Järvinen, 2007). However, even those who view them as strictly separate conclude that artifacts developed through design science research can be implemented in a real context with action research, preferably
after testing in a laboratory setting first (Iivari, 2007). Collatto et al. (2018) propose based on a literature review three different approaches to using both AR and DSR: either using AR by “taking into account the paradigm of design science”, or using DSR and AR in sequence as a mixed-methods approach, or for an instantiation-type of artifact combining them into action design research. They describe action design research to include “collaboration between the researchers and the participants since the artifacts conception up until its evaluation”, the iteration of development and evaluation to be repetitive and its purpose to be generating “knowledge about the implementation of the artifact in the real context”. Sein et al. (2011) originally proposed the research method combining AR and DSR to be called “action design research” (ADR), which joins the focus on solving practical problems by using interventions with the members of the organization with the focus on building and evaluating IT artifacts. They describe the ADR process to entail three phases that are iterated until formalization of the learning is done with generalized outcomes. The three iterative phases are: first problem formulation, second building, intervention, evaluation (BIE) and third reflection and learning (Sein et al., 2011). The key aspects in the ADR process are that the research is inspired from practice, the artifact is created with a theoretical approach, the BIE cycle should include influence from all stakeholders and the research done should be generalized (Sein et al., 2011).

These views of AR, DSR and ADR can be used to inform the research done in this dissertation. While ADR has been proposed for IS research on IT artifacts, it seems the principles and process of combining AR and DSR could be adapted also for method artifacts.

3.2 Research Context

This dissertation research was done in relation to a public procurement project where several municipalities and a hospital district aimed to procure an integrated health and social welfare IS together. In the project the negotiated procedure, based on the law in effect at the time in Finland (Act 348/2007), was used. The selection was based on EMAT with price-quality evaluation and the scoring utilized several criteria including usability. A project office was established specifically for the project and was part of the organization of the largest municipality involved in the procurement. The product comparisons for the procurement were planned in a subproject with approximately ten team members. The team members were also responsible for executing the
plans. The planning of the usability evaluation, measurement and scoring was done in this subproject.

The research described in this dissertation was initiated when the project sought for help in including the usability perspective in the procurement, and as a result one usability researcher from academia was employed by the project. Once a review of previous literature identified a lack of research and existing methodology and methods for measuring usability to be readily implementable as is in the project, the research problem presented in section 1.4 was identified. One of the central reasons for needing new methodology for the procurement, were the vast array of use contexts from both outpatient and inpatient healthcare to the different services of social welfare that the procured IS should cover, as well as the great variety of tasks performed by a multitude of professionals as future end users. Traditional evaluation methods used to quantify usability (such as usability testing) would have required an enormous number of test topics and test sessions which would not have been feasible from a resource or timeline perspective.

After the first drafts of suitable methods based on established ones were developed, I was approached to participate in the piloting of the methods as an outsider usability researcher. Within a few months, I joined the project as a part-time employee working in the product comparison subproject. My role was to participate in the further development of the evaluation methodology as well as its implementation in the project.

Several aspects of the above described setting are definitive factors for an action research project (Herr and Anderson, 2005). First is the nature of the research problem being entirely based on practical issues. Second is the need to develop current procurement practices to serve the problem and employ them in a real-world setting (need for action). And third is the role of the researchers as being actively collaborating partners with other insiders working in the organizational setting studied as opposed to being outside observers (positionality of researchers).

At the same time, the research problem in question called for the development of new methodology and methods (artifacts) to solve the problem and to utilize them in practice to demonstrate them being a satisfactory solution (evaluation). Existing literature was to be utilized to design the artifacts (research rigor). The motivation from the procurement project and research sides was also on the developed solutions being applicable in other similar types of procurement projects (generalizability to a certain class). These are all aspects identified as design science research (Hevner et al., 2004; Järvinen, 2007; Collatto et al., 2018).
This combination of aspects from both action research and design science research required the ADR strategy. The process of starting with “problem formulation” from a practice perspective with a theory based artifact, then “building, integration and evaluation” with reciprocity and mutual influence of all members, followed with “reflection and learning” and “finally formalization of learning” with generalizing the outcome (Sein et al., 2011) is followed in the research design of this dissertation.

3.3 Research Design

As stated above, I employed a research strategy combining action research with design science research in my dissertation. This strategy is similar to that discussed above as ”action design research” (ADR) (Sein et al., 2011; Collatto et al., 2018), and this term will be used in the rest of the dissertation for the employed strategy. The time-horizon of this employed strategy can be considered longitudinal.

As described in the previous section, I held a dual role as an employee in the procurement project and as a researcher with the objective of simultaneously solving the research problem. In AR the positionality of the researcher can vary from outsiders doing research with insiders, outsiders joining an organization for the research and complete members of the organization doing research (insider researcher) (Coghlan and Casey, 2001; Brannick and Coghlan, 2007; Nosek, 2007; Coghlan, 2019). The research presented in this dissertation was collected along the journey of my transformation from an outsider to an insider as the employment started around the same time as starting the research. After years of working in the organization, I have transformed into an insider. This shifting of positionality as a researcher has presented challenges and opportunities for the analysis and contributions presented in this dissertation. The challenges of insider research are typically identified as pre-understanding, role duality, and organizational politics while having access to materials is identified as an opportunity (Coghlan and Casey, 2001; Coghlan, 2019). Relevant to this dissertation work are role duality, insider access to materials and gained deeper understanding.

The ADR research done followed the timelines and overall phases of the procurement project. First, existing literature and knowledge was searched. Then the first versions of the methods were developed, and they were piloted in the summer of 2013. After the first piloting, the methods were further developed and piloted, the overall evaluation framework was developed, and the scoring
was planned in detail. The first phase of the product comparisons was executed in spring of 2014, and the first two developed usability evaluation methods were used. The final piloting for the second phase of the product comparisons happened in early autumn of 2014 and the second phase of usability evaluation was executed in late autumn. The final analysis and reporting on the usability evaluation methodology was done from 2015 to 2019 in scientific publications (Papers I-V). Figure 3 presents the timeline of the dissertation research: the ADR research followed by a complementary survey study which I will present next.

During the procurement project, literature on the usability of HIS and their procurement was utilized when developing the methods and evaluation procedure. In the project, the same usability evaluation methodology covered also the system procured for social welfare. However, this decision was done without prior knowledge on comparability between the two contexts of use or studies on usability of CIS. The lack of knowledge on the user experiences of social welfare professionals on their existing CIS and how these experiences related with those of healthcare professionals led to the development of a separate national survey study that was implemented in the year 2019. After the procurement selection phase was over, I participated in the adaptation of the validated NuHISS instrument (Hyppönen et al., 2019) into the social welfare context as part of a multidisciplinary research group (Ylönen et al., 2020). An analysis of this survey together with earlier answers of physicians and nurses to similar surveys in the healthcare context was done in years 2020-2021 and
is reported in Paper VI published in 2022. At the time of the surveys the procured system was not yet in use by the surveyed user groups. The paper presents results from six user groups: outpatient physicians and nurses working in primary care and specialized care as well as social workers and other social welfare professionals working in open services settings. This paper is utilized in this dissertation as complementary information to the action design research project.

All the published papers and their relation to the dissertation research described above and the scope of the thesis (presented in section 1.4) is summarized in Figure 4.

### 3.4 Building the Artifacts: Iterative Methods Development

The actions in this methods development phase are outlined in Figure 5 (next page) and discussed in more detail below. The literature search and the development of the first versions of the methodology framework and methods were conducted before the author of this dissertation joined the project. The literature that was used during the project is presented as part of Chapter 2 *Theoretical foundation*, specifically in sections 2.2 and 2.3. However, these sections also include additional references. One of the main challenges at this stage was to devise an appropriate way to present the results for two different comparison purposes, i.e., to ensure a minimum level of usability and to rank the vendors in the first phase of product comparisons as well as to score usability for the final selection in the second phase. Moreover, efficient time and
resource allocation for those involved in the selection process needed to be considered. Therefore, the same sessions were used for evaluating several aspects of the products. Methodology for evaluating for a broad spectrum of functionalities during vendor demonstrations was thus needed, especially for the purpose of narrowing down the number of vendors in the first phase of the selection process. The more resource intensive and focused usability evaluation that could possibly give stronger evidence on usability but allowing the assessment of a narrow spectrum of functionalities, i.e., usability testing (Kushniruk et al., 2010), was planned for the second phase of selection. In this phase fewer vendors would remain and the minimum level of usability for a large scope of use contexts had been assessed.

In June 2013, the initial versions of the methods for product comparisons were piloted with systems that were in use at the time. My role in these sessions was an outsider usability researcher. The feedback and experiences from this pilot were analyzed and considered when further developing the methods for first phase of product comparisons in spring 2014. After the first phase of product comparisons, the usability evaluation methodology and scoring for the second product comparison phase were further developed and piloted. The second phase took place in late autumn 2014.

The usability evaluation method development was led by the other usability researcher and the author of this dissertation. This included planning the detailed methods for assessing usability, the metrics used for measuring usability attributes, and the scoring and weighting of these attributes. The selection of use contexts, user groups and their tasks used in the usability assessment was led by the domain experts working in the procurement office. The scale of the procurement was very large i.e., it covered both healthcare and social

Figure 5. The iterative methods development phase.
welfare in a multitude of use contexts and user roles in areas such as labor and
delivery, operating theatres, emergency departments, intensive care units,
child welfare services, and home care, and the timeline of the procurement
was strict. Thus formal user research conducted by usability researchers was
not considered a feasible approach since this was estimated to have resulted
in hundreds of different user scenarios. The domain experts run workshops
with user representatives for the identification of the appropriate user scenar-
ios for the demonstrations and the selection of tasks and creating the initial
versions of test scripts for the second phase.

The procurement project was conducted with the negotiated procedure.
Some constraints for the usability evaluation process were set by the project.
Such constraints were the timeline, tender scoring with a price-to-quality
scoring, the weight of each criterion, organizing the product comparisons in
two phases, the overall goals of these comparisons, and defining health and
social welfare professionals as well as citizens as the overall user groups for
the evaluation. Other criteria evaluated in the product comparisons based on
seeing the offered information systems in use were functional quality and
scope as well as configurability of the products. The practices planned for
these other evaluations influenced the choices made for usability evaluation,
for example using scenario-based demonstrations. The requirements of the
legislation on public procurement for equal treatment of vendors, objective
criteria, outlining the evaluation beforehand, such as announcing the mini-
mum requirements in the very beginning, and planning the procedures for at
least three vendors were to be followed in the planning of the usability evalua-
tion.

3.5 Intervening and Evaluating: Implementing the Methods

The actions in this implementation phase are outlined in Figure 6 (next page)
and discussed in more detail below. The developed usability evaluation meth-
odology and methods were implemented in the procurement project in 2014
in two phases. The overall evaluation framework for both phases is described
in Paper I. In the first preliminary assessment phase there were four compet-
ing vendors. In the second short-list evaluation phase the number of vendors
had been reduced to two.

Preliminary Assessment Phase
The goal of the preliminary assessment phase was to ensure the vendors’ sys-
tems fulfilled minimum quality requirements and to rank the vendors in order
to reduce the number of candidates. The evaluation in this phase was based mainly on comprehensive user scenarios which described the work of typical user groups, their context of use as well as their key tasks. In total, nine scenarios were selected for the evaluations that were considered to cover the main contexts of the client and patient information system use. The vendors were required to build the demonstration of their system to follow these scenarios. The demonstrations took 3.5 days for each competing vendor.

The usability evaluation utilized the developed methods: heuristic evaluation during demonstrations (HED), demonstration based perceived usability questionnaire (DPUQ) and a more traditional heuristic evaluation that followed same scoring practice as the HED method. During six of the demonstrations (for 2.5 days per vendor), I together with the other usability researcher applied the HED method. In addition, we performed a more traditional heuristic evaluation for one area of the system separately. Future professional end-users filled in the DPUQ questionnaires during all nine demonstrations. The role of these methods as part of the overall procedure can be seen in Paper I. Papers II and III describe the new methods developed for this phase in more detail.

**Short-List Evaluation Phase**

The goal of the short-list evaluation phase was to score the vendors on each quality criterion for the final selection. The usability evaluations in this phase were mainly based on user testing scenarios. There were six scenarios for user testing, five for professional user groups and one for citizens as future users of the client/patient portals. The method used for professional user groups was paired-user testing and for citizens traditional usability testing. Several usability metrics were measured with user testing. Additionally, group
evaluations of vendor demonstrations with users (DPUQ) and rating the usability of the system by a usability researcher were used.

The facilitation of the user testing sessions was split between the usability researchers, in total 15 paired-user testing and 10 usability testing sessions were organized per vendor. The group evaluation sessions were likewise split between the usability researchers. The role of all these methods as part of the whole can be seen in Paper I. Papers IV and V describe the user testing, mostly paired-user testing for healthcare professionals.

*Appeals to the Market Court*
During both phases of the selection process the decisions were appealed to the Market Court. These appeals did not concern or affect the implementation of the product comparisons for usability described in this dissertation. The market court decisions did not thus find any problems with the usability evaluation process or methodology.

*Data Collection and Analysis*
During the product comparison subproject detailed documentation was created by the usability researchers and other experts working in the project, which described the decisions, plans and the execution of the product comparisons. This was done both for the project itself, for practical use later and for research purposes. Due to the nature of a public procurement project having to be upfront, transparent, and equal towards all tenderers, there was an ample amount of documentation produced during the project. This documentation was available for the reporting and analysis of this research. This dissertation and the research articles included utilized this documentation (written reports and documents as well as numerical data) when reporting on the usability evaluation methodology, methods, and analysis of their use. This included all documentation from using HED and traditional heuristic evaluation, filled DPUQ and SUS questionnaires and all quantitative data tracked during user testing as well as their analysis. In addition, all documentation provided to the vendors during the procurement process regarding the usability procedure was gathered, such as instructions and descriptions on scoring criteria. Moreover, the internal documentation on the product comparisons was consulted.

One way to evaluate the value of the developed usability evaluation methods for IS procurement is by their ability to determine the level of usability of the evaluated system correctly, comprehensively, and efficiently. Reliability and validity are two often used measures for examining an evaluation method (Hartson, Andre and Williges, 2003), and these are included in correctness.
Thoroughness (Bastien and Scapin, 1995) can be likened to comprehensiveness which is related to examining the system as broadly as possible. Efficiency (Hartson, Andre and Williges, 2003) of a method can be evaluated by examining the resources that are needed for getting a feasibly comprehensive evaluation of the system.

These three measures were examined for HED thoroughly in Paper II. Correctness was assessed with statistical analysis of the grade data. For reliability, percent agreement as an index of inter-rater agreement (IRA), “the extent to which different raters assign the same precise value for each item being rated” (Gisev, Bell and Chen, 2013), was used together with inter-rater reliability (IRR). IRR refers to “the extent to which raters can consistently distinguish between different items on a measurement scale” (Gisev, Bell and Chen, 2013), and for it weighted kappa and intra-class correlation coefficient (ICC) were used. For analyzing validity, correlation with DPUQ was calculated. The R statistics software was used for all the calculations. The comprehensiveness of HED was evaluated both against traditional heuristic evaluation and DPUQ. The efficiency of HED and DPUQ were also compared in Paper II. For DPUQ, correctness has not been evaluated with statistics. All these statistical analyses were done by the author of this dissertation.

The usability measures collected with paired-user testing for the healthcare topics were examined and the systems compared in Paper IV. MS Excel was used for calculating task success, errors, and satisfaction during testing. The author of this dissertation carried out these analyses. In Paper V, the correlation of the different measures in revealing the difference between systems is calculated by the main author of that paper. In the procurement project, due to low numbers of users, we run elaborate statistical tests based on those recommended (Sauro and Lewis, 2012) to ensure the differences revealed between vendors in the final scoring of each metric were justified. This statistical analysis is not part of the dissertation.

3.6 Quality Criteria for the Research

Reliability and validity of research are generally used concepts in relation to perform good scientific research. The terminology used for the criteria of good research can, however, differ depending on the research strategy used as do the criteria itself. The research in this dissertation follows the action design research approach, as presented in section 3.1, which draws from both action research and design science. To determine the appropriate quality criteria for this dissertation work I will thus look to the criteria of these two approaches...
first. The quality criteria for action research is different from that of quantitative and qualitative research, because the goal of the research is different (Herr and Anderson, 2005, chap. 4). While generally accepted criteria for its quality do not exist, Herr and Anderson (2005, chap. 4) have proposed their own validity criteria based on the goals of action research: dialogic validity, outcome validity, catalytic validity, democratic validity and process validity.

For design science research, the situation is similar as no commonly agreed upon validity criteria seem to exist beyond evaluating the utility of the produced IS artifact (Larsen et al., 2020). Suitable criteria for evaluating the IS artifacts include “functionality, completeness, consistency, accuracy, performance, reliability, usability, fit with the organization and other relevant quality attributes” (Hevner et al., 2004). A recent literature based framework of examining validity of DSR lists three viewpoints: design antecedent validity (theoretical congruence and requirements), development and use context validity (data input and internal artifact validity), as well as design outcome validity (criterion and relative improvement validity) (Larsen et al., 2020). One argument is that the quality criteria of design science differ based on the type of knowledge production phase of the project: whether general (nomothetic design, nomothetic science), or situational/specific (idiographic design or idiographic science) (Baskerville, Kaul and Storey, 2015). Examples of quality criteria for nomothetic design are “applicability, generalizability, external validity, transferability, consistency, reliability”, “production of acceptable similarity between expected and observed performance, inventiveness, innovativeness, and originality” (Baskerville, Kaul and Storey, 2015).

In action design research, characteristics of DSR and AR are combined and the key characteristics are: the purpose is to generate generalizable knowledge on implementing an artifact that solves a problem in the actual context, there is total collaboration between the researchers and participants, the collaboration occurs from the beginning until evaluation of developed artifact and the development and evaluation is iterative (Collatto et al., 2018). Based on this, it would seem that most above listed validity criteria can be deemed possibly relevant for an action design research project. Some of the criteria evaluate the artifacts and others the overall research. There is also some overlap between different criteria names and their definitions between the two research approaches.

Based on the above, for this dissertation work, I have selected the following quality criteria for evaluating the research done with the ADR approach: process validity, design antecedent validity, development and use context validity, democratic validity, outcome validity and dialogic validity. With process
validity it is evaluated whether the problems have been “framed and solved in a manner that permits ongoing learning” (Herr and Anderson, 2005, chap. 4) and if the process is reflective and iterative. With design antecedent validity it is evaluated whether the artifact development background is valid and how congruent the development done is with previous theory. With development and use context validity the internal validity of the artifacts is evaluated, their construction and structure, how they represent the environment (evaluation of usability in this case) and how valid the data input for the artifacts is. With democratic validity the level of collaboration within the research is evaluated, as well as the relevance and applicability of the research to the context, called local validity. With outcome validity it is evaluated whether the action resulted in solving the problem or redefining the problem, but also the artifacts quality is evaluated. With dialogic validity it is evaluated whether the research has been peer-reviewed and communicated within the scientific community.
4. Results: Usability Evaluation Methods and Procedure for Public Procurement

This chapter presents key results from the research work for this dissertation in order to answer the research questions. The full results are presented in the six research papers that can be found as attachments to this dissertation. Some details have also been added that were not part of the original research papers, and these are called out clearly in the chapter. The results are organised in this chapter following five themes related to the research questions. All relevant results from each research paper are combined under each theme. Table 2 (next page) presents the mapping of the papers to each theme. Section 4.1 presents the methods, metrics, and measures for quantifying usability during procurement. Section 4.2 addresses three usability evaluation methods for procurement and their development in more detail. Section 4.3 deals with the details of quantifying usability with these three methods. In section 4.4, the implementation of these methods in the procurement project is reported, followed by an evaluation of each method’s success. Finally, in section 4.5 considerations of use context and user group when planning the evaluation procedure are reviewed, the complete evaluation framework is presented and the combined scoring of usability for the procurement decision is addressed.
### Table 2. Themes in this chapter and the research papers’ relevance to the themes

<table>
<thead>
<tr>
<th>Section</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>4.1 Methods, Metrics and Measures for Quantifying Usability</td>
<td>x</td>
</tr>
<tr>
<td>4.2 Introducing HED, DPUQ and Paired-User Testing Methods</td>
<td>x</td>
</tr>
<tr>
<td>4.3 Quantifying Usability Attributes with the Proposed Methods</td>
<td>x</td>
</tr>
<tr>
<td>4.4 Implementing and Evaluating the Proposed Methods</td>
<td>x</td>
</tr>
<tr>
<td>4.5 Constructing the Procedure for Combined Scoring of Usability</td>
<td>x</td>
</tr>
</tbody>
</table>

### 4.1 Methods, Metrics and Measures for Quantifying Usability

In Paper I, we presented a framework for usability evaluation in procurement. The framework included six different usability attributes that were evaluated and quantified with several different methods and produced in total 15 distinct measures for usability. The usability attributes included aspects from the ISO (2018) and Nielsen (1993) definitions of usability: efficiency, effectiveness, satisfaction, lack of errors and learnability. Additionally, an attribute we termed as “quality of user interface design” was included. This final attribute viewed the UI design and interaction as a whole, covering efficiency, effectiveness, errors, and learnability through usability heuristics. The attributes represent a wide array of viewpoints on usability.

The evaluation methods include both established methods: usability test (Nielsen, 1993; Dumas and Fox, 2012), paired-user test (Hackman and Biers, 1992), system usability scale (SUS) questionnaire (Brooke, 1996), and preference ranking question (e.g. Agnisarman *et al.*, 2017), as well as those we developed during the action design research project: usability questionnaire for demonstrations in procurement (DPUQ), heuristic evaluation during demonstrations (HED), traditional heuristic evaluation utilizing the HED scoring procedure, expert evaluation of optimal paths and interactive scenario based demonstrations (group inspection sessions).

As mentioned in Papers I-V, the realities, restrictions, schedules, and methodological choices of the whole procurement project affected the selection of and requirements for the usability methods. There was both a need for methods giving an overall picture on system usability and ensuring the minimum level of usability broadly and efficiently, as well as a need for methods that
would give a more detailed view on the usability. The contexts of use, user groups and their tasks also affected the choice of methods. Most of the methods were used for evaluating the usability for different professional user groups, while more traditional heuristic evaluation and usability tests were used for evaluating the usability for citizen users. This was due to a smaller scope for the system to evaluate and simpler tasks when it comes to citizens. An overall need for all methods was to be able to quantify the usability in a way that would enable comparison between systems. Three main methods for procurement, HED, DPUQ and paired-user testing, and their associated measures and quantification are discussed in more detail in Papers II-V and are presented in sections 4.2, 4.3 and 4.4. All the measures relating to these methods can be seen in Table 3. The measures include both those that are more established (such as SUS score and task completion percentage) and ones that were defined during the procurement project (mostly for HED, DPUQ and group inspection methods).

Table 3. Usability evaluation methods, attributes, measures, and targeted user group (revised from Paper I). (Continues on the next page.)

<table>
<thead>
<tr>
<th>Evaluation method</th>
<th>Type/Session</th>
<th>Attribute</th>
<th>Measure</th>
<th>User Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>HED</td>
<td>Expert evaluation / Scenario-based demonstration</td>
<td>Quality of UI design</td>
<td>Documented usability issues: heuristic violations, missing functionalities, omitted parts of the user scenario and positive findings</td>
<td>Professionals</td>
</tr>
<tr>
<td>Task-based heuristic evaluation</td>
<td>Expert evaluation</td>
<td>Quality of UI design</td>
<td>Documented usability issues: heuristic violations, missing functionalities, and positive findings</td>
<td>Citizens</td>
</tr>
<tr>
<td>DPUQ</td>
<td>Questionnaire / Scenario-based demonstration</td>
<td>Satisfaction</td>
<td>Perceived usability: professionals’ responses during and after session</td>
<td>Professionals</td>
</tr>
<tr>
<td></td>
<td>Questionnaire / Group inspection session</td>
<td>Satisfaction</td>
<td>Perceived usability: professionals’ responses after session (DPUQ summative part)</td>
<td>Professionals</td>
</tr>
<tr>
<td></td>
<td>User testing</td>
<td>Effectiveness</td>
<td>Percentage of successfully completed tasks</td>
<td>Professionals, Citizens</td>
</tr>
</tbody>
</table>
## Results: Usability Evaluation Methods and Procedure for Public Procurement

<table>
<thead>
<tr>
<th>Evaluation method</th>
<th>Type/Session</th>
<th>Attribute</th>
<th>Measure</th>
<th>User Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired-user test</td>
<td>Errors</td>
<td>Errors</td>
<td>Errors made by the user during task completion</td>
<td>Professionals, Citizens</td>
</tr>
<tr>
<td></td>
<td>Satisfaction</td>
<td>Satisfaction</td>
<td>Positive and negative markers given by users during task completion</td>
<td>Professionals, Citizens</td>
</tr>
<tr>
<td>SUS</td>
<td>Questionnaire / User testing</td>
<td>Learnability</td>
<td>Perceived learnability: learnability factor of SUS* from user’s responses after task completion</td>
<td>Professionals, Citizens</td>
</tr>
<tr>
<td></td>
<td>Questionnaire / User testing</td>
<td>Satisfaction</td>
<td>User’s responses after task completion (SUS score)</td>
<td>Professionals, Citizens</td>
</tr>
<tr>
<td>Preference ranking</td>
<td>Questionnaire / User testing</td>
<td>Satisfaction</td>
<td>User’s rank of systems based on preference after task completion</td>
<td>Citizens</td>
</tr>
<tr>
<td>Expert evaluation</td>
<td>Expert evaluation</td>
<td>Efficiency</td>
<td>Number of steps in the optimal solution to tasks</td>
<td>Professionals, Citizens</td>
</tr>
<tr>
<td>Expert evaluation</td>
<td>Expert evaluation / Interactive scenario-based demonstration (group inspection session)</td>
<td>Efficiency</td>
<td>Usability specialist’s assessment of efficiency</td>
<td>Professionals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Efficiency</td>
<td>Usability specialist’s assessment of efficiency of configuring the system</td>
<td>Professionals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Learnability</td>
<td>Usability specialist’s assessment of learnability based on professional’s verbal answers</td>
<td>Professionals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Learnability</td>
<td>Usability specialist’s assessment of learnability of configuring the system</td>
<td>Professionals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quality of UI design</td>
<td>Usability specialist’s assessment based on usability heuristics and professionals’ discussions during demonstration</td>
<td>Professionals</td>
</tr>
</tbody>
</table>

* SUS learnability factor (Lewis and Sauro, 2009)

### 4.2 Introducing HED, DPUQ and Paired-User Testing Methods

In Papers II-V, we presented three usability evaluation methods, their development and implementation during the procurement project. The methods were **heuristic evaluation during demonstrations** (HED, Paper II),
usability questionnaire for demonstrations in procurement (DPUQ, Paper III), and paired-user testing (Papers IV-V). HED and DPUQ were developed during the procurement, while for paired-user testing the focus was on how to plan usability testing in a procurement setting as well as measure and quantify the usability attributes.

4.2.1 Developing the Methods

In Papers II and III we presented how the development of HED and DPUQ methods was done in four phases: literature review, development of first version, piloting, and development of final version. At the same time as the two other methods, also paired-user testing methodology (Papers IV and V) was adapted and piloted, although this is not described in the papers.

The HED and DPUQ methods were developed for the first phase of the procurement, where there were several vendors’ systems to be evaluated. At this phase the goal was to assure that a minimum level of usability was reached and to narrow down the number of candidates to continue in the procurement process. Other evaluations were planned to be based on comprehensive user scenarios, which the competing IS vendors would follow strictly during their demonstrations of their systems. End-user representatives, i.e., physicians, nurses, and social welfare professionals, from the procuring organizations were to be following these demonstrations to evaluate the functionalities. For efficiency, these scenario-based demonstrations were thus the basis for also evaluating the usability at this stage, as they would include most identified key contexts and critical and frequent tasks. The user scenarios were based on user stories written with the user representatives in workshops and were similar to clinical information processing scenarios (CLIPS) (Lincoln and Essin, 1993; Einbinder, Remz and Cochran, 1996; Belden, Grayson and Barnes, 2009). This approach originally developed for healthcare was also adapted for social welfare scenarios.

In the second phase of the procurement, the goal was to further evaluate the remaining vendors’ systems as part of the final decision. This required measuring and quantifying their usability in a way that could be presented as one score. In this phase, with fewer vendors remaining, it was possible to gather stronger evidence on usability from a narrower scope of use contexts and paired-user testing was selected as a main method. This method is presented in Papers IV and V.
Literature Review

The existing literature on usability evaluation methods in general and more specifically in the context of HIS procurement and healthcare context has been described in Chapter 2. The key take aways for the development of the methods are presented here.

A well-established expert evaluation method for usability, is heuristic evaluation (HE) (Nielsen and Molich, 1990) that was developed for UCD (Gulliksen and Göransson, 2001). It has also been presented as a method to use in healthcare IT procurement (Kushniruk et al., 2010). The known limitations of HE include lack of use context awareness, a narrow perspective on the whole system with only a limited number of UIs and not revealing missing functionalities (Chin, Diehl and Norman, 1988; Nielsen, 1993). Extending HE to be used with the user scenario-based demonstrations added the missing work context and process view and also extended the scope of evaluation. Additionally, as HE is mainly a qualitative method, there was a need to enhance it to enable score-based comparisons between competing systems.

To collect the views and perspectives of users on perceived usability and satisfaction, multiple standardized usability questionnaires have been developed, e.g. QUIS (Questionnaire for Interaction Satisfaction) (Chin, Diehl and Norman, 1988), SUMI (Software Usability Measurement Inventory) (Kirakowski, 1996), CSUQ (Computer System Usability Questionnaire) (Lewis, 1995) and SUS (System Usability Scale) (Brooke, 1996). These questionnaires are mainly aimed for use after usability testing. The critique towards these questionnaires includes the lack of being able to consider work contexts in complex settings, such as healthcare. For this purpose, a healthcare specific questionnaire has been developed for physicians (Viitanen et al., 2011), and it has later been validated (Hyppönen et al., 2019). After the procurement, the questionnaire has been modified also for nurses (Hyppönen et al., 2018) and social welfare professionals (Ylönen et al., 2020). The existing questionnaires presented a good starting point for gathering the users’ assessment of usability during the demonstration sessions.

Also user testing is used during procurement according to literature (Beuscart-Zéphir et al., 2002, 2005; Liljegren and Osvalder, 2004; Jorritsma, Cnossen and van Ooijen, 2014). It is especially suggested as a method giving good evidence on usability (Kushniruk et al., 2010). Usability testing is an established method (Nielsen, 1993), that is also used for summative purposes to measure a variety of usability attributes (Hornbæk, 2006; Sauro and Lewis, 2012). Based on this, it was included in the evaluation procedure planning from the beginning. However, literature emphasizes that usability testing is
resource intensive in procurement even for a narrow scope (Riihiaho et al., 2015) and controlled usability testing might not represent actual use accurately (Siegel, 2012). This influenced positioning the method in the second phase of evaluation along with other methods, and after evaluations that more broadly covered the systems. In this phase also fewer systems remained.

First Versions of Methods
The first version of HED was not described in the attached research papers. It followed traditional heuristic evaluation procedure, but during a demonstration. There was a paper form designed for documenting the heuristic violations during the demonstration, categorized by heuristic, and to collect the evaluation for their severity. The scoring based on severity ratings was not yet designed for the piloting.

The perceived usability questionnaires’ (DPUQ) development was presented mainly in Paper III, but also discussed in Paper II. The established usability questionnaires, CSUQ (Lewis, 1995), SUMI (Kirakowski, 1996), SUS (Brooke, 1996), and QUIS (Chin, Diehl and Norman, 1988), were used as a basis for the questionnaire together with the tailored usability questionnaire for EHR systems (Viitanen et al., 2011), which has later been validated (Hyppönen et al., 2019). The procurement setting and the user scenario-based demonstration influenced the questionnaire design from the beginning. First, the end users would evaluate also functional scope and quality at the same time. Second, the questions needed to be suitable for a range of different clinical scenarios. Last, some usability aspects were identified already in the beginning to be impossible to evaluate due to the demonstration setting, such as error recovery and terminology (as the demonstrated systems might not be in the native language of users). The first version of DPUQ had two parts: eight questions to be answered during a demonstration break and a standard SUS questionnaire to be answered in the end.

Use of paired-user testing was presented in Papers IV and V, but there is not an elaborate discussion on its development for the procurement. Usability testing with healthcare and social welfare professionals was planned for the second phase of procurement. Because the users were not trained to use the systems being evaluated and only familiar with them through seeing demonstrations, a paired-user usability testing method (Hackman and Biers, 1992) was selected for use so they could function as a team. This was seen to reduce the effects of possible individual users with poorer performance due to the aforementioned reasons. Beyond that, following traditional procedures of conducting the actual testing with users and a moderator was considered
applicable. The focus was placed on selecting and writing suitable tasks without knowing the evaluated system as well as on selecting and collecting usability measures and planning the scoring. The topics and tasks in the paired-user testing were selected to represent typical and frequent tasks with the IS for a few main user groups of professionals. Domain experts working in the procurement office together with the first usability researcher prepared the preliminary testing scenarios.

**Piloting**

The first round of piloting for the evaluation methods in the procurement, including usability methods, took place during summer 2013. For HED and DPUQ this is presented in the papers. For paired-user testing, the papers do not describe the piloting stage. In the first piloting round, the first versions of HED and DPUQ as well as paired-user testing with one scenario were piloted with one of the EHR systems in use at the time at one of the procuring organizations.

Usability researchers applied the first version of the **HED** method, and the author of this thesis was one of them. The documentation was done either on the paper form or on the computer, depending on preference. They were not familiar with the user scenario nor with the evaluated system beforehand. The purpose of the piloting was to gather actual experiences on documenting findings without a possibility to pause the demonstration, and to analyze how the findings would aggregate.

The piloting was a significant step in the development process of **DPUQ**. There were 28 user representatives participating in the piloting, who filled in the questionnaire after watching a demonstration. They provided feedback on the questionnaire.

During the **paired-user testing** pilot in round one, the usability researcher working in the procurement project acted as the moderator and documented the completion of tasks on paper. The users gave satisfaction feedback with a dedicated hardware solution during the completion of test tasks (described in Paper V, but not in the scope of this dissertation), and filled in a paper SUS questionnaire after the testing. The author of this thesis, along with a group of domain experts and procurement office staff observed and took notes on the testing procedure. Further piloting for paired-user testing with other scenarios, including social welfare, was done during fall 2013. The focus of the second round was on the scenarios and improved hardware solution.
Development of Final Versions of Methods

There were several findings made during the piloting that influenced the method development. The main findings for the piloted HED method were: 1) It was too slow to document on the paper form. 2) The fact that the evaluators did not know the user scenario nor the application domain beforehand and not all were as familiar with the heuristics influenced their results. Thus, training and practicing were deemed as important as well as familiarizing themselves with the domain: clinical work and EHR systems as well as social welfare. 3) It was noticed that the findings started to repeat during the demonstration, which should be considered as part of determining the scoring and grade boundaries for different length demonstrations. But also, witnessing the same issues emerging repeatedly helped in specifying the findings over time during the fast-paced demonstration.

For DPUQ, the following observations were made during piloting: There were wording changes needed for the SUS to reflect the demonstration situation where the system was not used by the users themselves, also some statements were entirely inappropriate: this included learnability, confidence, and efficiency related statements. Statements that evaluated the need for functionalities were not suitable because the scenario was not the object of evaluation. Also, the questions needed to reflect evaluating a system for complex work. Alternating positive and negative statements as in the original SUS seemed prone to error and confusing for participants. It was also discovered that a third set of questions was needed for longer demonstrations. Based on the observations, the first eight statements of the original DPUQ were reduced to six and it became part 2 of DPUQ. Questions were rearranged, and wordings were modified. Part 1 of DPUQ was added to include six new questions. The final part of DPUQ was modified so that from SUS three statements were replaced with new ones, only positive statements were used, and the statements were rearranged.

With paired-user testing, the piloting influenced mainly the writing of the test tasks and how the measures would be documented during testing. During piloting, it was observed that the moderator documenting the successful completion of tasks and errors made on paper was confusing, difficult to follow up on afterwards and prone to mistakes. The collected measures were also modified. Based on these findings, the technical solution for moderator documentation described in Paper V was developed. This solution, however, is not in the scope of this dissertation. The quantification of user satisfaction during and after test tasks was also modified based on the piloting. It was observed that users had uneven habits of giving their feedback with the devices, so we
focused on balancing this. An additional question by the moderator, for both users to give one final positive or negative feedback on their experience after completing each task was added to ensure that all tasks got at least one feedback from each user. Also, because of this, it was seen that simply calculating the sum of positive and negative feedback presses would not be equal between vendors, so the measure was modified.

The test tasks and scenarios were another main focus for development. First, based on piloting it was seen, that there were too few test tasks. Additionally, guiding the domain experts further in envisioning and writing of scenarios and test tasks from a goal and future use point of view was seen necessary. The test tasks also needed to be system independent, logical from the user goal point of view and suitable for users with limited familiarity with the systems beforehand. To achieve this goal, iterative writing of tasks with both usability researchers and domain experts was needed. As a distinctive difference to traditional usability test script writing, the role of the domain experts was integral in the writing because of the complex healthcare and social welfare context. Usability researchers were responsible for final phrasing of tasks and the flow of the test. A need for test moderators to practice and familiarize themselves with the system with the help of domain experts pre-testing was also observed during piloting.

Finally, considering the communication and information exchange with the vendors regarding the testing procedures was deemed necessary to plan in detail. During the development emphasis was put on identifying the needed information for successfully preparing the system and moderating the tests, including determining the successful completion of tasks. Providing the needed background information for tasks in enough detail and carefully planning the phrasing of tasks to be unambiguous required extra effort compared to normal test preparation. The principal of equal treatment of vendors also influenced the planning of these practices so that it could be ensured the same information was given to and received from all vendors.

4.2.2 Heuristic Evaluation during Demonstrations - HED

This method was presented in detail in Paper II. The usability attribute evaluated with this method is quality of UI design. The measure used is the documented usability issues: heuristic violations, missing functionalities, omitted parts of the user scenario and positive findings. The method produces one grade for the system seen during the demonstration. To utilize this method, at least two usability experts are needed, domain experts are needed, and users
are not involved in using this method. An overview of the method is given in this subsection and the measures are described in section 4.3.

Both HED as well as DPUQ, which is presented in the next subsection, are to be applied during a demonstration based on a user scenario that describes typical work processes (Einbinder, Remz and Cochran, 1996; Carroll, 1997) and should be written by domain experts. The demonstrations are required to strictly follow the scenarios without presenting anything outside it. The competing vendors are given the scenarios beforehand. The scenarios include predefined demonstration breaks and can last many hours in length when demonstrated.

There are three phases in applying HED: 1) Alignment on heuristics and scoring before the evaluation. 2) Actual evaluation that is done individually by usability experts during the demonstration. 3) Aggregation of results and determining the grade after the demonstration. The details of the phases are illustrated in Figure 7.

Any set of heuristics can be used with HED. It is however important, that the evaluators are familiar with the heuristics and have a similar understanding of them in the context of the system evaluated. This consensus can be achieved by defining examples of typical heuristic violations for the type of system evaluated. Before the evaluation, also a common understanding of rating the issues should be reached between evaluators. Practicing applying the method together should also be done. It is also important that the evaluators are familiar with the user scenarios before evaluation.

During the evaluation phase three other types of usability issues are documented in addition to heuristic violations: missing functionalities, omitted

---

**Heuristic Evaluation during Demonstrations (HED)**

**Before**
- Alignment between evaluators on heuristics and rating issues
- Familiarizing with user scenario
- Practicing
- Determining essential functionalities

**During**
- Individual evaluation
- Documenting viewed issues

**After**
- Consulting domain experts on missing functionalities and omitted parts of scenario
- Aggregation of results between evaluators
- Determining the final grade

*Figure 7. Three phases of the HED method.*
parts of the user scenario and positive findings. The most essential functionalities (e.g., summary views of data) are determined beforehand with domain experts, who can explain what is needed to complete the users’ goals in the system. To evaluate usability, the functionality must first exist. During the demonstration all four types of usability issues are given numeric scores, influence ratings. The scoring principles used in HED are described in section 4.3. A dedicated electronic spreadsheet is used for the documentation, to do it efficiently as the demonstration progresses. The difference to traditional heuristic evaluation (HE) (Nielsen and Molich, 1990) is revealing missing functionalities and uncompleted user goals. This gives a more truthful view of what the overall usability of the system would be when using it. Another difference to traditional HE is not documenting which heuristics are violated, as it is not relevant to the goal of the process.

After the demonstration, domain experts who also followed the demonstration should be consulted to add the same missing functionalities and omitted parts of the user scenario to the individual evaluations. Finally, the evaluators share their individual results with each other and negotiate the final usability grade for the scenario. This is also described in more detail in section 4.3.

### 4.2.3 Perceived Usability Questionnaire for Demonstrations - DPUQ

This method was presented in detail in Paper III. The usability attribute evaluated with this method is satisfaction and the measure used is perceived usability (professionals’ responses during and after a demonstration session). To use this method, usability expertise is not needed, and users do the evaluation. An overview of the method is given in this subsection and the measures are described in section 4.3.
The construction and application of DPUQ is illustrated in Figure 8. DPUQ is meant for evaluating usability by users when not using the system themselves but rather watching it being used. It is intended to be used in user-scenario based demonstrations as described in the previous subsection. The questionnaire covers themes that are universal for the usability of all systems (consistency, logic, status, complexity, and visual appearance of the system) as well as those more specific to healthcare and social welfare (compatibility of system and subject domain work tasks, support for collaboration in subject domain work). The former questions’ design utilized several existing questionnaires (Chin, Diehl and Norman, 1988; Kirakowski, 1995; Lewis, 1995; Brooke, 1996; Sauro and Lewis, 2011; Viitanen et al., 2011). The latter questions were designed based solely on an EHR system usability survey (Viitanen et al., 2011), which has later been validated (Hyppönen et al., 2019). The questions were considered appropriate also for social welfare. The questionnaire has three sub questionnaires; two separate question sets (six statements each) for demonstration breaks and a third, summative set (10 statements), for the end of the demonstration. The short questionnaires include same number of both general and EHR-specific usability statements, while the summative questionnaire is more similar to SUS (Brooke, 1996) but includes also three EHR-related questions. Examples of the statements in DPUQ are: “I perceive the arrangement of the fields and functionalities on-screen logical”, “The system supports collaboration and information exchange between involved parties” and “This system is highly suitable for my daily work tasks”. The statements are rated using a four-point Likert scale: Fully agree (3) – Fully disagree (0). The questionnaire can be found in Paper III with references to existing questionnaires.
4.2.4 Paired-User Testing

This method was presented in Papers IV and V. The usability attributes used this method to evaluate were effectiveness, (lack of) errors and satisfaction. The measures used in relation to these attributes were: successfully completed tasks, errors made by the user during task completion, and positive and negative markers given by users during and after each task completion. Additionally, the SUS questionnaire (Brooke, 1996) was used after testing to evaluate satisfaction and learnability, and the measures were regular SUS score and learnability factor calculated from SUS responses (Lewis and Sauro, 2009). To use paired-user testing, in addition to needing users, usability experts and domain experts are also needed. Their expertise is used for preparing the scenarios. The moderator could also be someone with less usability expertise as user guidance is minimal and qualitative data is not collected. An overview of the method is given in this subsection and the measures are described in section 4.3.

The paired-user testing follows traditional moderated usability testing practices, except instead of one there are two users who solve the tasks as a team. For use in procurement the following considerations are important: The vendor is given the test tasks and required patient/client history data in advance and configures the system to best meet the user goals described in the test tasks. To help in running the tests, the vendor is asked to provide a manual of the correct paths to complete the tasks. This is used to prepare for the moderating by learning the correct use and to determine the criteria for successful completion of tasks together with domain experts. The vendor also prepares a short training video on the topic of the testing for the users to watch in the beginning of the testing session. The test tasks are read to the users by the moderator and given also in writing. The moderator does not assist the users with task performance. The task of the moderator is to document the data for the usability measures and ensure the time limits for each task and overall testing are kept. Details of the method use are presented in Papers IV and V.

To ensure equal conditions for vendors during procurement, the testing needs to be planned in a balanced way. A within-participants study design is suitable for a situation with two vendors with different systems, so same pairs use both systems. In this design, the order of the systems the pairs use should also be alternated. Users should be equally familiar with all systems they evaluate. A between-participants study design is suitable for situations where both vendors are offering the same system with different configuration, so there are different users for each system. To ensure the task completion is interpreted
in a similar way for both systems, the moderator should stay the same in one test topic.

4.3 Quantifying Usability Attributes with the Proposed Methods

In the following, I give an overview on how the selected usability attributes are scored and quantified with each of the three methods: HED, DPUQ and paired-user testing.

4.3.1 Scoring of HED

The quantification and scoring for HED was presented in Paper II with more elaborate reasoning. The scoring was based on influence ratings that were set for the documented usability issues and resulted in one grade. The influence rating was conceptually similar to severity rating in traditional heuristic evaluation, but also covered supportivity rating for positive findings. In HED a two dimensional approach was used for the influence rating, with non-correlated impact and frequency as the two dimensions (Nielsen, 1993, p. 104; Sauro, 2014) that were rated separately. The influence rating for each usability issue was calculated as the product of the impact and frequency ratings. The rating of these two dimensions differed based on the type of usability issue.

The demonstrated system was assumed to have the highest grade as a starting point, which equaled to 0 usability points. The usability issues were given points that were then added to or subtracted from the 0 starting points. The final points after the evaluation determined the usability grade, according to a predetermined grading scale. The deficiencies identified during the evaluation were given minus points, and as this was the main focus of HED, negative total points were assumed. Details for scoring and grading are presented below.

For heuristic violations, the impact of the violation on users was rated on a scale from -1 (minor usability violation) to -3 (major usability violation) and the frequency with which the violation occurred in the demonstration on a scale from 1 (single) to 3 (prevailing/frequent). It was also possible to give an impact score of -10 to a single usability catastrophe, in which case the frequency was 1. For positive findings, the impact on users was always evaluated as +1 and the frequency was evaluated on the same scale as for heuristic violations.

The last two usability issue types, missing functionalities and omitted parts of the user scenario, were scored similarly. When a predetermined main
functionality was missing from the demonstration, it was given a rating of either \(-5\) or \(-10\) (impact of \(-5/-10\) and frequency 1) depending on the functionality or level of deficiency. Omitted parts of the user scenario during demonstration were presumed to include the same average amount of heuristic violations as demonstrated parts and the points were added accordingly based on the percentage of user scenario missing from the demonstration. A resulting score \(US\) was calculated for the analysis of each evaluator.

\[
US(v_s) = \sum (I_n \cdot F_n)
\]

where \(I_n\) is the impact and \(F_n\) the frequency of each usability issue \(n\) for vendor \(v\) for user scenario \(s\).

It should be noted that missing functionality and omitted parts of the user scenario were not used to disqualify a vendor. Rather, the reasoning was that all vendors should be treated equally, and these issues should affect the usability score because all the functionalities in the user scenarios were not necessarily requirements for the final system.

Another important part of the scoring process for HED was determining grade boundaries because grades depicted the level of usability in a more easily understood format. It also simplified communicating the accepted minimum level for usability. The grade boundaries we used are not universal. Instead, they were determined by piloting the method and considering the length of the demonstration and complexity of the system evaluated. As an example, the grade Fail could be determined to be minus points of at least 90, equaling to ten major (impact \(-3\)) and prevailing (frequency 3) usability violations. The grade boundaries should include a point range around the boundary (a grey area) to determine those points that result in an unambiguous grade and need to be negotiated. Paper II included examples of grade boundaries.

After the evaluation, the evaluators compared their individual \(US(v_s)\) scores. They listed and compared the number of heuristic violations overall and violations for each severity rating for all evaluators, the number of positive findings and their ratings as well as notes on the other two subtraction types. The resulting individual grades were also compared. Evaluators should reach a joint conclusion on the final grade. Closer comparison of individual analysis and negotiations were needed if the individual grades differed from each other significantly and points were not within the grey area.
4.3.2 Scoring of DPUQ

The quantification of DPUQ was presented in Paper III. DPUQ was designed to produce two scores of perceived usability to represent the usability attribute of user satisfaction for an evaluated demonstration. The two six-question sub questionnaires filled during demonstration breaks formed the first score and the third summative sub questionnaire filled at the end formed the second score. The first score $DPUQ_1$ was an arithmetic average of all answers to the sub questionnaires from all participants. The second score $DPUQ_2$ was an arithmetic average of the sums of the answers to the ten questions by each user. Each question was rated using a four-point Likert scale: Fully agree (3) – Fully disagree (0). Thus, $DPUQ_1$ was between 0-3, and $DPUQ_2$ was between 0-30. The two scores were calculated as follows.

\[
DPUQ_1 \left( v_s \right) = \frac{1}{Q} \cdot \sum_{q=1}^{Q} R^q
\]

where $R^q$ is rating of a statement $q$ from the first two DPUQ sub questionnaires for vendor $v$ for user scenario $s$ and $Q$ is the total number of ratings for statements.

\[
DPUQ_2 \left( v_s \right) = \frac{1}{N} \cdot \sum_{n=1}^{N} \left( \sum_{q=1}^{10} R^q_n \right)
\]

where $R^q_n$ is the rating from participant $n$ to statement $q$ in the third DPUQ sub questionnaire for vendor $v$ for user scenario $s$ and $N$ is the total number of participants.

The calculation of the first score made it possible to use the first two sub-questionnaires multiple times if the demonstration length required it. The summative questionnaire design was based heavily on SUS questionnaire, and the calculation of the second score as a sum was inspired by SUS.

4.3.3 Scoring of Paired-User Testing

The quantification of the usability attributes from paired-user testing used in procurement was presented in Paper V, while Paper IV presented an
alternative way of considering the details of the measures outside of procurement context. In procurement, paired-user testing was used to measure effectiveness, (lack-of) error and satisfaction with a total of three measures: percentage of successfully completed tasks, points for deviations from optimal path and satisfaction feedback rate during task completion. The SUS questionnaire was used for measuring satisfaction and learnability with two additional measures: SUS score and learnability component of SUS.

For effectiveness, the goal of each task and the optimal path documentation from the vendor was used to determine successful completion of a task. The criteria for successful completion of each task were set beforehand based on the outcome in the system. Errors were allowed during task completion if they did not affect reaching the goal. For each testing topic and task there was a maximum time for trying to complete the tasks. If the maximum time was exceeded, the task was failed. For procurement, the calculation for the percentage of successfully completed tasks PSCT (as presented in Paper V) differs from the traditional way of calculating it based only on tasks attempted during the testing time (as presented in Paper IV). PSCT was calculated as follows over all testing topics.

$$PSCT (v) = \frac{\sum S_t}{\sum N_t} \cdot 100$$

where $S$ is the number of successfully completed tasks and $N$ is the total number of tasks in testing topic $t$ for vendor $v$. The sums are calculated over all testing topics.

To calculate the score for (lack-of) errors, the data for deviations from optimal path was needed. For procurement, errors were defined as deviations the users made from the optimal path (provided by the vendor) during task completion. Examples of deviations included e.g., navigating to wrong view, unintentional activity relative to the goal, mistake in interaction or ignorance of substantial information on screen. These deviations (errors) were divided into small (0.5 points) and large (1 point). Only successfully completed tasks were considered in the calculation. The total error points $EP$ were calculated as follows.
\[
EP(\nu_t) = \frac{1}{S} \cdot \sum_{S=1}^{S} \left( \frac{1}{N_s} \cdot \sum_{n=1}^{N_s} \sum_{s} e_n^s \right)
\]

where \( \sum e_n^s \) is the sum of errors for successfully completed task \( n \) in testing session \( s \) for vendor \( v \) in testing topic \( t \). \( N_s \) is the total number of successfully completed tasks in a session and \( S \) is the total number of sessions. If \( N_s = 0 \) for a session, then the average sum of errors for that session is 12 as default.

Satisfaction was measured with two measures in procurement. First, user satisfaction feedback data from task completion was needed. The satisfaction feedback rate was defined as the percentage of tasks with more positive than negative feedback markers from both users combined. During task completion, both users could spontaneously give either positive or negative markers to the use of the system as many times as they desired (e.g., subjectively positive event or intuitive functioning during use, a negative struggle or dissatisfaction towards the behavior of the system). Additionally, after each task the moderator requested both users to give one final marker each based on their feeling overall on task completion. Only successfully completed tasks counted for this measure. The satisfaction rate \( SR \) was calculated as follows.

\[
SR(\nu_t) = \frac{P}{N} \cdot 100
\]

where \( P \) is the number of successfully completed tasks with more positive than negative markers and \( N \) is the number of all successfully completed tasks for vendor \( v \) in testing topic \( t \).

The second satisfaction measure was the SUS score. The SUS score for each user was calculated as a standard measure from the SUS questionnaire (Brooke, 1996). The resulting score was an average of the individual SUS scores and was on a scale of 0-100. Perceived learnability was extracted as two questions from the SUS questionnaire (Brooke, 1996) based on the factor analysis presented by Lewis and Sauro (2009) and calculated as an average of the responses for each user and then averaged.
4.4 Implementing and Evaluating the Proposed Methods

The three developed methods, HED, DPUQ, and paired-user testing were used in the procurement project during year 2014. HED and DPUQ were part of the preliminary assessment phase where the primary goal was to ensure the IS candidates would reach a minimum level of usability. Also, there was a need placed by the project to put the vendors in order according to their scores. In this phase, there were four vendors, although the procedure was planned for six vendors. Paired-user testing was used during the second short-list evaluation phase of procurement where the goal was to inform the final decision and give the vendors points to be used in the final scoring that included also other aspects than usability. In this phase, there were two vendors remaining. The implementation of these methods and their evaluation was presented in Papers II-V. In this chapter, I present an overview of these results.

4.4.1 Preliminary Assessment Phase: HED and DPUQ

There were altogether nine user scenario-based demonstrations held per vendor, six for healthcare and three for social welfare. The scenarios ranged in length from 2.5 hours to 6 hours. HED was used in six of these demonstrations and DPUQ was used in all nine of them, which makes 18 hours of demonstrations per vendor for HED and 25.5 hours for DPUQ. I, together with the other usability researcher, participated in 24 demonstration sessions (six per vendor, four vendors) applying the HED method. The number of respondents for DPUQ varied between user scenarios. The total number of DPUQ questionnaires filled during demonstrations was 771. More details are presented in Table 4.

We as usability researchers had both experience with heuristic evaluation in general and in the use context. The choice of heuristics for HED was Nielsen’s 10 heuristics (Nielsen, 1993) which we were very familiar with and perceived aligned with those found in literature for the health informatics field (e.g. (Belden, Grayson and Barnes, 2009; Zhang and Walji, 2011)). As an additional resource we compiled a list of examples on usability issues specifically found with heuristic evaluation of health information systems. Some examples were presented in Paper II.
Table 4. The topics of the user scenarios, their lengths, and the used usability evaluation methods with number of respondents for DPUQ. The scenarios’ use contexts are presented in more detail in Paper III. Table combined from papers II and III.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Length (hours)</th>
<th>Methods</th>
<th>Respondents *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare (HC) scenario 1 **</td>
<td>6</td>
<td>HED, DPUQ</td>
<td>22-23 (P, N)</td>
</tr>
<tr>
<td>HC scenario 2</td>
<td>3</td>
<td>HED, DPUQ</td>
<td>9-10 (P, N)</td>
</tr>
<tr>
<td>HC scenario 3</td>
<td>2.5</td>
<td>HED, DPUQ</td>
<td>17 (P, N)</td>
</tr>
<tr>
<td>HC scenario 4</td>
<td>3</td>
<td>DPUQ</td>
<td>24-26 (P, N)</td>
</tr>
<tr>
<td>HC scenario 5</td>
<td>2.25</td>
<td>DPUQ</td>
<td>11-12 (P, N, SWP)</td>
</tr>
<tr>
<td>HC scenario 6</td>
<td>2.25</td>
<td>DPUQ</td>
<td>16-17 (P, N, SWP)</td>
</tr>
<tr>
<td>Social Welfare (SW) scenario 1</td>
<td>2.5</td>
<td>HED, DPUQ</td>
<td>29-31 (SWP)</td>
</tr>
<tr>
<td>SW scenario 2</td>
<td>2</td>
<td>HED, DPUQ</td>
<td>29-31 (SWP)</td>
</tr>
<tr>
<td>SW scenario 3</td>
<td>2</td>
<td>HED, DPUQ</td>
<td>29-31 (SWP)</td>
</tr>
</tbody>
</table>

* P=physician, N=nurse, SWP=Social welfare professional.
** Parts 1&2 of DPUQ were used twice in this scenario because of the length.

Before the evaluations started with HED, we determined grade boundaries for Us(v₃) in each user scenario on a scale of 3 (good) – 2 (fairly good) – 1 (acceptable) – 0 (fail) based on the length of the user scenarios and piloting the devised method. The grade boundary calculations were based on reasoning the absolute minimum points for a two-hour length scenario (the worst possible usability) and then setting the boundaries at 90% (fail), 60% (acceptable) and 30% (fairly good) of the points, while the grey areas were + -5% around the boundary. For longer scenarios standard number of minus points were added per each hour. These point and grade boundary details were presented in Paper II.

To ensure equal scoring between evaluators, a shared understanding was formed before the evaluation for using the impact and frequency ratings as well as what a usability catastrophe was. The scenario was also reviewed prior to the demonstration with domain experts from an IS point of view and the parts resulting in subtractions for missing functionality were determined. A final rehearsal of using HED was done with a currently used system a week prior.

During the demonstrations, we did the evaluations and ratings separately on our own laptops according to the process described in subsection 4.2.2. After each demonstration the subtractions for missing functionality and omitted
Results: Usability Evaluation Methods and Procedure for Public Procurement

parts of the scenario were decided with subject domain experts. A review of one’s own documentation was useful after the demonstration to combine multiple documentations of the same issue with an appropriate frequency rating. During the aggregation step, first the Us(v) scores were calculated as presented in subsection 4.3.1. The most severe documented violations were reviewed together in more detail, and it was noted whether both researchers had made the same observation. The actual grade negotiations were straightforward following the process laid out previously. In case of a difference in grading, the main reason was identified. For example, one researcher documenting more unique positive issues could be used as basis for the higher grade. The negotiations took on average one hour per demonstration.

For each of the nine user-scenario-based demonstrations per vendor there were two DPUQ scores calculated following the formulas described in subsection 4.3.2.

4.4.2 Short-List Evaluation Phase: Paired-User Testing

As presented in Papers IV and V, the paired-user testing comprised of three test topics in healthcare and two in social welfare. The scenarios and tasks represented typical tasks of the user groups that were highly important both when using the IS and for patient and client safety. The scenarios were prepared together with domain experts working at the procurement office. The number of tasks in the test topics ranged from 10 to 19. The vendors configured their systems to best meet the user goals in the scenarios and used identical fictitious patient and client data that was provided to them.

In healthcare, three pairs of users tested each test topic with both candidate system, totaling 18 participants in healthcare. In social welfare, separate pairs tested the systems, totaling the number of participants to be 24. This was due to anticipating the brand of the offered systems being identical. All users had been working in their respective fields and using IS for patient and client care for several years. The information on the testing sessions is presented in Table 5 (next page).

I and the other usability researcher in the project moderated the tests following the procedure outlined in subsection 4.2.4. The tests were run at the procurement office. User data (other than SUS questionnaire) was captured with a hardware solution presented in Paper V, that is not in the scope of this dissertation. Details of the testing setup are provided in Papers IV and V.
Table 5. The paired-user testing session information. All sessions were 90 minutes, and maximum length for each task was 12 minutes. The test topics’ use contexts for healthcare are presented in more detail in Paper IV. The social welfare topics were not presented in the papers.

<table>
<thead>
<tr>
<th>Test topic</th>
<th>User group</th>
<th>Study type</th>
<th>Pairs</th>
<th>Test tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare (HC) 1</td>
<td>Nurses</td>
<td>Within-participants</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>HC2</td>
<td>Nurses</td>
<td>Within-participants</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>HC3</td>
<td>Physicians</td>
<td>Within-participants</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Social Welfare (SW) 1</td>
<td>Social workers</td>
<td>Between-participants</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>SW2</td>
<td>Social workers</td>
<td>Between-participants</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

The data gathered from the tests was analyzed for each task, test session and test topics. The gathered data included information on whether the task was completed successfully, the small and large deviations from the optimal path made during task completion, satisfaction feedback from both users during task completion and right after, and a filled SUS questionnaire at the end of the session. The resulting scores for effectiveness, (lack-of) errors, satisfaction and learnability were calculated according to the formulas presented in sub-section 4.3.3. These scores were then transformed into grades.

For (lack-of) errors, the error points for each testing topic were transformed into grades on a scale of 0-5, where the highest grade was received with the least error points. The grade 0 was an exception if no tasks in any session were completed successfully, and the average would have been exactly 12 points.

The satisfaction rate for each testing topic was similarly graded on a scale of 1-5. These grade transformations were presented in Paper V. The error grades and satisfaction rate grades were summed for healthcare and social welfare separately to count the final points for these measures. For effectiveness, the percentage of successfully completed tasks, and for satisfaction the SUS score was calculated over all testing topics for healthcare and social welfare separately and transformed into grades on a scale of 1 to 5, but these transformations were not presented in any of the papers. All grade transformations can be seen in Table 6 (next page). For learnability the score was not transformed into a separate grade as it was on a scale of 1-5. In the procurement, the fairness of the different grades between vendors was ensured by calculating statistical significance for the results for each score that fell on adjacent grades. These calculations were based on those presented in Sauro and Lewis (2012) and are not part of this dissertation.
Results: Usability Evaluation Methods and Procedure for Public Procurement

Table 6. Grade transformations for effectiveness, (lack-of) errors and satisfaction.

<table>
<thead>
<tr>
<th>Attribute (measure) / Grade</th>
<th>Calculated</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectiveness (Percentage of successfully completed tasks ($v$))</td>
<td>Over all topics</td>
<td>$v &gt; 90$</td>
<td>$90 \geq v &gt; 80$</td>
<td>$80 \geq v &gt; 60$</td>
<td>$60 \geq v &gt; 30$</td>
<td>$v \leq 30$</td>
<td>-</td>
</tr>
<tr>
<td>(Lack-of) Errors (Error points ($v_t$))</td>
<td>Per topic</td>
<td>[0,1]</td>
<td>[1,2]</td>
<td>[2,4]</td>
<td>[4,6]</td>
<td>6,∞</td>
<td>12</td>
</tr>
<tr>
<td>Satisfaction (Satisfaction rate ($v_t$))</td>
<td>Per topic</td>
<td>$v_t &gt; 80$</td>
<td>$80 \geq v_t &gt; 60$</td>
<td>$60 \geq v_t &gt; 40$</td>
<td>$40 \geq v_t &gt; 20$</td>
<td>$v_t \leq 20$</td>
<td>-</td>
</tr>
<tr>
<td>Satisfaction (SUS score)</td>
<td>Over all topics</td>
<td>$score &gt; 78.8$</td>
<td>$78.8 \geq score &gt; 72.5$</td>
<td>$72.5 \geq score &gt; 62.6$</td>
<td>$62.6 \geq score &gt; 51.6$</td>
<td>score $\leq 51.6$</td>
<td>-</td>
</tr>
</tbody>
</table>

4.4.3 Evaluating The Methods for Procurement

The value of a usability evaluation method during IS procurement comes from being able to determine the level of usability in a way that the ultimate goal of comparing the systems can be fulfilled. In Paper II, we argued that the ability to assess the level of usability of the evaluated system correctly, comprehensively, and efficiently is important when evaluating the performance and suitability of the methods to IS procurement. Correctness includes reliability and validity, comprehensiveness means assessing the system as broadly as possible, and efficiency is determined by the resources used. Furthermore, other criteria and goals depending on the viewpoint could be set to the methods: level of expertise in usability (and in the domain) needed for use, user-involvement, and ability to produce information for the implementation phase post selection.

In Paper II, the performance of the HED method was assessed by calculating inter-rater agreement (IRA) and inter-rater reliability (IRR) based on the grades given by individual usability researchers applying the method. The usability researchers gave same or adjacent grades in all scenarios, and the whole range of grades was in use. The defined grey areas near the grade boundaries came also to use. The resulting IRA was high (70.8%) indicating substantial agreement and confidence in the grades being correct. For IRR, Kappa (0.696; $p < 0.001$ and $z = .000337$ and $z$-value = 3.59) and ICC (0.73; $p < 0.001$, lower-bound = 0.69 and upper-bound = 0.91) were both indicative of substantial reliability of the grades. For greater validation of the method the total points for usability issues should be considered.

In Paper II, we also compared the grades produced by HED with the answers to the DPUQ statements to determine if they were aligned. For the purposes of this comparison, all DPUQ statements were treated as equal, and all responses were included in the analysis. For each scenario an arithmetic mean
of all answers was calculated. The calculated correlation (0.79; p<0.001, df = 22 and t = 6.097) indicated that end-user's DPUQ answers and usability researchers HED grades were well aligned. Although, it seemed that the evaluations by end-users were slightly higher overall.

All three methods produced results revealing quantifiable differences between the evaluated systems, and thus satisfied the ultimate need for procurement purposes. In Papers IV and V the differences between the systems identified with paired-user testing were presented in more detail. Overall, the difference between the systems in both healthcare and social welfare was present in percentage of successfully completed tasks, error points and satisfaction feedback rate consistently in favor of one vendor. The calculated correlation for the measure comparisons in Paper V indicated good correlation. A more detailed look into the results in healthcare test topics in Paper IV showed that the difference between systems was revealed also within each test topic and measure, although not always so apparent. This held true also for learnability and SUS score.

However, it could also be seen, that the relationships between different measures during paired-user testing were not clear. For example, a similar rate of successfully completed tasks and errors in two different test topics and systems (59.0% and 2.6±0.6 vs. 62.9% and 2.9±0.2) was paired with very different levels of satisfaction rates (28.2% vs. 71.4%) and SUS scores (32.5±12.2 vs. 70.0±11.6). Perceived learnability differed somewhat (22.9±19.7 vs. 35.4±15.2). The satisfaction measures were, however, at a more similar level with each other in the same system between test topics. In another example with different test topics and the same system, the completion rates were again at a similar level (50.0% vs. 52.4%), but errors differed (3.0±0.7 vs. 1.3±0.3) and so did the satisfaction rates (47.2% vs. 26.2%) and SUS scores (25.4±16.1 vs. 41.3±12.1). However, perceived learnability stayed exactly the same (27.1±13.3).

**Comprehensiveness and efficiency** cannot be assessed independently of each other. If the preparative work that is naturally needed for all methods, to prepare the scenarios for example, is excluded, the evaluation and analysis can be considered as follows. In Paper II we analyzed these efforts for **HED** and **DPUQ**. Comprehensiveness of HED and DPUQ was the same, as the same scenarios and tasks were used to demonstrate the system. As for efficiency, with HED, the evaluation and aggregation of results per vendor with two evaluators took 48 hours and covered ten different use contexts. This made roughly five hours per use context. With DPUQ, the total hours for end-users attending the sessions were 396.5 hours per vendor when considering
the same ten use contexts. This made roughly 40 hours per use context. However, the number of end-users varied between scenarios and if we only considered the minimum number of nine end-users for each scenario the hours per use context would have been 16. (Although considering that some scenarios had multiple use contexts evaluated, this would have meant some end-users evaluating use contexts that are outside their primary expertise.) These calculations made HED more efficient than DPUQ by 68%. It should be noted that end-users did not solely use their effort during the demonstration for usability and evaluated another aspect used in the procurement. Still, if their time would have been divided in half, resulting in 8 hours per use context, the hours would still have been in favor of HED. In Paper II, HED was also compared to traditional heuristic evaluation (10 tasks evaluated), and it was revealed that HED was more comprehensive (more tasks per use context were evaluated) and efficient (26% more efficient even if number of tasks per context was not considered).

For **paired-user testing**, such calculations were not addressed in the attached papers. However, they can be easily derived from the information given in Papers IV and V. The comprehensiveness can be considered as follows: the evaluation covered five test topics (use contexts) with number of tasks ranging from 10 to 19 per topic where each session took 90 minutes. This is similar to traditional heuristic evaluation per scenario, and the coverage was significantly less than with HED and DPUQ. For paired-user testing, to evaluate one use context per vendor the effort of two users and one moderator was needed for three sessions (90 minutes each) which makes a total of 13.5 hours. As the data was gathered electronically during the testing and analysis was mostly automated, another 30 minutes per use context could be added in our project, adding up to 14 hours. As such, this makes paired-user testing less efficient than HED and DPUQ both. If documenting the gathered data for analysis needed to be done manually, then paired-user testing would have been even less efficient in this regard. However, paired-user testing did produce data for several attributes of usability, unlike either HED or DPUQ which each evaluated one attribute and produced 1-2 measures. In our case, we evaluated four attributes and five measures with paired-user testing.

The three methods differed in level of expertise in usability and domain needed for use, user-involvement, and ability to produce information for the implementation phase. The first two aspects for each method were presented in the earlier sections of this chapter. Both HED and paired-user testing needed expertise in usability and knowledge of the domain for preparation and implementation. DPUQ could be used without either expertise if scenario
preparation is excluded. HED is the only one of the methods which does not include users. The documentation in HED could be used for the implementation phase, while neither paired-user testing nor DPUQ provide qualitative data. Paired-user testing could however be used to leverage qualitative data on the encountered usability issues.

4.5 Constructing the Procedure for Combined Scoring of Usability

The selection phase of procurement can include multiple stages of evaluation with different objectives. The objectives and goals for the evaluation and the procurement drive the selection of measures and methods as well as the definition of scoring and weights. In this section, I will go through how the evaluation procedure was constructed for the procurement. I will also go through what was the combined scoring for usability based on several usability attributes.

4.5.1 Context and User Group when Planning the Evaluations

Identifying relevant use contexts, including user groups, is central to assessing usability. The two main user groups for the evaluation had been identified in the very beginning of the project: professionals providing the services and citizens using the services. These main user groups can be further divided into more specific user groups, and especially for professionals there are a multitude of specific user groups that could still be further specified into subgroups: such as physicians, nurses, and social workers. In the procurement project, there were two major use context areas: healthcare and social welfare. Within these both there were several more precise use contexts, such as intensive care and child welfare.

In Paper I, we presented how usability attributes were selected for evaluation in the procurement project by defining objectives of usability for the user groups based on the established overall goals of the procurement. The usability objectives for the main user groups were defined before I joined the team in collaboration with the first usability researcher in the project and domain experts working in the procurement. These objectives and their linkage to the goals of the procurement are presented in Table 7 (next page). In the table, the usability attributes can be seen expressed within the objectives for the two main user groups.
Table 7. Objectives for the evaluation linked to the goals of the procurement (revised from Paper I). The numbers signify the priority of the goals linked to the objective.

<table>
<thead>
<tr>
<th>Goals of the procurement</th>
<th>Unified service and care pathways</th>
<th>Cost-effectiveness and quality</th>
<th>Data driven management and development</th>
<th>Client / patient in the center</th>
<th>Satisfied users</th>
<th>New innovative services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved efficiency and effectiveness</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced number of errors</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency of taking the system in use: learnability</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased user satisfaction</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Objectives for professionals

Objectives for citizens

- Increased user satisfaction
- Increased efficiency
- Fluency of taking the system in use: learnability
- Reduced number of errors
- Accessibility of electronic services

The importance of understanding the use context, user groups and their tasks and goals when planning the user scenario-based demonstrations and the scenarios for the paired-user testing was highlighted in Papers II-IV. As presented in the papers, domain experts working at the procurement office were largely responsible for making decisions related to the context of use. Also, for example, in HED domain experts were needed to identify and explain the most essential functionalities to be demonstrated with the user scenario to be then taken into consideration during the usability evaluation. Additionally, domain experts were needed to support in preparing to moderate paired-user tests, especially to determine the successful completion of tasks. In addition to domain experts, user representatives were also involved in identifying critical and frequent tasks through writing of user stories describing working with clients and patients as mentioned in Paper I.

The methods presented in this dissertation were developed first using previous knowledge and literature on usability evaluation and procurement in the healthcare context. The primary reason, as also evident in the literature
presented in Chapter 2, was the lack of similar studies in the social welfare context. For overall methodology selection and the development of HED (presented in Paper II), the recommendations from procurements in healthcare context were utilized. When developing DPUQ (presented in Paper III), one of the existing questionnaires utilized was the questionnaire instrument developed for usability in a clinical context for physicians (Viitanen et al., 2011; Kaipio et al., 2017). During procurement DPUQ was used also for social welfare.

In Paper VI, we explored the experiences of physicians, nurses, and social welfare professionals on their IS on a national scale in Finland. These systems included also the systems being replaced by the procurement project. The users were working in similar open healthcare and social welfare contexts. All user groups expressed dissatisfaction with the overall usability of the used information system, although there were differences between them on the details. Physicians especially, and healthcare workers overall, rated the usability of their IS lower than social welfare professionals. With regards to traditional UI design aspects (logical placement of functions and terminology) there were no major differences between the groups. The study also found that the ISs failed in the contextual aspects of usability, with supporting users in their work, for all user groups. In these aspects, there were differences on how dissatisfied the groups were. For example, support for routine tasks was considered very low by physicians, as was documenting patient information by specialized care healthcare professionals, compared to social welfare professionals. For IS support on collaboration and information exchange, healthcare professionals rated it to be good within the same organization while social welfare professionals did not agree. For cross-organizational collaboration all groups were dissatisfied, however, social welfare professionals were considerably more so.

4.5.2 Selecting Methods and Measures

In the procurement, the evaluation procedure was organized in two phases as presented in Paper I. The preliminary assessment (phase A) and the short-list evaluation (phase B) had differing goals for the evaluation. In the first phase the goal was assessing the relative usability of the products and ensuring that it reached a certain level for key functionalities. In the second phase the goal was to go deeper in the evaluation, measure the usability so that the systems could be compared and present the results so that they could be part of the final points that the procurement decision was based on. The overall goals of
the evaluation in each phase affected method selection together with the usability objectives of different user groups (presented in the previous subsection) which in turn affected how much focus was put on the different usability attributes and what the selected measures were.

In phase A, it was considered important to cover several central areas of system use for the major user groups in key use contexts and involve users extensively. The details of the systems were not as important because of the main objectives of this phase. Because of this, methods that extensively covered the system and involved users while still being relatively efficient were deemed needed. This led to development of HED and DPUQ. In phase B, the goals were getting a more detailed evaluation on key functionalities and getting better quantified evidence on usability. This led to choosing paired-user testing as the main method. All chosen usability attributes, methods and measures in each phase are presented in Table 8.

### Table 8. The usability attributes, methods, targeted user group and measures (revised from Paper I)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Evaluation method</th>
<th>User Group</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction</td>
<td>DPUQ (Usability questionnaire in scenario-based demonstration)</td>
<td>Professionals</td>
<td>Perceived usability: professionals’ responses during and after session (DPUQ score 1 and score 2)</td>
</tr>
<tr>
<td>Quality of UI design</td>
<td>HED (Heuristic evaluation in scenario-based demonstration)</td>
<td>Professionals</td>
<td>Documented usability issues: heuristic violations, missing functionalities, omitted parts of the user scenario and positive findings (score)</td>
</tr>
<tr>
<td></td>
<td>Task-based heuristic evaluation</td>
<td>Citizens</td>
<td>Documented usability issues: heuristic violations, missing functionalities, and positive findings (score)</td>
</tr>
<tr>
<td>Phase B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effectiveness</td>
<td>Paired-user / usability test</td>
<td>Professionals / Citizens</td>
<td>Percentage of successfully completed tasks</td>
</tr>
<tr>
<td>Errors</td>
<td>Paired-user / usability test</td>
<td>Professionals / Citizens</td>
<td>Errors made by the user during task completion</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Expert evaluation</td>
<td>Professionals, Citizens</td>
<td>Number of steps in the optimal solution to tasks</td>
</tr>
<tr>
<td></td>
<td>Interactive scenario-based demonstration (Group inspection session)</td>
<td>Professionals</td>
<td>Usability specialist’s assessment of efficiency</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Usability specialist’s assessment of efficiency of configuring the system</td>
</tr>
</tbody>
</table>
Attribute | Evaluation method | User Group | Measure |
---|---|---|---|
Learnability | SUS questionnaire (in paired-user / usability test) | Professionals / Citizens | Perceived learnability: learnability factor of SUS* from user’s responses after task completion |
| Interactive scenario-based demonstration (group inspection session) | Professionals | Usability specialist’s assessment of learnability based on professional’s verbal answers |
Satisfaction | Paired-user / usability test | Professionals / Citizens | Useability specialist’s assessment of learnability of configuring the system |
| SUS questionnaire (in paired-user / usability test) | Professionals / Citizens | Positive and negative markers given by users during task completion |
| Preference ranking (in usability test) | Citizens | User’s responses after task completion (SUS score) |
| DPUQ (summative part in group inspection session) | Professionals | Perceived usability: professionals’ responses after session (DPUQ score 2) |
Quality of UI design | Interactive scenario-based demonstration (Group inspection session) | Professionals | Usability specialist’s assessment based on usability heuristics and professionals’ discussions during demonstration |

* SUS factor structure (Lewis and Sauro, 2009)

4.5.3 Combining the Scoring within Attributes and between Attributes

In both phases of the procurement usability was only one of the evaluated aspects of the vendors. As stated in Paper I, in phase A 15 % of the points came from the usability evaluations and in phase B 40 % of the points for evaluations came from usability (equaling 12 % of the total points). In both evaluation phases during the procurement project, the total score for usability was determined by giving each usability attribute a relative weight.

In phase A, there was the additional requirement of reaching a minimum level of usability which was defined as reaching a grade above 0 (Fail) in each evaluated user scenario for the “quality of user interface design” attribute. This attribute was evaluated with the HED method and traditional heuristic evaluation depending on user group. Both methods utilized the same grading system which is presented for HED in subsection 4.4.1 and in Paper II. Essentially, it can be thought to correspond reaching 10 % of maximum points (if maximum points equal best quality). The total usability points were calculated by weighting the two usability attributes for each user scenario so that “quality of user interface design” was given a weight of 2/3 and “satisfaction” a weight of 1/3 (see Table 9, next page). For satisfaction the two scores from DPUQ were weighted equally. The calculation required scaling all scores to the same scale. The scenarios had also individual weights. While this was not presented
Table 9. Weights given to usability attributes in phase A

<table>
<thead>
<tr>
<th>User scenario</th>
<th>Weight for “quality of UI design”</th>
<th>Weight for “satisfaction”</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCD1</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>HCD2</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>HCD3</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>HCD4</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>HCD5</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>HCD6</td>
<td>2/3 (from traditional HE)</td>
<td>1/3</td>
</tr>
<tr>
<td>SWD1</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>SWD2</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>SWD3</td>
<td>2/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

In any of the papers, the sum of points over all user scenarios was calculated to get the final score.

In phase B, the usability attributes were given different weights depending on user group. These weights were presented in Paper I and can be seen in Table 10. Within those attributes that had several measures (efficiency and learnability, and satisfaction), the measures were given different weights within the attribute. Likewise, as in phase A the scores were scaled to the same scale and then summed for each user group. The final score in phase B was calculated by weighting the results for professionals in healthcare context, professionals in social welfare context and citizens using the patient portal differently. These weights and calculations were not presented in any of the papers.

Table 10. Weights given to usability attributes in phase B (as presented in Paper I)

<table>
<thead>
<tr>
<th>Usability attribute</th>
<th>Weight for professionals</th>
<th>Weight for citizens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effectiveness</td>
<td>30 %</td>
<td>25 %</td>
</tr>
<tr>
<td>(Lack-of) Errors</td>
<td>20 %</td>
<td>10 %</td>
</tr>
<tr>
<td>Efficiency and Learnability</td>
<td>20 %</td>
<td>10 %</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>20 %</td>
<td>25 %</td>
</tr>
<tr>
<td>Quality of UI design</td>
<td>10 %</td>
<td>-*</td>
</tr>
</tbody>
</table>

* 30 % from accessibility; evaluated to complement usability evaluation
5. Discussion

This dissertation highlights the methods, procedure, and considerations for including usability measurement in the public procurement of IS. By taking an action design research approach for the study, it was possible to develop a satisfactory solution to the research problem emerging from a real-life project as well as apply and evaluate the solution in a real context. In the following, I will present and discuss the overall contributions of this dissertation.

5.1 Answering the Research Questions

In Chapter 4, the main results of this dissertation work were presented. In the following, I will answer the research questions based on the findings.

5.1.1 Suitable Methods and Metrics

The first research question was:

*RQ1: What are effective evaluation methods and metrics for measuring usability during the selection phase in public procurement of IS?*

The first research question was addressed both based on studying previous literature and by implementing both existing and new methods and metrics in the action design research study of a public procurement project. Quantitative usability evaluation includes many commonly used measures for quantifying the usability attributes, especially with users, by using questionnaires for satisfaction and usability testing (Hornbæk, 2006; Sauro and Lewis, 2012). There are several established usability evaluation methods that are suggested and have been previously applied in procurement of HIS. Especially standard usability questionnaires (HIMSS EHR Usability Task Force, 2010), heuristic evaluation (Carvalho, Borycki and Kushniruk, 2009; Kushniruk et al., 2010),
usability testing and on-site testing (Beuscart-Zéphir et al., 2005; Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014) were mentioned, with the latter two of the four providing strongest evidence on usability (Kushniruk et al., 2010). Utilizing clinical information processing scenarios to enhance used evaluation methods was also promoted (Kushniruk et al., 2010). However, the literature was lacking on concrete examples of measures in this context and partly utilized procedures that are not suitable for public procurement which requires transparency and unambiguous scoring of criteria, as discussed in the Synthesis of the Theoretical Foundation in section 2.4. Also, the suggested methods are resource intensive to cover sufficiently the early stage of public procurement that might include numerous vendors to evaluate and in the case of complex IS environments multiple use contexts with several user groups and tasks.

We developed two new methods, HED and DPUQ, for the preliminary assessment stage to address the challenges presented above and implemented them into use in the procurement project. Both these methods were designed for use in vendor demonstrations that were based on prescribed user-scenarios. This enabled the evaluation to cover a very substantial number of use contexts and user tasks in a confined timetable. HED modified the well-known heuristic evaluation method to include aspects that are not traditionally accounted for, such as missing functionalities, but are important in the procurement context. It also produced a score and grade for each evaluated scenario to rank the systems in their quality of UI design. HED was also efficient to use, requiring only two evaluators and covering a larger scope than traditional HE. DPUQ was based on established usability questionnaires and transformed the point of view of potential end users from using the system to perceived usability based on seeing the system being used. It also produced numerical scores. Compared to established usability questionnaires there are also added questions related to context of work.

In the second short-list evaluation phase, we used usability testing and paired-user testing, SUS and DPUQ questionnaires, group inspection sessions, and expert evaluation to collect data. The measures used included users’ views, usability experts’ assessments and users’ performance. While user testing itself is an established method, the scoring functions for the usability attributes for selection purpose were presented in more detail in this research than previously.

Overview of all the methods, metrics, and measures in the procurement project is presented in section 4.1 of Results: Usability Evaluation Methods and Procedure for Public Procurement. HED, DPUQ and paired-user testing are
presented in more detail in section 4.2 and their scoring functions for the usability attributes in section 4.3. The application and suitability of the methods for this purpose is evaluated in section 4.4.

Based on the conducted research, all three more closely examined methods, and the metrics and measures they used, are suitable for public procurement (see Table 11 for comparison). The results of different metrics seem overall to be aligned with each other, for both HED and DPUQ in the preliminary assessment phase, as well as user testing in the short-list evaluation phase. However, there are differences especially between objective and subjective metrics in paired-user testing as was to be expected based on literature (Bailey, 1993; Kissel, 1995; Hornbæk and Law, 2007; Jorritsma, Cnossen and van Ooijen, 2014). For more details, see subsection 4.4.3.

Table 11. Comparison of the three usability evaluation methods for procurement.

<table>
<thead>
<tr>
<th>Aspects of the method / Method</th>
<th>HED</th>
<th>DPUQ</th>
<th>Paired-User Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usability attributes</td>
<td>Quality of UI design</td>
<td>Satisfaction</td>
<td>Effectiveness, Errors, Satisfaction, Learnability</td>
</tr>
<tr>
<td>Usability metrics</td>
<td>Documented usability issues</td>
<td>Perceived usability during and after demonstration</td>
<td>Successfully completed tasks, errors made, feedback during task completion + overall and learnability SUS scores</td>
</tr>
<tr>
<td>Usability measures</td>
<td>One score</td>
<td>Two scores</td>
<td>Percentage, points, rate, SUS scores</td>
</tr>
<tr>
<td>Comprehensiveness, Efficiency (excluding preparation)</td>
<td>Large scope with minimal resources, ~5 h / use context</td>
<td>Large scope with larger resources, ~8-16 h / use context</td>
<td>Focused scope with larger resources, ~14h / use context</td>
</tr>
<tr>
<td>User involvement</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Level of usability expertise needed</td>
<td>High + needs domain experts</td>
<td>Very limited (+ needs domain experts for scenarios)</td>
<td>High + needs domain experts</td>
</tr>
<tr>
<td>Ability to produce information for implementation</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The methods do differ in their efficiency and comprehensiveness. HED can be used to cover a large scope quite efficiently, while using both DPUQ and user testing would need more resourcing for the same scope. Efficiency and comprehensiveness are important when there are possibly many systems with multiple and complex use contexts to evaluate. All the methods also review different attributes of usability with their metrics. Furthermore, other valid viewpoints for the methods include user-involvement, level of expertise in usability needed for use, and ability to produce information for the implementation phase. The latter two methods have the advantage of involving future end users, which is needed for a successful procurement according to literature (Moe, 2014). The level of usability expertise needed for using the methods differs. For HED and paired-user testing, high level of expertise is needed for conducting the evaluation. For DPUQ, however, the level needed is limited to only determining suitable placement of questionnaire within the scenario. Additionally, also domain experts are needed for all methods, for preparing the scenarios in all and additionally for evaluation considerations in HED and paired-user testing. Ability to produce information of usability issues for implementation is also something that sets HED and paired-user testing apart from DPUQ, although for paired-user testing this would need additional resourcing for documenting usability issues. All these aspects need to be considered when choosing suitable methods for each procurement case, so that the choices are in alignment with the goals of the procurement.

5.1.2 Construction of Evaluation Procedure

The second research question was:

*RQ2: How to construct the evaluation procedure supporting the quantification and comparison of usability for this use?*

To answer the second research question, we explored recommendations and insights from literature and then devised the evaluation procedure used in the procurement project. In the project, the choice had been made to use two different phases for the evaluation and include the usability criterion as one of the minimum requirements in the first phase. A two-stage procedure is also suggested in the literature. The goals of these two phases were different for the quantification of usability. The goals of first ensuring the minimum requirement for level of usability was achieved, and then ranking the systems to enable reducing their number before entering the second phase, generated certain demands for the used methods and their scoring.
With HED (and traditional heuristic evaluation with similar scoring) ensuring the required level is achieved, is possible by determining beforehand the score/grade needed for each scenario. In the first phase, it is also necessary to get an extensive overview of the system, so that passing the minimum requirement is not based on a small part of the system. Overall acceptance from future users can be achieved with DPUQ, since professionals from many different user groups can be included. These methods targeted quality of UI design and satisfaction in the preliminary assessment phase. These are both also needed so that a more focused evaluation is possible in the second phase.

For both phases, the attributes of usability to evaluate are determined based on the goals of the procurement. Existing literature suggested using several different metrics (Sauro and Lewis, 2012), both objective and subjective, to ensure that possibly contradictory evaluations would be accounted for. We used this approach specifically in the short-list evaluation phase of the procurement project, where altogether thirteen measures were used to evaluate usability of six different usability attributes. The evaluation framework used in the procurement project is presented in detail in subsection 4.5.2.

From a legislative perspective, the planned evaluations need to be transparent and treat the vendors equally and in a non-discriminatory way. This includes planning the evaluations and their details, for example the guidance for and information exchange with vendors, with ample time before the implementation. The evaluation criterion also needs to be stated in the beginning of the procurement and its weight in the overall evaluation be disclosed according to the legislation. This includes stating the minimum requirements for the first phase. Literature also argues that this weight cannot be determined in a predictable way that reflects the procurer’s values without first knowing the scoring rules behind each criterion and then using this information to determine the trade-off relationships of the criteria with existing methods and mathematically solve the equation system (Mateus, Ferreira and Carreira, 2010). This same thought process could be applied to determining the weights of the usability sub criteria (usability attributes) and their different metrics.

The action design research done in the procurement project presents one unique way of quantifying overall usability based on the individual metrics. The weighting scheme for usability in the first phase was fairly simple, with quality of UI design and satisfaction having the same weights for each scenario (except those that only evaluated satisfaction), and the grades from HED and scores from DPUQ scaled to the same scale before calculation. In the second phase the weighting scheme was more complex and included several levels: weights between user groups, weights between attributes for each user group
and weights between measures within attributes for each user group. The scores for the measures were scaled based on the weight. Some details of this weighting are presented in subsection 4.5.3. The weighting presented in this dissertation is only an example, and the weighting scheme is unique for each procurement to reflect the goals of the procurer and for overall usability, as is also stated in literature for the overall scoring of criteria (Mateus, Ferreira and Carreira, 2010).

In the next section, I present a process model for defining and planning the usability evaluations for a procurement project. This model also addresses the additional theme introduced in section 1.5 Research Questions: the implications of contexts of use on the procedure.

5.2 Usability Scoring Model

I have created a Usability Scoring model (USco) combining the findings from the literature review and results from the action design research done in the procurement project. This model was influenced by depiction of the tender evaluation methodology building blocks in Mateus et al. (2010). The USco process model includes the components for defining and planning a two-stage usability evaluation for the selection phase. The model assumes that a price-to-quality scoring is used in the procurement, as it is the common approach.

Usability goal setting, where the contexts and general user groups for the procurement are first identified, forms the basis for the model. The usability attributes important to these identified contexts and groups are then linked to the goals of the procurement to form the basis for the usability evaluations.

For a two-stage evaluation procedure, two different scoring models are needed: one for the ranking of systems based on usability and one for the usability measurement. The usability ranking model defines the scoring for the preliminary assessment stage. The usability measurement model defines the scoring for the short-list evaluation stage. The USco model also includes the steps for more detailed planning of the usability evaluation in the short-list evaluation stage. The USco model is presented in Figure 9 (next page). In the following, a general overview of the steps in each of these components are presented.
Defining: Usability Ranking Model

The purpose of the usability ranking model is to define how the systems are ranked in order for usability and what is the minimum level of usability required. The suggested methods for evaluating usability at this stage are HED and DPUQ, and the assumption is that satisfaction and quality of the UI design are usability attributes that are relevant for the procurement. In a minimal evaluation procedure, this could be the only usability evaluation procedure and influence the final selection.

1. Selecting contexts and user groups: Because not all identified contexts and user groups can be covered during evaluations, the first step is to select a wide enough and representative range of those identified in the previous phase (Usability goal setting) for this preliminary evaluation to cover the system. It is advisable to do this evaluation in conjunction with other evaluations concentrating on the functionalities, so this step is a collaborative step. It is also assumed that the user scenarios for the demonstrations are defined by domain experts and creating these is thus not included in the process model.

2. Defining points & grade boundaries (HED): Once the user scenarios are scripted, the points and grade boundaries for HED can be determined for each scenario. This can be based on piloting the method with
currently used systems. The grade boundary estimation process used in the procurement project seemed to work quite universally for different length scenarios for different user groups in the healthcare and social welfare context. More details can be found in subsections 4.3.1 and 4.4.1.

3. Defining the minimum required level: The minimum required level of usability is determined based on the grade boundaries for HED. As an example, the grade ‘fail’ was set in the procurement project to reaching 90% of the absolute minimum points for each scenario. Thus, the minimum required level was receiving a grade above ‘fail’ for each scenario. This sets the bar quite low for usability, and the level could be higher in other projects, especially if the final selection is done based on this evaluation or there are no subsequent evaluations for usability.

4. Defining the weights for HED and DPUQ: Finally, the weights for HED and DPUQ need to be defined for each scenario to be able to compile the evaluation into one score and rank the systems in order. Each scenario could also have its own weight. By applying the recommendations in literature for weighting different criteria (Mateus, Ferreira and Carreira, 2010) to determining weights for sub criteria, the weights should reflect the trade-off relationship between the attributes and scenarios. This requires posing specific questions to the decision makers of the procurement and as an example comparing plausible examples of tenders with significantly different performance levels using the scoring. More details on the suggestions for how this should be done are in subsection 2.1.2 Determining Weights.

Defining: Usability Measurement Model
The purpose of the usability measurement model is to define the specifics of quantifying usability into one score to be used in the final selection. The suggestion is to use at least usability testing for measuring usability in this stage.

1. Selecting contexts and user groups: For this more focused evaluation, a scope of use contexts and user groups need to be selected and defined from those identified. This could be narrowed down further than in the previous stage, depending on the complexity of the procurement. The most relevant, common, and representative contexts should be considered.

2. Selecting usability attributes for evaluation: The usability attributes that are to be evaluated are selected from those identified in the goal setting stage. The importance of each attribute and evaluating it could
be different for different user groups and contexts and this determination should be based on domain knowledge.

3. Defining the usability metrics and measures: Next step is defining meaningful metrics and measures for each usability attribute (and for each user group and context), that are also possible to evaluate during a procurement. These should be selected to evaluate each attribute as extensively as is considered feasible. In addition to typical measures for usability, such as rate of successfully completed tasks, other options are presented in subsection 4.5.2.

4. Defining the usability methods: In this step, suitable methods for the procurement context are selected to evaluate the metrics and measures reliably. The efficiency and extensiveness of the methods are also considerations. Some other methods besides the most common, usability testing, are presented in subsection 4.5.2.

5. Defining the scoring rules: Once the metrics and measures as well as the methods are defined, the scoring rule of each method-measure combination should be set. The purpose of a scoring rule is to convert a measure into a partial score on a numeric scale that describes the level of performance on that measure. This is discussed in more detail based on literature in section 2.1.2 (in general) and section 2.4 (related to usability). Scoring rules used in this dissertation for the selection phase (for DPUQ and paired-user testing methods) are presented in section 4.3.

6. Defining the weights within/between usability attributes: Finally, determining the weights between the different measures within an attribute, between the attributes, and also between the different user groups and contexts is needed. This determination should be based on modeling the end-result with the pre-defined scoring rules, so that it reflects the procurer’s preferences for valuing the different usability attributes (as also discussed in step 4 of defining the usability ranking model). The general process for determining weights discussed in literature is presented in subsection 2.1.2 Determining weights.

Planning: Usability Evaluation Procedure
This component is meant for supporting the more detailed planning of usability evaluation in the short-list evaluation stage. It outlines the key steps needed.

1. Defining contexts and sub-user groups: To start planning the evaluations in more detail, a group of focused sub-contexts and sub-user
groups need to be selected and described. This selection should be based on domain knowledge and the desired coverage (and available resources) for the evaluation during this stage of the procurement.

2. Identifying key user goals and scope: Central user goals (critical and most frequent) should be identified for each user group and use context and the evaluation should focus on these. Also, goals that are generalizable to other user groups should be considered. This selection should be based on domain knowledge, and preferably also user research.

3. Defining scenarios and tasks for the evaluation: When the goals and scope of evaluation are selected, the actual scenarios and user tasks should be scripted. This scripting should utilize domain expertise. The tasks should be written to be system-independent and include enough background information.

4. Piloting the procedure: If possible, all scripts should be piloted with the system(s) currently in use with users representing the specified user group. This piloting will inform the finalising of the scenarios and tasks as well as the preparations needed for the evaluated system.

5. Specifying the details for/from vendors: During procurement, all information needs to be specified beforehand. The piloting stage will inform what background information needs to be in the system before evaluation can be done. This information needs to be provided to the vendor by the procurer. To perform the evaluations on unfamiliar systems, instructions on how to use the system as designed should be received from the vendor. These instructions can be used to guide the moderating of usability tests and user performance assessment. Also receiving training materials for users can be necessary.

6. Familiarizing with vendors’ systems pre-evaluation: If usability testing is used, the moderators need to plan for time to familiarize themselves with the system before the evaluation so they can assess the users’ performance in achieving the goals. This can be done with one-on-one sessions with the vendor and based on the provided instructions.

Other Considerations
The relevance of understanding the use context, including user groups and their tasks and goals, cannot be emphasized enough when planning which areas should be evaluated in the first place. It is essential to identify the key use contexts and critical and frequent tasks for users, especially considering what areas are most affected by using the IS to be procured. This understanding and
complexity can also affect the choice of suitable evaluation methods, as mentioned in section 4.1. The dissimilarities in the deficiencies related to usability experienced by different user groups (described in Paper VI and outlined in subsection 4.5.1) indicate that the needs for improvement with the procured system are non-identical. This could mean focusing on different aspects of the systems during usability evaluation based on user groups or context.

The relationship and collaboration between usability researchers and domain experts (physicians, nurses, and social workers, for example) working in the procurement project is also essential in several points of the process. In a complex systems context, such as healthcare settings, this is even more important, as domain expertise is highly specialized and not transferrable to usability professionals. In addition to domain experts, also working with end users is necessary.

5.3 Practical Implications

For Procuring Organizations Overall

The results of this dissertation work can benefit all procuring organizations to some extent. Based on the results, usability is a possible measurable criterion for selection of systems. While the methods, framework for evaluation and the scoring of usability were developed for public procurement purposes, they could be adapted to also private procurements. The methodology is mainly context independent, so most of it could be used also outside the healthcare and social welfare sector in other complex procurements.

All individual methods produced numeric results that could be used to compare the systems and the results from applying the process showed alignment between the different used usability evaluation methods and metrics overall, although there were contradictions with metrics used in paired-user testing. This would allow for selecting even only one method to get a reasonable enough end-result for determining which system scores highest on usability. However, based on literature using a few different methods is typical (see Table 1 in subsection 2.2.2), or recommended, especially in complex projects (Beuscart-Zéphir et al., 2005). The experience from the procurement project highlights that the selection of methods should also consider factors other than purely the efficiency of the methodology, such as including end users in the selection process and getting a more in-depth understanding of the system and its potential usability problems. In line with previous literature (Kushniruk et al., 2010), our recommendation is using usability testing in the later stages of selection. The number and selection of specific measures, however,
Discussion

depends on several factors such as the importance of the usability attributes for the different user groups.

For Procuring Public Organizations
The procurement project in question showed that including usability evaluation as a selection criterion can be done in a public procurement and that the evaluation can have a significant effect on the selection in both stages. During the procurement, there was a market court appeal on the selection, however the usability methodology did not give reason for the appeal nor was it questioned.

For public organizations, this dissertation work provides concrete examples, models and methods of how to include usability evaluation in the procurement process of any IS. The methodology presented in this dissertation could be adapted to other public procurements easily, and it gives a practical example of using a two-stage procedure to limit the number of vendors. The proposed metrics and scoring rules could also be adapted if the procuring organization considers them aligned with their usability goals. For supporting the step-by-step planning of the evaluation procedure, this dissertation presented the USco model. And finally, specifically for the context of healthcare and social welfare, the dissertation work introduced a suitable user questionnaire, DPUQ.

For Vendors
The dissertation work has also practical implications for vendors. If including usability and minimum requirements for its level in the selection of IS in the public sector become standard practice, to succeed in acquiring contracts the products need to have good usability. This means emphasis on UCD is needed already in the software development phase.

For Professionals Working in Healthcare and Social Welfare and Society in General
If usability of IS is taken seriously from the beginning of public procurement projects, there exists greater potential for the end results of IS implementations to also be more usable for their users. As the efficiency of use is improved, this can mean reduced costs in the form of less time spent on using technology for completing tasks. Specifically in the context of healthcare and social welfare, this time could then be spent on caring for patients and clients. Better usability can also mean less user errors, which in healthcare improves patient safety (Koppel et al., 2005; Kushniruk et al., 2005). Problems with usability also add to the cognitive load of users which has been linked to
physician burnout (Iskander, 2019). However, the selection phase of the procurement is just the basis for the rest of the IS project and focus on usability needs to remain throughout the implementation to achieve these results.

5.4 Scientific Implications

Knowledge on Public Procurement

Previous literature on public procurement (Moe, 2014) has highlighted the lack of concrete studies on projects in this area. This dissertation adds to that scarce knowledge base by presenting the work of including usability evaluations into the selection phase of a real-life public procurement.

Usability Evaluation Methodology

The need for usability evaluation methods that are suitable for the public procurement context has been brought up in previous literature (Jensen, Rasmussen and Lyng, 2013). While there exist recommendations for methodology in procurement of HIS in literature (e.g. Beuscart-Zéphir et al., 2005; Carvalho, Borycki and Kushniruk, 2009; Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014), these do not generally consider the context and specific requirements of public procurement. In this dissertation, two new evaluation methods, HED and DPUQ, have been developed for this purpose, along with using paired-user testing, and applied in practice. The use of demonstrations is a standard practice in procurement (Beuscart-Zéphir et al., 2002; Howcroft and Light, 2002) as is users evaluating the system functionality based on these demonstrations (e.g. Holbrook et al., 2003). Using usability questionnaires was also common (Beuscart-Zéphir et al., 2002; Holbrook et al., 2003; Liljegren and Osvalder, 2004; Jensen, Rasmussen and Lyng, 2013) in various phases of procurement and they have been recommended for this use in EHR procurements (HIMSS EHR Usability Task Force, 2010). However, the questionnaires used during demonstrations previously were either not available or did not evaluate usability, as is presented in subsection 2.2.2. Standard usability questionnaires, on the other hand, have been developed from the point of view of using the system as is introduced in subsection 2.3.1.

DPUQ answers the need for a usability questionnaire that makes evaluating usability based on seeing the system used possible, as opposed to using it before answering. The other developed method, HED, adapts and expands traditional heuristic evaluation (HE) (Nielsen and Molich, 1990; Nielsen, 1993), which has also been recommended for procurement purposes (Kushniruk et
It adds to the evaluation the context and recognizing missing functionalities which HE has been criticized for not considering (Chin, Diehl and Norman, 1988; Nielsen, 1993). Both these methods also utilize clinical information processing scenarios (CLIPS), which are suggested for HIS procurement (Kushniruk et al., 2010), by basing the demonstrations on them. The use of CLIPS and demonstrations broadens the possible scope of the evaluation while also being efficient, as is presented in subsection 4.4.3. The work in this dissertation also supports using of usability testing in this context, which is widely recommended in literature (Beuscart-Zéphir et al., 2005; Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014).

Also, using a two-stage model for evaluating usability in HIS procurement has been previously suggested (Schumacher, Webb and Johnson, 2009; Kushniruk et al., 2010; Jorritsma, Cnossen and van Ooijen, 2014). The work in this dissertation supports this approach also in a public procurement context, when negotiated procedure is used as the legislative basis. In the first phase the literature suggests that a minimum requirement for level of usability should be set, and relative usability should be evaluated (Beuscart-Zéphir et al., 2005; Schumacher, Webb and Johnson, 2009). The HED method supports both these goals, and DPUQ can be used for evaluating perceived relative usability. For both stages, the evaluation framework presented in section 4.5 adds details to selecting methods and measures with this approach compared to previous literature. Finally, this dissertation proposes a new process model, USco, for the steps needed in defining and planning the two-stage usability evaluation, quantification of usability and its scoring in a public procurement. The planning of a repeatable, vendor neutral and representative procedure for the evaluation is essential.

Quantifying Usability in Procurement

The quantification of usability in a comparable way is essential for the selection phase in public procurement, however, literature for implementing this is not readily available. Heuristic evaluation, which is suggested as one suitable method for preliminary evaluation (Carvalho, Borycki and Kushniruk, 2009; Kushniruk et al., 2010), is not inherently quantifiable, and existing attempts to quantify it in this context do not produce one score (Beuscart-Zéphir et al., 2002, 2005). HED presents a way to get a unified score for the usability of the evaluated system. For usability testing, which is commonly recommended for procurement, averaging the score of usability measures over all tasks is presented in relation to procurement, while traditional usability
literature considers primarily individual tasks with summative evaluation. In this dissertation work, this idea is supported by the evaluations done.

The weighted sum is commonly used in public procurement for creating the overall score from different criteria (Pictet and Bollinger, 2003; according to Mateus, Ferreira and Carreira, 2010) and it has also been discussed for combining results from usability measures (Jorritsma, Cnossen and van Ooijen, 2014). Also other ways to combine usability measures have been proposed (Lewis, 1991; Sauro and Kindlund, 2005) although generally treating them separately is recommended (Hornbæk, 2006). Procurement purposes require a way to unify the results, and this dissertation gives a concrete example of using weights for usability attributes. However, as presented in Paper V, the results of the usability attributes in this procurement seemed to correlate and weighting the components might not have been as critical. In previous literature, the idea of different weights between usability attributes depending on context of use is raised (ISO, 2018), however there is no consensus on whether this is a suitable approach (Gillan and Bias, 2014).

Planning of how the performance levels in the scoring reflect the procurer’s desires is deemed important in overall procurement literature (Mateus, Ferreira and Carreira, 2010). This dissertation presents the scoring rules used in the procurement case for the different usability measures and examples of transforming them into grades, where not all levels are equally distributed.

In the USco model, all these aspects and considerations raised in general procurement literature for weighting and scoring criteria are included in the components for defining and planning the quantification and scoring of usability evaluation.

5.5 Reliability and Validity of the Research

The relevant reliability and validity criteria for this research have been presented in section 3.6. In this dissertation work, new usability evaluation methods and a methodological framework have been iteratively developed, applied, and evaluated with the aim of solving a real-life problem of including usability in the selection of IS in public procurement while the usability researchers worked in a real-world project. This aligns with the purpose and characteristics of action design research. In the following, I will go through how the conducted research meets the relevant validity criteria for evaluating the research quality.
Process Validity
The research process in this dissertation work follows an iterative cycle that is a requirement for action design research (Sein et al., 2011). Chapter 1, Introduction, presents the problem and research questions as well as the motivations for the dissertation work that stems from a real-life context. Chapter 3, Action Design Research: Developing Usability Assessment, describes the iterative development cycles for the research. Chapters 2 and 4, Theoretical Foundation and Results: Usability Evaluation Methods and Procedure for Public Procurement, then present the actual research, the foundations and actual design, development and evaluation of the usability evaluation methodology for procurement. Finally, section 5.2 in this Discussion presents a combined solution, a model, for the research problem of including usability evaluation in the selection phase of public procurement. This model, however, has not been validated or evaluated yet.

Design Antecedent Validity
The development of the artefacts, the developed usability evaluation methods and methodological framework is based on the foundation of existing literature on previous attempts and suggestions for quantifying usability during procurement in a similar use context and the best practices of usability evaluation and measuring usability in general. The special requirements of the public procurement context have also been considered in the development. This is described in subsection 4.2.1 Developing the Methods. The process model presented in section 5.2 of this Discussion is based on the Results as well as additional previous literature presented in Chapter 2 the Theoretical Foundation.

Development and Use Context Validity
The internal validity of the artifacts, the developed usability evaluation methods and the methodological framework, presented in Chapter 4 the Results, as well as the process model presented in section 5.2 of this Discussion, can be examined by the information presented on how the artefacts are based on previously existing established usability evaluation methods, the definition of usability including the understanding of the use context, and how they measure and quantify usability. This information is presented in Chapter 4 the Results and evaluated in subsection 4.4.3. The section 5.2 of this Discussion goes through how the process model was constructed. It has been argued that for heuristic evaluation and usability testing the identified usability problems are dependent on the individual evaluators and not repeatable (Nielsen and Molich, 1990; Hertzum and Jacobsen, 2003; Dumas and Fox, 2012).
methodology the individual problems are not relevant; in HED the method concentrates on the general number and severity of problems while in paired-user testing the results gathered are quantitative based on user performance, and not qualitative.

The data input validity of the developed artefacts can be evaluated by the information on how the evaluated areas were selected and how the scenarios for the demonstrations and the paired-user testing scripts were prepared by or in collaboration with domain experts. This information is presented in the Results, in subsections 4.5.1 Context and User Group when Planning the Evaluations, 4.2.1 Developing the Methods and 4.4 Implementing and Evaluating the Proposed Methods in Procurement.

Democratic Validity
The action design research done in this dissertation was conducted by the researchers while working in the procurement project. My role as a researcher shifted from outsider to insider during the project and included extensive collaboration with other participants and stakeholders within the project. This research design and collaboration has been described in section 3.2 Research Design.

The local validity of the research can be considered strong, as the research was initiated from the procurement project and the research problem emerged from their reality. The developed solutions have also been applied in this real-life context and were found to be suitable for the purpose.

Outcome Validity
Throughout the research iterative action was taken to develop a solution to the problem: how usability can be measured during the selection phase of public IS procurement to inform the final selection. These actions are presented in Chapter 4 the Results. The understanding of the problem complexity grew during the research. The actions resulted in new developed usability evaluation methods especially suitable for this context, and a methodological framework and process model to support the planning of the evaluations. The methods were applied during the project successfully to solve the problem. Ultimately, the applicability of the results to other real-life procurement projects and their usefulness to practitioners in the field are what determine the validity of the results. The methodology developed in this dissertation work has already been partially applied (preliminary assessment phase with HED and DPUQ methods) in one other procurement project by the other usability researcher and a new practitioner.
The reliability and validity of the applied individual usability evaluation methods can also be addressed separately. This evaluation has been incorporated into subsection 4.4.3 of the Results where the applicability of the methods is assessed. While the correlation between HED and DPUQ, presented in the subsection, can be considered to be strong generally, the rather large difference could still raise the question of what is considered strong enough in the procurement context. However, in this context the most relevant consideration is the ability to determine an order between the evaluated systems. In this regard, both methods produced the same overall end result and the usability specialists’ evaluation was only less positive than users’ on the level of usability in some scenarios.

**Dialogic Validity**

The results of the research have been published in several peer-reviewed scientific papers (Papers I-V) included in this dissertation. The process model presented in section 5.2 of this Discussion has, however, not been validated formally yet.

### 5.6 Limitations

This research also has limitations. The research on procurement usability evaluation methodology is based on work done in one, although large, procurement project. The scope of this research is limited to the selection phase of the procurement project and does not consider how the results of this phase are related to the later phases of the procurement. Furthermore, I and the other usability researcher worked inside the organization developing and applying the methodology of usability evaluation and we were not objective observers. Both of these approaches, however, are in line with the action design research methodology applied. While the application and data for evaluating the methodology was gathered in one procurement project, the analysis is based on quite extensive amount of data. For HED, this means six demonstrations and 18 hours per vendor for total four vendors, comprising 48 individual evaluations by usability researchers and resulting in 144 hours of evaluated demonstration of use. For DPUQ, 771 questionnaires were filled in nine demonstrations per vendor for total of four vendors. For paired-user testing, five scenarios were evaluated per vendor by three pairs of users for two vendors which makes a total of 30 usability testing sessions.

The identification and selection of use contexts, user groups, and tasks that formed the basis of all evaluations were not based on formal user research
done by usability professionals. The project utilized domain experts for identifying and selecting the scope of evaluation. This was in part due to the strict time restrictions imposed by the procurement context. Also, the fact that the evaluation focused on a desired future state by the procuring organizations rather than the current situation played its part. By not utilizing user research methodology, the selected use contexts, user groups and tasks could have been biased or missing important details because they relied on the existing knowledge and interpretations of the domain experts. Similarly, the user scenarios or CLIPS employed in the demonstrations which formed the basis of evaluation for the developed HED and DPUQ methods were selected and written by domain experts based on their own interpretations of current work practices and the desired future state. This could mean that the scenarios were not entirely realistic representations of actual work. The planning of usability testing also relied on domain experts preparing the scripts in the beginning with the help of instructions before the usability researchers took responsibility for finishing them together with the domain experts. This means that while the final tasks and their presentation were prepared to be aligned with usability testing best practices, their premises could have been biased. These practices are, however, aligned with the suggestions in previous literature for the need of utilizing expertise from domain specialists with usability efforts when complex systems are in question (Lauesen, 1998; Jorritsma, Cnossen and van Ooijen, 2014). It should be noted, also, that the selection and depiction of users and their tasks were not the topic of research in this dissertation, but rather supporting material for applying the methodology.

The determining of weights for the usability attributes in the short-list evaluation phase of the procurement did not follow the mathematical approaches suggested in the general procurement literature of first determining the scoring rules and then the weights based on procurer preferences. Evaluating the possible effects on the outcome has not been done. Also, considering different weighting strategies and their possible effects on the outcome are not in the scope of this dissertation. However, it is noteworthy that one vendor’s solution consistently received better scores for all the measures and usability attributes evaluated in the two professional use contexts. Thus, the selection result reflects this, and the weights did not affect the overall result of the usability evaluation in this case. Furthermore, the overall weighting of usability against the other criteria in the procurement is not considered as part of this dissertation.

The definition of usability that formed the basis in the procurement project for the methodological choices was a combination of the ISO standard (2018) and Nielsen (1993) definitions. This can be considered to represent primarily
a situational viewpoint to usability (Hertzum, 2010). Focusing on this definition could have inadvertently dismissed other relevant viewpoints to usability and limited the evaluation. However, it should be noted that at least the viewpoints of perceived usability and some aspects of organizational usability were evaluated with the DPUQ and HED methods. Also, accessibility as one aspect of universal usability was evaluated for the citizens’ user group, although it was not included in the usability metrics in this dissertation.

The evaluation procedure includes over 10 usability measures and several usability evaluation methods. This is quite an extensive approach. No analysis has been done on what effect simplifying this procedure by leaving some metrics or methods out would have had on the usability evaluation outcome. The developed methodology for the preliminary assessment phase has been since applied in another procurement project in the same domain by the other usability researcher. The results in this other project supported the findings presented in this dissertation, although they have not been scientifically published. However, to better determine the generalizability of the developed methods and the methodological framework, they should be studied in further public procurement projects by other usability researchers. The application of the methods should be studied in procurements both in healthcare and social welfare as well as in other use contexts. The presented USco process model has been developed afterwards for this dissertation and is based on a combination of previous literature and the research done in the procurement project. The model should be validated and evaluated with other usability practitioners and researchers as well as implemented in new real-life procurement projects to determine its generalizability.

Finally, this dissertation work considers usability evaluation during the selection phase only from the procurer’s point of view. This is, however, in line with the action design research approach and the focus of the thesis. The considerations and effects of applying this methodology from the vendor’s viewpoint should be considered in the future.

5.7 Recommendations for Future Research

This dissertation is an attempt to describe the quantification of usability as a selection criterion in a complex public procurement context in much more detail than previous research. However, this work is based on one procurement project. As mentioned earlier, some of these methods and an alike approach have been applied also in another similar procurement in Finland by the other usability researcher. However, as these results have not been reported
scientifically, further, more detailed, research should be done on the developed methods HED and DPUQ, examining the points and validating the questions. Research on applying the overall methodological framework, and especially the short-list evaluation phase, in other procurements, would also be needed. The presented USco model should also be validated, tested, and possibly further developed. Moreover, the developed methods HED and DPUQ extended traditional usability evaluation methods to evaluate a larger scope of the system more efficiently based on seeing the system being used. For large-scale complex systems this type of evaluation would also be needed in the implementation and validation phases in addition to focusing on narrow parts of the workflows. Further development and application of the methods for this purpose should be explored.

Additionally, more research overall on including usability and usability evaluations as a selection criterion in public procurement within the EU is needed. Specifically, complex contexts such as healthcare and social welfare should be further regarded in this research. Also, the considerations emerging from literature on how to define valid weighting functions and scoring rules in public procurement should be studied in relation to usability and its attributes, metrics, and measures. This dissertation viewed the topic purely from the perspective of the procurer. Critical viewpoints of one vendor to usability evaluation practices during procurement have been reported (Tarkkanen and Harkke, 2016; Tarkkanen, Koskinen and Harkke, 2016). Other vendors and their perceptions based on more cases should be studied to understand further how they view including usability aspects overall into the procurement, and usability evaluation specifically into the selection phase. I have gathered this type of research data from the winning vendor in the researched procurement project, but the publication of the results is pending analysis and writing.

Moreover, there are also other interesting directions for future research that are not as directly related to the topic of this dissertation. They have rather emerged from the experience of working in a real-life public procurement and implementation project for several years. In the researched procurement project, another evaluated selection criterion was the configurability of the IS. The relation of these two concepts is perhaps not self-evident, at least during the selection. However, the configuration done during the implementation phase has been studied to affect the usability of the IS significantly (Ratwani et al., 2016). Thus, the question remains what is the effect of configurability on the usability evaluation results for the selection? Further research into the connection between these concepts and its implications to the procurement selection phase should be done.
Furthermore, the selection phase is only one part of solving the puzzle for providing end users with information systems with better usability and creating a positive user experience (UX) for them. This phase is the basis that is built upon in the later stages of a procurement project. Complex procurement and contractual factors have been identified to affect usability efforts during implementation customization negatively (Lee, Williams and Sheikh, 2016). Usability engineering efforts during the implementation stage, and how contractual requirements around it can have an effect on the efforts, should be studied further.

When looking at the bigger picture of a large-scale procurement project, what are the aspects of the implementation stage that make it difficult to achieve good usability? This situation has been realized recently in both Denmark and Finland (Hertzum, Ellingsen and Cajander, 2022). A related question is how these obstacles could be conquered in real life? One affecting reason could be the viewpoints on usability (Hertzum, 2010) used for evaluating and engineering usability during and after the selection phase. The process of considering the different images of usability and selecting those relevant in the context of the procurement case already in the beginning could have a positive impact. Hertzum (2018) presents how organizational usability can be evaluated during the later phases of procurement. However, research on the methodology of including the measurement of different images in the selection phase of complex information systems would be needed.

Finally, the relation of the concepts of usability and UX within the healthcare and social welfare professional IS context has not been studied scientifically to my knowledge. While the experiences of these users on the usability of their systems have been studied, also in Paper VI included in this dissertation, the public discussion seems to also involve viewpoints that can be considered closer to the concept of UX, such as emotions and expectations. Good usability, while central in efficient and effective utilization of these tools in the work, might not be enough to create satisfied end users for health and social welfare IS.
6. Conclusions

There are multiple viewpoints that affect planning and implementation of usability evaluation for selection of IS in public procurement. They emerge from the requirements presented by the regulatory environment and the goals of each project. This adds complexity to the procedure compared to regular summative evaluation of usability. In other contexts, the summative evaluation is also usually more confined in scope.

Based on this dissertation work, there are no methodological, procedural, or processual reasons not to plan for measuring usability during the selection phase, even when procuring a large-scale IS. Additionally, it is recommended to include usability in the contractual requirements that vendors need to agree to fulfill. However, based on literature, there are challenges with both writing verifiable requirements and requirements that vendors will agree on fulfilling. Also, writing usability requirements that will cover a large-scale IS to an appropriate extent takes effort and can be challenging. It can also be assumed that verifying those requirements with summative usability evaluation during implementation will take a substantial amount of effort from the procurer when large-scale systems are in question. Therefore, based on this dissertation work, an approach that combines usability both in the contractual requirements and as a selection criterion would be recommended.

In the following, I summarize the answers to the research questions and the devised usability scoring model, and present related conclusions from three different viewpoints: methodological, procedural, and processual.

*Methodological Viewpoint: Methods and Metrics for Quantifying Usability*

The first research question addressed the measurement of usability on a methodological level: what are the suitable usability evaluation methods and metrics to use during the selection phase of public procurement of IS?

Answering this question was done based on previous literature and by implementing new and existing usability evaluation methods in the action design
research project. Most used usability evaluation methods in previous HIS procurements included established methods such as satisfaction questionnaires after use and performance measures from usability testing or on-site testing. Concrete examples of measures and scoring suitable for public procurement was however lacking. The methods are also resource intensive for early phases of procurement with numerous vendors in complex use contexts of healthcare and social welfare. Based on the recommendations in literature for heuristic evaluation and using clinical information processing scenarios, two new methods, HED and DPUQ, were developed. They were used in the project with demonstrations in the early stage. The first method quantifies heuristic evaluation to rank the quality of UI design and the second method is a perceived-usability questionnaire for users. In the later stage, several measures and methods were used to conduct a more focused evaluation. An overview of all these is presented in the results. The main method used in this stage was paired-user testing and the dissertation presents scoring functions for it in more detail than previously. All three methods and the used measures were deemed suitable for public procurement based on the assessment in the results. The methods differ in their efficiency, comprehensiveness, user involvement, level of usability expertise needed, and ability to produce information for implementation.

From a methodological viewpoint, the dissertation work led to the following conclusion: There are several usability evaluation methods that can be used during public procurement to reliably produce comparable results on usability. Both methods with and without users are suitable to measure usability, although the objectivity demand of public procurement must be taken into consideration. The comparison between methods in our case indicate that from a pure ranking standpoint even just the developed heuristic evaluation method for demonstrations, HED, could suffice in evaluating usability. HED also outperforms traditional heuristic evaluation. However, user involvement is recommended for a successful procurement overall. When demonstrations are used to evaluate other aspects than usability by users, also adding the developed perceived usability questionnaire, DPUQ, is natural. Based on our results, when involving users, using only a few metrics and measures can reveal the differences between vendors’ products regarding usability.

The recommendations based on the dissertation work are as follows. Ultimately the selection of methods and number of metrics (and measures) should depend on the resources and goals of the procurement and the expectations for the utility of the evaluation material after selection. If resources are scarce, planning elaborate evaluation is not feasible. Using HED and DPUQ with
extensive scenario-based demonstrations can cover a large-scale system efficiently. However, if resourcing is not an issue, and there are plans to further develop the IS after selection, user testing with extensive scenarios, that gives also qualitative data on usability, are a good addition.

Procedural Viewpoint: Construction of the Usability Evaluation Procedure

The second research question addressed the quantification of usability from a procedural level: how to construct the complete evaluation from these different methods and metrics?

This question was answered by utilizing recommendations in previous literature and constructing the evaluation procedure for the researched procurement project. The usability evaluation was constructed to occur in two phases, as had been already decided in the procurement as the overall structure for other evaluations. This was also suggested by literature. In the first, preliminary assessment, phase a minimum requirement for usability was set and evaluated. Additionally, the vendors’ systems were ranked for usability as a maximum number of vendors was set to enter the second phase. The demands for the methods and their scoring of usability required getting a broad evaluation of the system use contexts and setting a minimum level beforehand. HED and DPUQ were mainly utilized in this phase. The second, short-list evaluation, phase followed the recommendation from literature to use several metrics to overcome contradictory results that are typical for subjective and objective metrics.

Legislation for public procurement requires the evaluations to be transparent and treat vendors equally. Also, the minimum requirements and weights of criteria used in the procurement need to be disclosed in the beginning. According to literature, the weights should reflect the procurement goals and be set based on the scoring for each criterion. Setting the weights thus requires first establishing the scoring rules for each metric within the criterion. This introduces detailed planning of evaluations beforehand for unfamiliar systems. The dissertation work presents one unique way of quantifying usability based the goals of the researched procurement project. The approach included several metrics and methods that covered evaluation of usability for multiple user groups and use contexts.

From a procedural viewpoint, the dissertation work led to the following conclusion: Based on the results of this research a two-phase usability evaluation procedure is indeed practically suitable for, as well as effective in, public procurement of IS. The first phase should evaluate the system as broadly as possible and the second phase should focus on a narrower, yet still as widely
representable as possible, scope. The usability attributes that are relevant to measure and their relative weights depend on the procurement goals. Establishing these goals is very important in the beginning. And as setting the weights correctly requires first establishing the scoring rules behind them, this can get complex as the number of metrics increases. If usability attributes’ relevance also differs based on user group or context of use, as in the researched procurement, this adds to the complexity.

**Processual Viewpoint: Preparing for Usability Evaluation**

Based on both the research work done in this dissertation and going through the related literature, a process model for scoring usability was constructed. The model presented in the dissertation is the first version and has not been validated. The USco model includes the components for including usability evaluation in the selection of IS in public procurement. First, it includes the preceding analysis stage of usability goal setting for the procurement where use contexts and user groups and their relevant usability attributes are identified and linked to the goals. This stage informs the following stages. Then, the model consists of process steps for three different stages: defining the usability ranking and usability measurement models as well as planning the more focused evaluation.

The purpose of the usability ranking model is to describe the ranking details and the minimum level requirement in the first phase of the evaluation. The purpose of the usability measurement model is then to describe the scoring of usability in the final selection in detail. Finally, the purpose of the planning stage is to choose the areas to evaluate, as well as to prepare the scripts, evaluators, and vendors for the actual implementation of the evaluation. Additionally, the model emphasizes understanding of the context of use where the IS to be procured will be implemented. This understanding is needed in all stages.

From a processual viewpoint, the dissertation work led to the following recommendation: The results of this dissertation implicite, that including usability as a minimum requirement in the selection and as a final selection criterion, can influence the procurement result. This can happen by excluding systems with poor usability early on before spending more time on the procurement and by not selecting a system with inferior usability. Both these aspects are included in the components of the USco model. Usability evaluation of unfamiliar systems that reflect future needs rather than current state, together with the requirements of public procurement, requires very detailed
planning and utilization of both usability and domain expertise. A process model, such as the one described above, helps with this planning.

Finally, this dissertation contributes to the scientific body of work from three perspectives. First, by providing concrete knowledge of public procurement and including usability as a selection criterion. Second, by enhancing existing usability evaluation methods further to include previously lacking aspects in order to utilize them in evaluating a large-scale system. And third, by introducing new ways to quantify usability in public procurement as well as giving supporting evidence and examples for approaches presented in literature. This dissertation work is, however, based on one procurement project. Future work should further evaluate the developed methods, the presented procedure, and the formed process model for scoring usability in other public procurements.


References


References


References


References


References


References


158


